Model-Based Clustering Multivariate EMA Time-Series Data

Dumitru Verşebeniuc
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—In recent times, smartphones and watches have facilitated researchers in collecting real-time through ecological momentary assessment (EMA). This data enables the exploration of human behavior, emotions, and psychological processes in their natural environments. Clustering is one of the techniques to identify different subgroups of individuals based on extracted information from EMA items. This study focuses on model-based clustering, which involves extracting information/coefficients from linear or non-linear machine-learning models and inputting extracted coefficients to the traditional clustering algorithm, such as K-Means or hierarchical clustering. Primarily, linear models are used that utilized linear parameter estimates as coefficients and non-linear models, such as decision models, where feature importance is used instead of coefficients. Results demonstrate the predictive performance of linear and non-linear models as well as the clustering performance and analysis of distinct clusters of individuals. Moreover, Linear Lasso Regression (Lasso) with K-Means illustrates a balance cluster distributions for 2 clusters, highlighting special characteristics of individuals within each cluster. Nevertheless, Vector Autoregression (VAR) with K-Means exclusively assigns cluster based on a single attribute of individuals. Generally, Model-based clustering techniques, such as VAR and Lasso K-Means demonstrate the ability to capture behavioral patterns among individuals, which are represented as coefficient matrices, for grouping them into distinct clusters.**

***Index Terms*—Model-based clustering algorithm, Multivariate time series, Ecological Momentary Assessment**

I. Introduction

Ecological Momentary Assessment (EMA) is a method of collecting real-time data on individuals' behaviors, feelings, and psychological processes in their natural environments. EMA data allows researchers to gain a more complete understanding of health-related behaviors and mental disorders. Clustering EMA data can help identify different subgroups based on extracted features, potentially providing crucial insights into individuals' feelings, addictions, and behaviors that vary over time. This, in turn, can support researchers in studying and understanding disorders and helping individuals dealing with mental and wellness issues.

Analyzing EMA data can be challenging due to its high dimensionality, missing data points, and complexity, making it difficult to identify patterns and make inferences. Clustering is a useful technique for exploring the structure of high-dimensional data and grouping similar observations together. Multivariate time-series clustering can be classified into three types: raw-based, feature-based, and model-based approaches [1].

Raw-based approach in multivariate time-series clustering involves directly analyzing the raw data points of the time series. It focuses on the similarity or dissimilarity between the individual data points. Clustering algorithms such as k-means with dynamic time warping (DTW) can be applied to group similar time series together based on the proximity of their raw data points. In this approach, the clustering is based on the values and patterns observed in the time series.

Feature-based approach in multivariate time-series clustering performs by extracting relevant features from the time series data and then applying clustering techniques to group similar time series based on these extracted features. Feature extraction methods, such as statistical measures (e.g., mean, standard deviation, variance) or dimensional reduction (e.g., principle component analysis, principal feature analysis) [2], are used to capture important characteristics of the time series. These extracted features are then used as inputs to clustering algorithms.

Model-based clustering involves three approaches: machine learning-based (ML), representation-based, and distribution-based (i.e., generative models). The ML approach uses extracted coefficients from the model or the model itself as input for a clustering algorithm based on distance metrics, such as k-means or hierarchical clustering, or maps the original standardized data to model representation and performs a cluster assignment by the model itself. Representation-based approaches transform the data from a high-dimensional space to a space of fewer dimensions before fitting the model. The data transformation may be linear, as in principal component analysis (PCA), but many nonlinear dimensionality reduction techniques also exist. Distribution-based clustering is a method of clustering data points using probabilistic models, such as Gaussian Mixture Model or Hidden Markov model, to estimate the data distribution and perform clustering based on distributions or characteristics of the statistical model.

Since ground truth labels are not available, the evaluation of clustering models must rely on the model itself. Addition-

This thesis was prepared in partial fulfillment of the requirements for the Degree of Bachelor of Science in Data Science and Artificial Intelligence, Maastricht University. Supervisor(s): Jerry Spanakis (G.), Mado Ntekouli

ally, this study aims to evaluate the performance of machine learning models, which may be complicated due to temporal dependencies and potential non-stationarity of the data.

The objective of this research is to compare the performance and results of different model-based clustering methods and a raw-based clustering methods in terms of their ability to efficiently group similar individuals and identify significant patterns in the data. This study can assist practitioners in selecting the optimal method for their particular EMA datasets and analysis objectives.

Research questions:

1) How can different machine learning models be used to extract information for clustering multivariate EMA time-series data?

2) How does the choice of different model-based clustering methods affect the performance of clustering for multivariate EMA time-series data?

II. Methods

The Methods section describes data and techniques used for clustering multivariate EMA time-series data through model-based approach. This includes information on the data pre-processing steps and a description of model-based approach for clustering the extracted coefficients. The model-based approach consists of two components: a predictive machine learning model for extracting the coefficient matrix from individual, and a clustering method for grouping these coefficient matrices into distinct clusters. Additionally, this section also describes a raw-based approach as baseline clustering approach for further comparison with model-based approaches.

A. Data

The original dataset comprises data from 277 individuals who completed eight surveys per day for 28 consecutive days. In other words, there are at most 224 prompted assessments nested within an individual. The dataset contains 56,499 data points across 61 features, including user ID, device ID, issued/response time, and survey responses (EMA items). Responses are represented as integers on a 7-Likert scale, where 1 indicates 'Not at all' and 7 indicates 'Extremely'.

B. Data preprocessing

To reduce the dimensionality of the data and simplify the model fitting process, the 61 features were grouped into similar sub-categories, resulting in 12 behavior features, some of which represent the average number of positive and negative states.

The data was cleaned by removing the missing questionnaire responses, rows with an 'Expired' title in the Duration column and rows with NaN values. Additionally, individuals with a small number of data points were excluded from the data set. The threshold for inclusion in the final data set was at least 112 data points, which is 50% of total prompted assessments, per individual. As a result, the final dataset used includes data from 187 individuals, with 12 features, according to domain experts, and 31,291 data points. Additionally, indices were changed to the time representation in the format of YYYY-MM-DD HH:MM and added a column with the unique ID of participants.

In order to meet method requirements such as Lasso regularization, Ridge regularization, and Polynomial regression, and to avoid biased application of diverse data across individuals, standardization [3] is applied to each individual independently.

C. Data Evaluation

Data were checked for several attributes in order to obtain a concise overview of the complexity of the data structure.

1) Response time-delays: Every individual was checked for day delays, and it was found that the maximum non-response delay was two days. Therefore, it is possible to miss some important behavior patterns of individuals.

2) Constant features: Some individuals have constant and quasi-constant features[1]. In "Tab. I", we can observe, how many columns into individual data have constant or quasi-constant with 99% of the same observation.

TABLE I
Constant and Quasi-constant features in the individual data

Feature	Number of appearance	
	Constant	**Quasi Constant 99%**
Crave Food	11	10
Crave Other	56	34
Impulsivity	10	11
Self esteem	1	–
Worried	1	–
Somatic negative	–	2

When constant or quasi-constant features are present in the data, they may dominate the clustering process due to their zero variance, which can enhance similarity within subjects with zero variance. Consequently, the resulting clusters may primarily reflect the characteristics of the constant features, rather than effectively capturing meaningful patterns in individuals.

3) Non-stationary effect: Data is a collection of real-time data from individuals from their natural environment. Therefore, individuals' experiences and behaviors are influenced by various situational and contextual factors that may change from one moment to another, which can arise non-stationary effects. Non-stationary effects arise when relationships between variables fluctuate or vary across time points [4].

[1]Quasi-constants – are those that show the same value for a great majority of observations in the dataset.

D. Clustering by coefficients

An individual's data representation within the model can be understood using coefficients extrapolated from a Machine Learning model. These coefficients are frequently linked to the significance or input that particular characteristics or variables have when making predictions. While coefficients may not be able to fully define a person, they can capture the underlying patterns and relationships in the data. Therefore, clustering by coefficients can be considered one of the approaches to group individuals, who share similar feature importance patterns or demonstrate similar relationships with the target variable. In "Fig. 1", we can see the raw representation of how individuals can be clustered by extracted coefficients of the Predictive ML model "Section II-F".

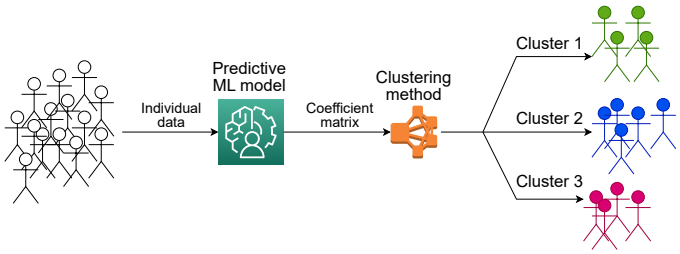

Fig. 1. Example illustrating the basic principle of clustering based on coefficients.

E. Coefficient matrix

The coefficient matrix is a representation of regression coefficients or feature importance, in the case of decision tree models, extracted from a machine learning model. To generate this matrix, individual data points are used to fit one or more machine learning models. The resulting output is an $n \times c$ matrix, where n denotes the number of individuals and c represents the number of coefficients, where the size of c depends on the applied model. The process of fitting individual data points depends on the type of model used: *regression* or *autoregression*.

1) Regression models: Regression models relate one or more explanatory variables to a scalar response. Consequently, each feature requires its own model. To fit a model with all individuals' data points as independent variables, the dependent variable, which is one of the features, should be shifted by one time-step. This shifting is necessary for the model to interpret the individual by predicting their behavior.

In "Fig. 2", we can observe a process of extracting coefficients from Linear Regression models for one individual. Individual data is split into 2 sets: predictors set (all individual data points with all features) and response (one-dimensional shifted set for every feature) sets. After splitting data, Linear Regression model is fitted with predictors set, and response set one by one for each model independently. In the end, all models are extracted coefficients and appended together creating a coefficient matrix.

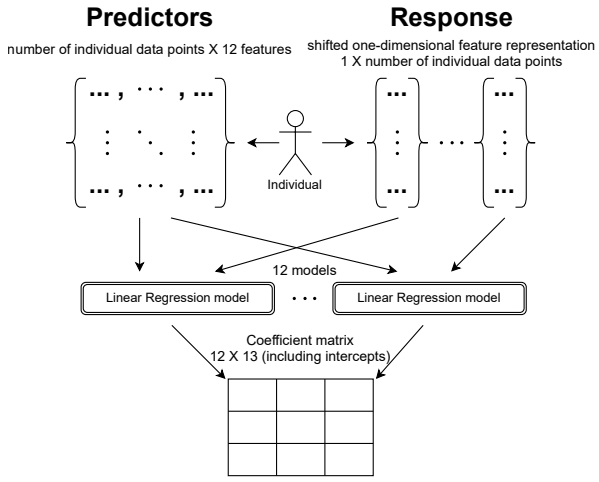

Fig. 2. Example of creating a coefficient matrix for an individual using Linear Regression model.

2) Autoregression models: In Autoregression models the output variable depends on its own previous values, on a stochastic term, and on correlation with other variables in the case of multivariate data. However, this approach is similar to regression models, autoregression also considers time-step, which can be useful for time-series data. Therefore, one autoregression model can be fitted with all individuals' data points at once and generate a coefficient matrix.

F. Predictive Machine Learning models

Predictive Machine Learning models for extracting coefficients can be divided into two types: *linear* and *non-linear* models. These models have distinct characteristics and are applied in different scenarios, which we cover further.

1) Linear models: Linear models describe a continuous response variable as a linear function of one or more predictor variables [5].

a) Standard Linear Regression: Standard Linear Regression is a statistical method used to create a linear model. The model describes the relationship between a dependent variable y (also called the response) as a function of one or more independent variables X_i (called the predictors) [6]. The general equation for a linear model is:

$$y = \beta_0 + \sum \beta_i X_i + \epsilon_i \tag{1}$$

where β represents linear parameter estimates to be computed and ϵ represents the error terms. Linear regression was taken as a basic linear model to compare other Machine Learning models.

b) Regularization: Linear machine learning models can be utilized with regularization techniques: *Lasso* and *Ridge regularization*.

- *Lasso regularization:* Lasso is a regularization technique that combines variable selection and regularization to improve prediction accuracy and interpretability in statistical regression models [7].

- *Ridge regularization:* Ridge regularization is a method used to estimate the coefficients of multiple regression models in situations where the independent variables exhibit high correlation [8].

c) Polynomial Regression: Polynomial Regression is a regression analysis technique that models the non-linear relationship between the independent variable x and the dependent variable y as a polynomial function of degree n [9]. Moreover, it is used when Linear Regression models may not adequately capture the complexity of the relationship in the data [10]. Therefore, Polynomial Regression can capture more complex and unique behavior patterns of an individual, which can be useful for further clustering.

d) Vector Autoregression: Vector Autoregression (VAR) is a statistical model used to capture the relationship between multiple quantities, known as endogenous variables, as they evolve over time. It enables the incorporation of input multivariate time-series data into the model. [11]. Moreover, VAR model is widely used for calculating coefficients for Ecological Momentary Assessment (EMA) data. [12], [13].

2) Non-Linear models: Non-Linear models describe non-linear relationships in data.

a) Random Forest Regression: Random Forest Regression is a supervised learning algorithm that utilizes an ensemble learning method called bagging. Bagging involves training multiple decision trees simultaneously and combining their outputs through a majority vote to determine the final model [14]. Random Forest is an easily interpretable model and can be an example of a decision model for extracting feature importance from the data.

b) XGBoost Tree Regression: XGBoost Tree Regression is a supervised learning algorithm that also employs an ensemble learning method similar to a random forest. However, in the case of Extreme Gradient Boosting (XGBoost), it utilizes a boosting technique where decision trees are created sequentially and each subsequent tree aims to improve upon the mistakes made by the previous ones [14]. The final model is obtained by summing the outputs from all the individual trees. Since, XGBoost is used a different ensemble learning method from Random Forest, it allows a comparison between two decision models.

In future work, it would be beneficial to consider the inclusion of other machine learning models, such as Support Vector Regression, Neural Networks, or Probabilistic models, but due to their complexity and sensitivity to parameter settings, they were excluded from this study.

G. Clustering methods for coefficients

Coefficient matrices can be grouped by equal variance or covariance, since they represent the behavior pattern of individuals. Sub-groups' coefficient matrices with equal variance and high covariance could be generated in case where such behavior patterns are similar to each other. To cluster such matrices, we utilized several methods including *K-Means*, *K-Medoids*, *Agglomerative hierarchical clustering with ward linkage*, and *Gaussian mixture clustering*, which aim to group the data into n clusters of either equal variance or covariance, as in the case of the Gaussian mixture model (GMM).

Additionally, the performance of k-means, k-medoids and GMM heavily depends on the initial choice of cluster centers. The *k-means++* algorithm addresses this issue by improving the initialization of the centroids [15]. In k-means++, the initial centroids are chosen in a way that maximizes the distance between them, reducing the chances of selecting centroids that are too close to each other. This helps to avoid convergence to a suboptimal solution and leads to better clustering results.

1) K-Means: K-Means works by dividing a dataset into k clusters, where each data point belongs to the cluster with the nearest mean.

2) K-Medoids: K-Medoids is a non-parametric clustering algorithm that uses representative points, known as medoids, to define clusters. Unlike k-means, which uses the mean of all points in a cluster as the centroid, k-medoids uses the actual data point that is closest to the center of the cluster as the medoid [16]. This makes k-medoids more robust to outliers and noise, as the medoid is less sensitive to extreme values than the mean.

3) Agglomerative hierarchical clustering with ward linkage: Agglomerative hierarchical clustering with ward linkage is a hierarchical clustering algorithm that aims to minimize the variance of the clusters being merged at each step. It is a variance-minimizing approach and in this sense is similar to the k-means objective function, but tackled with an agglomerative hierarchical approach [17].

4) Gaussian mixture clustering: Gaussian mixture clustering assumes that the data is a mixture of multiple Gaussian distributions with different means, covariances, and weights [18]. The goal of this algorithm is to find the parameters of the Gaussian mixture that best fit the data. This is done by optimizing the likelihood function - Expectation-Maximization (EM) algorithm [19], which is the probability of observing the data given the parameters of the Gaussian mixture.

H. Baseline raw-based clustering model

K-Means DTW: K-Means clustering with Dynamic Time Warping can be applied to our multivariate time-series data as a baseline method. Dynamic Time Warping is used to calculate the distance between time-series of similar shapes. Therefore, each individual data set should have an equal size of data points with 12 features, if an individual does not have the desired number of data points, we can fill the gap with NaN values. Cluster centroids, or barycenters, are computed with respect to DTW. A barycenter is the average sequence from a group of time series in DTW space. The DTW Barycenter Averaging (DBA) algorithm minimizes the sum of the squared DTW distance between the barycenter and the series in the cluster. As a result, the centroids have an average shape that mimics the shape of the members of the cluster, regardless of where the temporal shifts occur amongst the members [20].

III. Experiments

In Experiments section discusses methods for evaluation and analysis of model-based clustering results. This includes assessing the performance of predictive models and clustering methods, as well as analysis of obtained clusters from the model-based approach. Predictive model evaluation can be measured by quantifiable errors in the model's predictions, enabling a comparison between different models and demonstrating their performance across varying data sizes during the model fitting process. Clustering results can be compared with indicators that do not rely on ground labels in their calculations and can be applied to all clustering outcomes. Furthermore, cluster analysis needs to explain the composition of individuals within clusters, either partially or completely.

A. Predictive model evaluation

Performance evaluation of models can be measured by the difference between values predicted by a model and the values observed, also called the root-mean-square error (RMSE). This measurement is applied to the training, test, and 5-fold time-series cross-validation data split, as depicted in "Fig. 3".

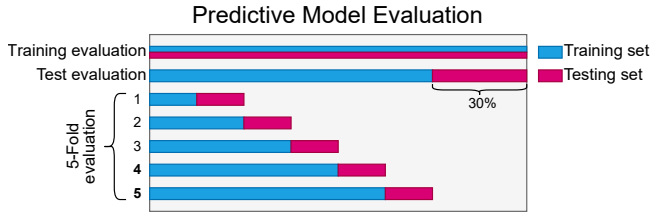

Fig. 3. Example of training, test, and 5-fold time-series cross-validation evaluation.

1) Training evaluation: Training evaluation demonstrates the model's ability to accurately replicate individual data.

2) Test evaluation: Test evaluation illustrates the model's capacity to make precise predictions for forthcoming individual data points.

3) 5-Fold time-series cross-validation: 5-Fold time-series cross-validation shows the model's ability to accurately predict future individual data points when a model is trained on varying data sizes.

Overall, these evaluations aim to effectively reproduce individuals' behavioral patterns in various situations. The final result displays the overall mean and standard deviation of the root-mean-square errors for each model evaluation. Moreover, the average root-mean-square error was calculated for each feature in the dataset, and decision tree models are set up with static initialization to have reproducible results in every run.

B. Cluster evaluation

Performance evaluation of clustering results can be measured by indicators such as the Silhouette coefficient, Calinski-Harabasz Index and Davies-Bouldin Index.

1) Silhouette coefficient: Silhouette coefficient measures the similarity of each data point to its own cluster compared to other clusters, providing a score ranging from -1 to 1, where higher values indicate better clustering results [21].

2) Calinski-Harabasz Index: Calinski-Harabasz Index measures the ratio of the between-cluster variance to the within-cluster variance and can be used to evaluate the model, where a higher score relates to a model with better-defined clusters [22].

3) Davies-Bouldin Index: Davies-Bouldin Index measures the average similarity between each cluster and its most similar cluster, relative to the average dissimilarity between each cluster and its least similar cluster [23]. The index provides a score ranging from 0 to infinity, where lower values indicate better clustering results.

These indicators can help in understanding the quality of clusters and facilitate the comparison of clustering results obtained from different models with the clustering methods listed in this study.

C. Cluster Analysis

Clustering methods, such as k-means, k-medoids and Gaussian Mixture, depend on the initial position of cluster centers, which makes every run of these methods random. Therefore, in order to have robust and reproducible results, evaluation is performed over 100 iterations with different established initialization for each clustering method.

1) Quality: Cluster results represent the average and standard deviation of cluster evaluation methods "Section III-B" over 100 iterations.

2) Individual clusters: Individual clusters are obtained as final clusters over 100 iterations. After each run of the clustering methods, the cluster labels are saved, and these labels are subsequently fitted to the K-Medoids method, as we can see in "Fig. 4". This process results in final clusters over all iterations.

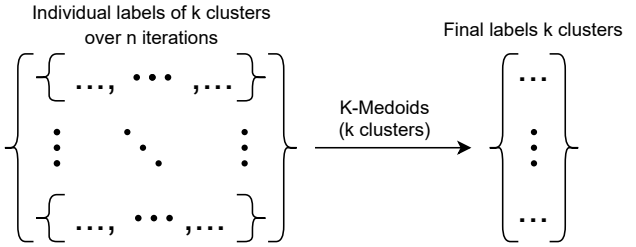

Fig. 4. Example of calculating final k clusters over n iterations

Additionally, individuals who were assigned to different clusters throughout the iterations can be identified by calculating the frequency of distances to centroids of K-Medoids. The k most frequent distances, corresponding to the number of clusters k used in the iterations, determine the individuals who consistently belonged to the same cluster across all iterations. On the other hand, the remaining distance frequencies represent individuals whose cluster assignments fluctuated over the iterations.

3) Coefficient representation: Metric Multi-Dimensional Scaling (MMDS) is a dimensionality reduction technique used to represent the coefficient matrix of individuals in a three-dimensional (3D) plot. In a two-dimensional (2D) plot, it can be challenging to separate and distinguish coefficient matrices that are close together or overlap. Moreover, with three-dimensional, we can capture more information and explore coefficient matrices from different viewpoints and angles.

MMDS aims to preserve the pairwise distances between coefficient matrices while mapping them to a lower-dimensional space [24]. Additionally, to facilitate a better understanding of the clustering of coefficient matrices, each individual data point in the 3D plot is color-coded according to its assigned cluster.

4) Cluster characteristic: Cluster assignments can also be explained by the characteristics of the original individual data. Dissimilarities and similarities between different clusters can be represented by the average, standard deviation, and correlation matrix of the original features.

Cluster Analysis can help in interpreting and understanding cluster assignments of individuals and their coefficient matrices.

IV. Results

The Results section presents information regarding the performance of predictive models and clustering methods, as well as a cluster analysis of two model-based clustering approaches. The predictive models are examined in three different evaluations "Section III-A", which results are illustrated in boxplots and a table showing the mean and standard deviation of RMSE for each model. Additionally, the performance of clustering methods "Section II-G", which are fitted by extracted coefficients is demonstrated in tables with clustering evaluations "Section III-B" and final clusters for each predictive model and in figures with clustering evaluations over different k clusters. Finally, two model-based approaches, which exhibited interesting results in both predictive and clustering evaluations, are selected for further cluster analysis "Section III-C".

A. Predictive model evaluation

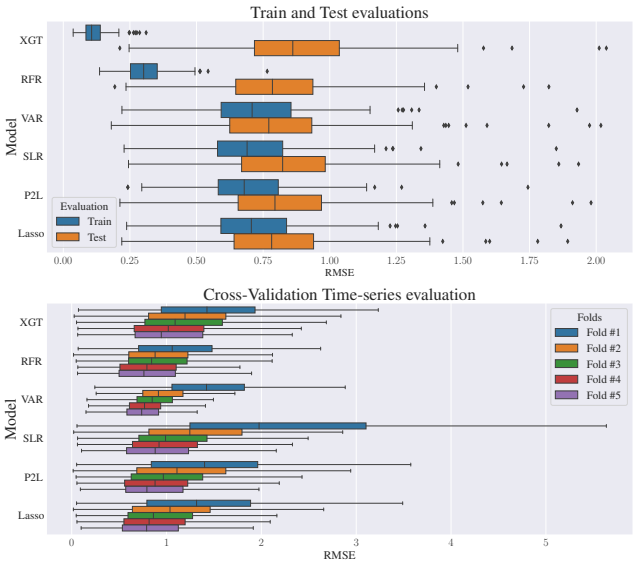

Fig. 5. Boxplots of RMSE for three evaluations approaches

"Fig. 5" illustrates boxplots of the predictive model evaluations "Section III-A". Training and test evaluations are depicted in the same plot, including outliers, which helps to understand the difference between models and evaluations. However, for the boxplots of cross-validation time-series evaluation, a separate plot is used without outliers, as their high values can make the plot less readable. Additionally, boxplots are added for each fold for CV time-series evaluation to have a clear picture of how data size is influenced the predictive performance of the model.

"Tab. II" demonstrates the results of the predictive model evaluation "Section III-A". Every evaluation has its mean values μ and standard deviation values σ. Additionally, predictive models have a column labeled *'n coef'*, which denotes the number of coefficients for one individual in a

TABLE II
RMSE MEAN AND STANDARD DEVIATION OF PREDICTIVE MODELS EVALUATION

Model	*Train*		*Test*		*CV Time-series*		n *coef*
	μ	σ	μ	σ	μ	σ	
XGT	**0.117**	**0.048**	0.898	0.275	1.270	0.708	144
RFR	0.309	0.083	0.808	**0.248**	**0.987**	**0.568**	144
VAR	0.735	0.220	**0.806**	0.280	1.034	0.704	144
SLR	0.716	0.211	0.851	0.262	1.484	1.266	156
P2L	0.702	0.193	0.833	0.268	1.188	0.711	1092
Lasso	0.730	0.212	0.813	0.257	1.096	0.623	156

coefficient matrix.

1) XGBoost Tree Regression (XGT): From "Tab. II", Extreme Gradient Boosting Tree Regression (XGT) shows the lowest average root-mean-square error for the training evaluation. However, mean μ values of testing and cross-validation (CV) time-series evaluations are one of the highest values among predictive models in "Tab. II". This suggests that XGT tends to overfit and perform worse against unseen data compared to other models.

2) Random Forest Regression (RFR): Random Forest Regression, which also belongs to the decision tree family like XGT, exhibits good performance in all three evaluations based on their mean μ and standard deviation σ values.

3) Standard Linear Regression (SLR): Standard Linear Regression demonstrates relatively good results in training and testing evaluations compared to other linear models. Nonetheless, it performs poorly in the CV time-series evaluation showing the longest whiskers in "Fig. 5" and the highest standard deviation in "Tab. II", indicating the dataset size plays a significant role in prediction accuracy for an individual.

4) Polynomial Regression: Polynomial Regression shows a limited ability to predict the future states of individuals in testing and cross-validation time-series evaluations. Therefore, it is not listed in "Tab. II".

5) Lasso regularization (Lasso & P2L): Linear Regression and Polynomial second-degree with Lasso regularization (Lasso and P2L respectively) show improvements in all three evaluations compared to their standard version of the models. Thus, using Lasso regularization can shrink features, which can positively influence the accuracy of the model. Furthermore, increasing the polynomial degree does not improve the accuracy of the model.

6) Ridge regularization: Ridge regularization, when applied to models like Linear Regression and Polynomial Regression, does not demonstrate substantial improvements in accuracy. Consequently, it is omitted from "Tab. II".

7) Vector Autoregression (VAR): Vector Autoregression illustrates the lowest results in testing and CV time-series evaluation among linear models in "Tab. II", indicating its ability to handle and forecast the future states of individuals. However, it performs the worst terms of training RMSE among all the models in "Tab. II", while still maintaining an average RMSE error among the linear models.

Predictive models, listed in "Tab. II", are employed in clustering methods due to their demonstrated good accuracy.

B. Clustering methods

"Fig. 6, 8" and "Appendix Fig. 7, 9" demonstrate the average values of cluster evaluations over 100 iterations with different numbers of clusters. All clustering methods "Section II-G" show a gradual decrease in quality "Section III-C1" as the number of clusters increases. Additionally, Linear Regression and Polynomial Regression with Lasso Regularization (Lasso & P2L respectively) produce similar results, which could be related to equal feature shrinkage.

Figures also demonstrate interesting observation that all clustering methods illustrate high clustering performance with 2 clusters. Consequently, the results of the 2 clusters are presented in "Tab. III, V, VI" and " Appendix Tab. IV"

1) K-means & K-Medoids: K-Means and K-medoids show nearly identical results in all cluster evaluations (see "Fig. 6" and "Appendix Fig. 7") and slightly different cluster distributions of final clusters (refer to "Tab. III" and "Appendix Tab. IV").

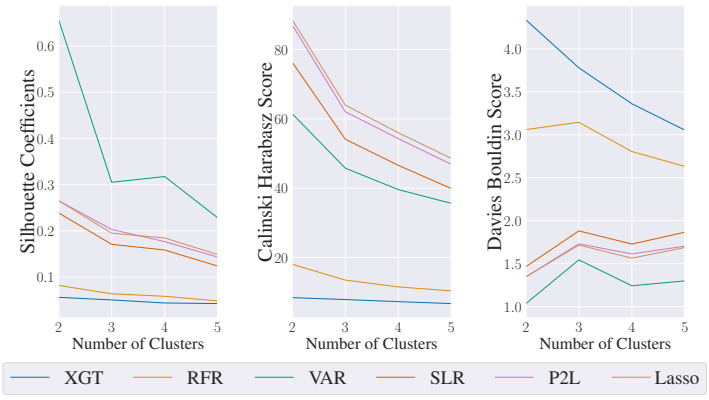

Fig. 6. K-Means clustering

From "Tab. III", we can observe that each predictive model's coefficients clustering has a small standard deviation across 100 iterations, indicating their robustness against the randomness of centroid initializations.

In "Tab. III", all predictive model's coefficients clustering (except VAR) shows a balanced number of individuals in both

TABLE III
RESULTS OF K-MEANS WITH 2 CLUSTERS OVER PREDICTIVE MODEL'S COEFFICIENTS (BASELINE RAW-BASED CLUSTERING INCLUDED).

Model	*SC*		*CH*		*DB*		*CLUS 1*	*CLUS 2*
	μ	σ	μ	σ	μ	σ		
VAR	0.656	0.002	61.442	0.293	1.036	0.034	179	8
P2L	0.265	0‘	87.038	0‘	1.347	0‘	100	87
Lasso	0.265	0‘	88.530	0.008	1.347	0.002	98	89
SLR	0.239	0‘	76.309	0.010	1.466	0.002	100	87
DTW*	0.141	0.036	17.283	0.03	0.134	0.134	101	86
RFR	0.082	0.001	18.029	0.013	3.060	0.004	105	82
XGT	0.056	0.013	8.410	0.121	4.335	0.229	93	94

* Baseline raw-based clustering method.
‘ value is close to zero.

clusters. However, VAR coefficients clustering demonstrates the highest Silhouette Coefficient (SC), indicating a good separation of individuals by groups from each other. Additionally, coefficients of decision tree models, such as Random Forest Regression (RFR) and XGBoost Tree (XGT), show relatively bad results in separation among all model's coefficients and baseline clustering method Dynamic Time Warping K-Means (DTW).

2) Agglomerative hierarchical clustering: Agglomerative hierarchical method with ward linkage illustrates relatively similar results with K-Means and K-Medoids, but several interesting differences are noticeable.

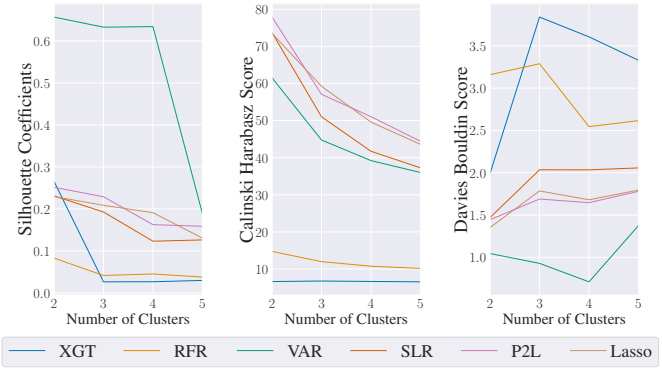

Fig. 8. Agglomerative hierarchical clustering with ward linkage

The clustering of Vector Autoregression coefficients shows a stable high Silhouette Coefficient (SC), decreasing Davies Bouldin (DB) Score, but decreasing Calinski Harabasz (CH) Score over 2, 3 and 4 number of clusters (see "Fig. 8"). This is due to the fact that Agglomerative hierarchical clustering splits the smallest cluster, which contains 8 individuals as shown in "Tab. V", into smaller sub-clusters.

Additionally, clustering XGT features importance show relatively good SC, but the lowest CH for 2 clusters, this occurs due to the creation of imbalanced clusters, as indicated in "Tab. V" and 3D plot of coefficient matrices "Appendix Fig. 10 [Coefficient matrices]" shows that these 6 individuals

TABLE V
RESULTS OF AGGLOMERATIVE CLUSTERING WITH 2 CLUSTERS OVER PREDICTIVE MODEL'S COEFFICIENTS (BASELINE RAW-BASED CLUSTERING INCLUDED).

Model	*SC*	*CH*	*DB*	*CLUS 1*	*CLUS 2*
VAR	0.657	61.516	1.045	179	8
XGT	0.266	6.622	1.994	181	6
P2L	0.252	77.891	1.445	100	87
SLR	0.231	73.703	1.470	105	82
Lasso	0.229	73.487	1.353	120	67
DTW*	0.141	17.283	0.134	101	86
RFR	0.083	14.733	3.159	130	57

* Baseline raw-based clustering method.

are allocated on the edges of the density sphere.

3) Gaussian mixture clustering: Gaussian mixture clustering shows similar average results of cluster evaluations, but with higher standard deviation (refer to "Tab. VI") in comparison to other cluster methods. This is particularly noticeable in the case of VAR coefficients, the increased standard deviation can be attributed to the clustering of coefficient matrices in a single region of high density, as depicted in "Appendix Fig. 17". Moreover, in "Appendix Fig. 17", we can observe how almost all individuals are colored as grey triangles, indicating that individuals in each iteration of Gaussian mixture clustering are assigned to different sub-group of people. Therefore, the final clusters are not calculated in "Tab. VI" due to strong fluctuations in cluster distributions and a high standard deviation in results.

TABLE VI
RESULTS OF GAUSSIAN MIXTURE CLUSTERING WITH 2 CLUSTERS OVER PREDICTIVE MODEL'S COEFFICIENTS (BASELINE RAW-BASED CLUSTERING INCLUDED).

Model	*SC*		*CH*		*DB*	
	μ	σ	μ	σ	μ	σ
VAR	0.463	0.231	29.558	20.217	1.646	1.046
Lasso	0.232	0.042	57.620	21.095	1.457	0.434
P2L	0.231	0.044	56.085	20.904	1.460	0.345
SLR	0.216	0.033	51.220	18.410	1.512	0.281
DTW*	0.141	0.036	17.283	0.03	0.134	0.134
RFR	0.100	0.076	7.786	3.212	3.334	1.038
XGT	0.097	0.085	3.875	1.364	3.590	1.594

* Baseline raw-based clustering method.

C. Cluster Analysis

The Cluster Analysis section covers the clustering results of two model-based approaches: VAR and Lasso with K-Means. Clustering results are analyzed by cluster distributions, coefficient matrices in 3D plot using MMDS, boxplots of feature mean values of individuals and feature correlation matrices within clusters.

1) Vector Autoregression: The clustering of VAR coefficients into 2 clusters with K-means shows stable clustering over iteration and high scores in cluster evaluations.

In "Fig. 11", we can observe cluster distributions and coefficient matrices in three-dimensional representation using MMDS, where each individual data point, which was always stable across iterations, is color-coded according to their assigned cluster. It is evident that there is a high density of individual coefficient matrices assigned to one cluster, indicating that most of the extracted coefficients of individuals have similar coefficients.

Fig. 11. K-Means with VAR coefficients: Cluster distributions of final 2 clusters & Representation of coefficient matrices on three-dimensional(3D) plot

Conversely, individuals allocated above the high-density area are assigned to another cluster. Additionally, one individual, which fluctuated in two clusters over clustering iterations, is labeled by the grey triangle in the three-dimensional plot using MMDS "Fig. 11 [Coefficient matrices]".

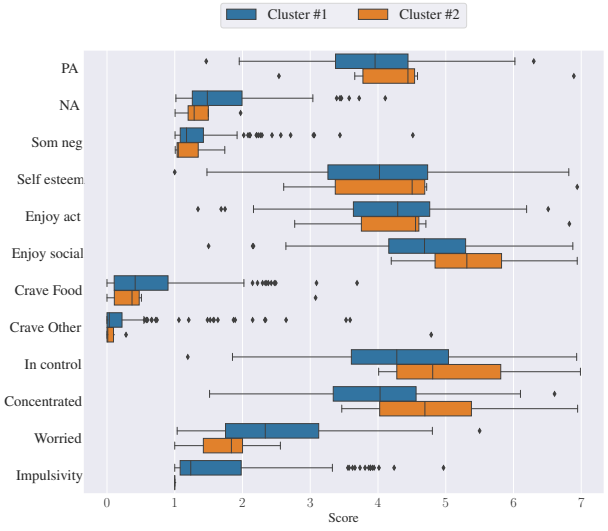

Fig. 12. Boxplots demonstrates each feature mean values of individuals within 2 clusters

"Fig. 12" demonstrates boxplots of feature mean values of individuals within each cluster. We can observe that a boxplot of Impulsivity in the second cluster is a line without box and whiskers, which means that every individual within the second cluster has Impulsivity: 1. Moreover, in the second, individuals have larger mean values for positive features and smaller mean values for negative effects. This suggests that the second cluster consists of more individuals with higher positive effects, a higher level of concentration, self-control, and less negative effects and worried.

On the other hand, the first cluster has individuals, which are more likely to crave food or other. However, individuals from the first cluster are spread out with feature mean values more widely than in the second cluster.

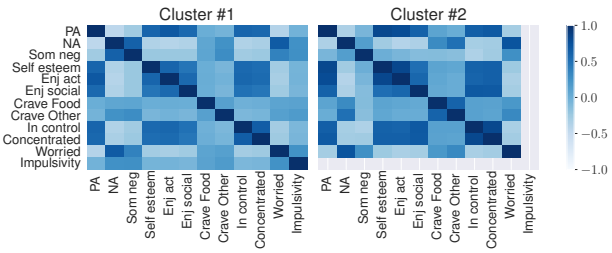

Fig. 13. Mean value of feature VAR coefficients

We also can compare clusters using the correlation matrix of each cluster, which allows us to observe how individuals time-points are correlated with each other (see "Fig. 13"). In the first cluster, we can notice that Somatic negative (Som neg) has a higher correlation with Negative affects (NA) around 0.62 in comparison to the second cluster, where it is 0.14. Conversely, the second cluster demonstrates a higher correlation between In control and Concentrated approximately 0.82 as compared to the first cluster with a correlation 0.65. Additionally, the second cluster illustrates higher correlations between In control−Enjoy social and Concentrated−Enjoy social with values of 0.69 and 0.68 respectively against 0.50 and 0.46 in the first cluster. However, we cannot observe Impulsivity in the second cluster, as it has a constant term.

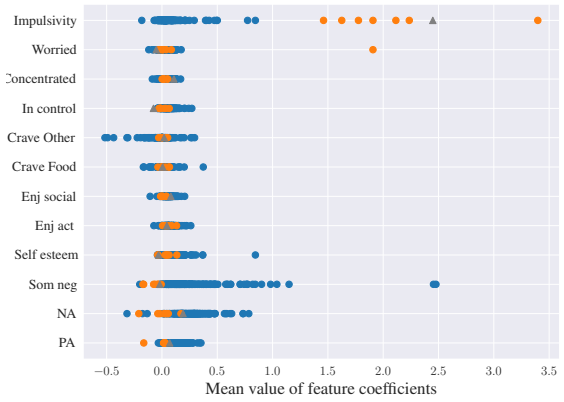

Fig. 14. Scatter of mean values of feature VAR coefficients

The clustering result of VAR K-Means illustrates that the second cluster has higher scores in positive features and less scores in negative features compared to the second cluster. Also, the second cluster of individuals have constant Impulsivity and are less likely to crave. However, it is noticeable that the second cluster contains only 8 individuals, making cluster distributions imbalanced. Therefore, "Fig. 14" demonstrates the mean values of feature coefficients with colors according to the cluster assignment, which can give insights on the clustering assignments of individuals.

In "Fig. 14", we can observe that the mean values of Impulsivity coefficients for 8 individuals are higher than those of individuals from the first cluster. This can suggest that the second cluster has a constant value in the Impulsivity column, which can be considered as an intercept for certain features in the model.

2) Linear Lasso Regression : The clustering of Linear Lasso Regression (Lasso) coefficients demonstrates a well-balanced distribution of individuals across various clustering methods and ranks second place after VAR in clustering evaluations. Therefore, K-Means clustering into 2 clusters based on Lasso coefficients is selected for further analysis.

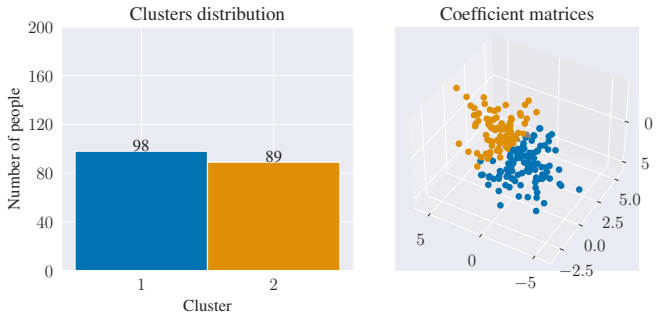

Fig. 15. K-Means with Lasso coefficients: Clusters distribution of final 2 clusters & Representation of coefficient matrices on three-dimensional(3D) plot

In "Fig. 15", we can observe how the K-means technique separates individual coefficient matrices into two groups of almost equal size, with a line down the center. Moreover, we do not observe any fluctuations in individuals across clusters, which indicates stable clustering across different initializations.

Individuals in the first cluster demonstrate a higher prevalence of negative effects and lower scores in Self-Esteem, Enjoy action & social, In control, and Concentrated (see "Fig. 16"). Conversely, individuals in the second cluster show an opposite pattern compared to the first cluster. Furthermore, there is a clear distinction between the clusters based on the variance of individual features.

The correlation matrices of clusters "Fig. 17" do not illustrate any significant difference in features correlations among clusters. However, within the first cluster, there is a higher correlation between Impulsivity and other features compared

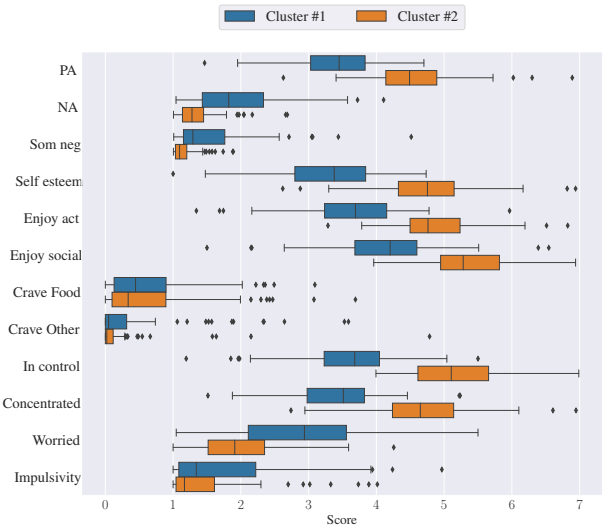

Fig. 16. Boxplots demonstrates each feature mean value of individuals within 2 clusters

to the second cluster. Moreover, in the first cluster, Somatic negative (Som neg) has a higher correlation with Negative affects (NA) around 0.62 in comparison to the second cluster, where the value is 0.42.

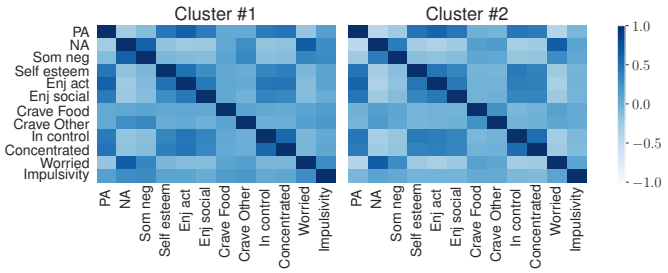

Fig. 17. K-Means with Lasso coefficients: Features correlations within 2 clusters

The clustering results of Lasso K-means demonstrate a well-balanced separation of individuals into two distinct groups. The first cluster contains individuals with fewer values in positive features, In control and Concentrated and higher values in negative features and Worried, Impulsivity features in comparison with the second cluster individuals. Therefore, this implies that one group exhibits a higher frequency of positive signals in their natural environment, while the other group demonstrates a greater occurrence of negative signals.

V. Discussion

In this study, our objective was to investigate different model-based approaches for clustering multivariate Ecological Momentary Assessment (EMA) time series and compare their

results with each other. We assessed these approaches using various criteria to gain insights into their effectiveness in handling complex data structures and performing clustering. We introduced seven predictive models for extracting coefficient matrices that represent the behavioral patterns of individuals and four clustering methods to group these matrices into distinct clusters. Combinations of predictive models and clustering methods can give interesting results in clustering.

Decision tree models, such as XGBoost Tree (XGT) and Random Forest Regression (RFR), demonstrated good performance in predicting individual data points. However, their coefficient matrices tend to under-perform in clustering assignments. This limitation may arise from the reliance on feature importance measures, which fail to highlight the significance of certain features and lead to insignificant differences among the feature importance values.

Polynomial Regression revealed RMSE μ close to zero during training evaluation, indicating the successful recreation of individuals with high accuracy. However, when applied to testing and cross-validation time-series data, it demonstrated a limited ability to predict the future states of individuals. Additionally, increasing the polynomial degree leads to strong overfitting and increasing the number of coefficients, which makes it harder to identify similarities between individuals due to high dimensionality.

Lasso regularization with a statistical model can be used as a feature selector for individuals, enabling shrinkage of certain features for better capturing individual patterns without noise. Prediction with Lasso regularization can positively influence the accuracy of predicting individual time-points and demonstrate better handling prediction with smaller data sizes compared to the standard version of the model. Moreover, feature selection can force clustering methods to prioritize relevant features of individuals and perform clustering based on selected features only.

Vector Autoregression (VAR) has shown a strong performance in forecasting individual time-points for all three evaluations compared to other predictive linear models. Furthermore, VAR is more flexible in dealing with multivariate time-series data due to the ability to consider time-steps and characteristics of the data.

Gaussian Mixture clustering (GMM) showed significant variance in cluster assignment across iterations, indicating difficulties in accurately separating coefficient matrices into distinct clusters. This inconsistency leads to different cluster distributions across iterations. Consequently, we can infer that GMM lacks robustness in stable clustering assignments.

K-Means, K-Medoids, and Agglomerative Hierarchical Clustering produced almost similar results in clustering coefficient matrices of models, with minor standard deviations across clustering results with the different k clusters. Therefore, these clustering methods can be considered reliable in effectively capturing similar variances in the clusters of coefficient matrices.

In this study, K-Means method with Linear Lasso Regression (Lasso) and VAR coefficients was taken for cluster analysis, but it is important to acknowledge the relevance of other clustering methods, such as K-Medoids and Agglomerative Hierarchical Clustering with the ward linkage, that have shown intriguing findings in cluster distributions.

Lasso K-Means with 2-k clusters "Section IV-C2" showed relatively a well-balanced separation of individuals into two clusters, where one cluster exhibits a higher prevalence of positive signals and another demonstrated a higher number of individuals with negative signals. Additionally, it was performed Lasso K-Means with 3-k clusters (See "Appendix B1"), where we can observe how additional cluster borrows almost equal data-points of the coefficient matrix from the initial two clusters, making the third cluster and second cluster are more separate in terms of positive and negative signals among individuals. Moreover, the first cluster assumed an intermediate position between these clusters. Nevertheless, there is still can be individuals in the first cluster, who more likely to relate to the different clusters, thereby indicating imperfection in the assignment of individuals to clusters.

In comparison, VAR K-Means can identify outlier individuals who do not closely resemble the main group and assign them to distinct clusters. In Cluster Analysis "Section IV-C1", K-Means with VAR coefficients demonstrated strong separation in favor of high impulsivity coefficients, indicating that VAR model may be more sensitive to the constant term rather than effectively capturing meaningful patterns in individuals.

Basically, assignment into 2-k clusters is an optimal value for clustering individuals based on their characteristic, where one cluster represents a high number of positive signals, while another demonstrates a high number of occurrence of negative signals. Consequently, the utilization of 2-k clusters enables the identification of two groups with distinct behavioral patterns.

VI. Conclusion

Our study explored the utilization of different machine-learning models for extracting information for clustering multivariate EMA time-series data. While our focus was primarily on linear models that utilized linear parameter estimates to create coefficient matrices representing individual patterns, we also explored decision models that employed feature importance instead of linear parameters as coefficients. The main idea behind extracting information from machine learning models is to capture both the stationary and non-stationary effects of an individual. In our study, the extracted coefficients described predominantly stationary (regular) effects with non-stationary noise. Therefore, the choice of different model-based clustering methods can affect the performance.

Multivariate EMA time series involves specific challenges from missing data points to hundreds of data points for every individual. Furthermore, model-based clustering needs to handle all these cases and highlights special attributes of individuals. Linear Lasso Regression (Lasso) with K-Means can cover the basics of these challenges and perform relatively

good clustering. On the other hand, Vector Autoregression (VAR) can cover basic, consider time steps, and can be modified to handle special cases in data.

In conclusion, by utilizing model-based clustering for multivariate EMA time-series data, researchers can gain deeper insights into individual behavior and it can lead to a better understanding of human actions, patterns, and decision-making processes in various contexts such as health, psychology, and social sciences.

VII. Future Work

Support Vector Regression (SVR) or probabilistic models, such as Hidden Markov Model (HMM) also can be considered for extracting information. EMA data has a complex data structure and psychologist typically analyze such data with linear mixed-effect models, which contains fixed and random effects [25]. Fixed effect refers to factors that are unchanged within the cluster and random effect are factors that are assumed to fluctuate randomly from one cluster to another.

SVR provides us with the ability to specify the level of tolerance for error in our model and create a linear function with linear estimators [26], where linear estimators/coefficients can be considered as a fixed effect and the level of tolerance for error as a random effect.

HMM is a probabilistic model of Markov chain, where each state of the chain independently can produce emissions according to emission probabilities or densities [27]. HMM can be fitted with individual time-points, creating a model, which can describe the behavior pattern of an individual. Then, the obtained HMMs can be compared using a distance measure, which is calculated using the forward-backward algorithm [27], where the likelihood can be used to define the distances [28].

References

[1] P.-Y. Zhou and K. C. C. Chan, "A model-based multivariate time series clustering algorithm," in *Lecture Notes in Computer Science*, ser. Lecture notes in computer science. Cham: Springer International Publishing, 2014, pp. 805–817.

[2] D. Tiano, A. Bonifati, and R. Ng, "FeatTS: Feature-based time series clustering," in *Proceedings of the 2021 International Conference on Management of Data*. New York, NY, USA: ACM, Jun. 2021.

[3] P. Muhammad Ali and R. Faraj, *Data Normalization and Standardization: A Technical Report*, 01 2014.

[4] E. Poitras, K. R. Butcher, and M. P. Orr, "Modeling interactive behaviors while learning with digitized objects in virtual reality environments," in *Cognitive and Affective Perspectives on Immersive Technology in Education*. IGI Global, May 2020, pp. 215–234.

[5] MatLab. (2023) Linear model. [Online]. Available: https://www.mathworks.com/discovery/linear-model.html

[6] Yale. (2023) Linear regression. [Online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[7] Wikipedia. (2023) Polynomial regression. [Online]. Available: https://en.wikipedia.org/wiki/Lasso_(statistics)

[8] D. E. Hilt, D. W. Seegrist, Northeastern Forest Experiment Station (Radnor, Pa.), and United States., *Ridge, a computer program for calculating ridge regression estimates /*. Upper Darby, Pa :: Dept. of Agriculture, Forest Service, Northeastern Forest Experiment Station,, 1977.

[9] Wikipedia. (2023) Polynomial regression. [Online]. Available: https://en.wikipedia.org/wiki/Polynomial_regression

[10] Analyticsvidhya. (2023) Master polynomial regression with easy-to-follow tutorials. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/07/all-you-need-to-know-about-polynomial-regression/

[11] Wikipedia. (2023) Vector autoregression. [Online]. Available: https://en.wikipedia.org/wiki/Vector_autoregression

[12] L. van der Krieke, A. C. Emerencia, E. H. Bos, J. G. Rosmalen, H. Riese, M. Aiello, S. Sytema, and P. de Jonge, "Ecological momentary assessments and automated time series analysis to promote tailored health care: A proof-of-principle study," *JMIR Res. Protoc.*, vol. 4, no. 3, p. e100, Aug. 2015.

[13] A. F. Ernst, "Dynamic clustering," Ph.D. dissertation, 2021.

[14] Qwak. (2023) Random forest regression. [Online]. Available: https://www.qwak.com/post/xgboost-versus-random-forest

[15] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," vol. 8, 01 2007, pp. 1027–1035.

[16] X. Jin and J. Han, *K-Medoids Clustering*. Boston, MA: Springer US, 2017, pp. 697–700.

[17] J. Song. (2021) Agglomerative clustering, model item, opengms. [Online]. Available: https://geomodeling.njnu.edu.cn/modelItem/1149cf58-cf96-4682-b2ee-ef95d41253b7

[18] C. Maklin. (2023) Gaussian mixture models clustering algorithm explained. [Online]. Available: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

[19] D. Reynolds, *Gaussian Mixture Models*. Boston, MA: Springer US, 2009, pp. 659–663. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_196

[20] A. Amidon. (2023) How to apply k-means clustering to time series data. [Online]. Available: https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3

[21] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[22] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

[23] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[24] "Metric and nonmetric MDS," in *Springer Series in Statistics*. New York, NY: Springer New York, 2007, pp. 199–225.

[25] J. Dora, C. McCabe, C. J. Van Lissa, K. Witkiewitz, and K. M. King, "Analyzing ecological momentary assessment data in psychological research using bayesian statistics: A tutorial," Nov. 2022.

[26] T. Sharp. (2023) Support vector regression. [Online]. Available: https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

[27] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[28] S. Ghassempour, F. Girosi, and A. Maeder, "Clustering multivariate time series using hidden markov models," *Int. J. Environ. Res. Public Health*, vol. 11, no. 3, pp. 2741–2763, Mar. 2014.

Appendix

A. Clustering methods

"Appendix Fig. 7, 9" demonstrates the average values of cluster evaluations over 100 iterations with different numbers of clusters. All clustering methods "Section II-G" show a gradual decrease in quality "Section III-C1" as the number of clusters increases. Additionally, Linear Regression and Polynomial Regression with Lasso Regularization (Lasso & P2L respectively) produce similar results, which could be related to equal feature shrinkage.

Figures also demonstrate interesting observation that all clustering methods illustrate high clustering performance with

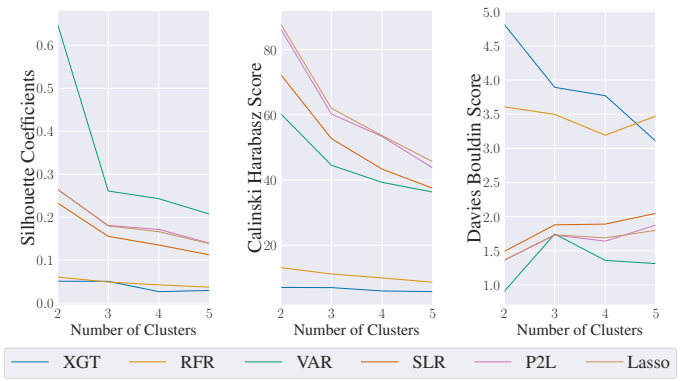

Fig. 7. K-Medoids clustering

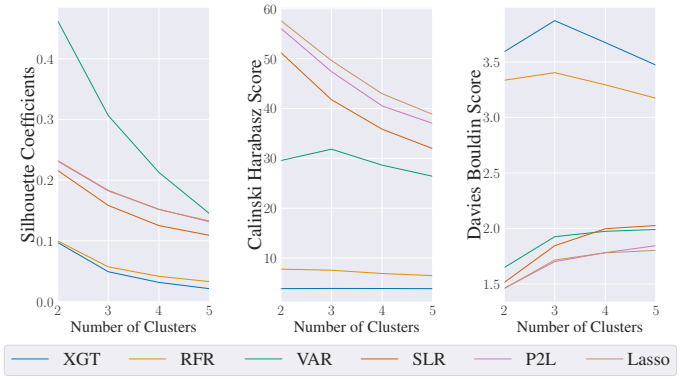

Fig. 9. Gaussian mixture clustering

2 clusters. Consequently, the results of the 2 clusters are presented in "Tab. III, V, VI" and " Appendix Tab. IV"

TABLE IV
RESULTS OF K-MEDOIDS WITH 2 CLUSTERS OVER PREDICTIVE MODEL'S COEFFICIENTS (BASELINE RAW-BASED CLUSTERING INCLUDED).

Model	SC		CH		DB		CLUS 1	CLUS 2
	μ	σ	μ	σ	μ	σ		
VAR	0.649	0‘	60.281	0‘	0.902	0‘	180	7
P2L	0.265	0‘	86.268	0	1.364	0‘	95	92
Lasso	0.264	0‘	87.779	0‘	1.358	0‘	96	91
SLR	0.233	0.004	72.293	0.465	1.489	0.027	111	76
DTW*	0.141	0.036	17.283	0.03	0.134	0.134	101	86
RFR	0.061	0.004	13.149	1.540	3.608	0.260	89	98
XGT	0.051	0‘	7.081	0‘	4.818	0‘	111	76

* Baseline raw-based clustering method.
‘ value is close to zero.

1) Agglomerative hierarchical clustering: Additionally, clustering XGT features importance show relatively good SC, but the lowest CH for 2 clusters, this occurs due to the creation of imbalanced clusters, as indicated in "Tab. V" and 3D plot of coefficient matrices "Appendix Fig. 10 [Coefficient matrices]" show that these 6 individuals are allocated on the edges of density sphere

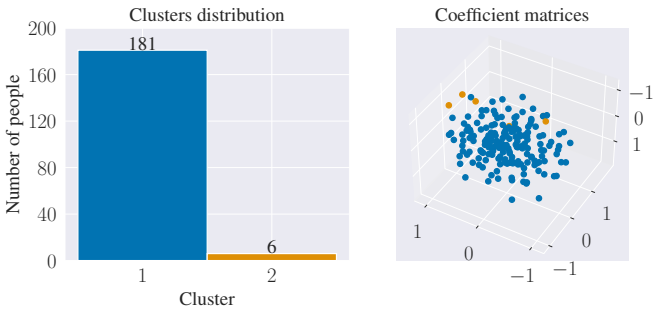

Fig. 10. Agglomerative hierarchical clustering with XGT coefficients: Clusters distribution of final 2 clusters & Representation of coefficient matrices on three-dimensional(3D) plot

2) Gaussian Mixture clustering: Moreover, in "Appendix Fig. 17", we can observe how almost all individuals are colored as grey triangles, indicating that individuals in each iteration of Gaussian mixture clustering are assigned to different sub-group of people. Therefore, the final clusters are not calculated in "Tab. VI" due to strong fluctuations in cluster distributions and a high standard deviation in results.

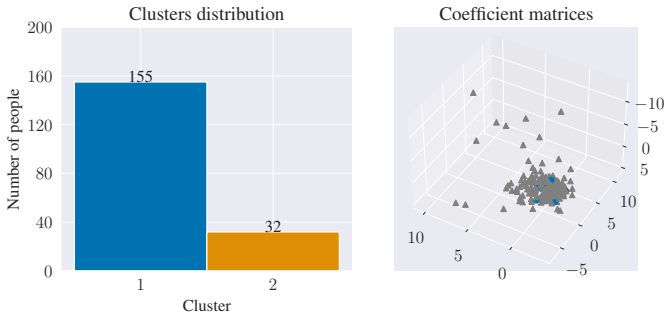

Fig. 17. GMM with VAR coefficients: Clusters distribution of final 2 clusters & Representation of coefficient matrices on three-dimensional(3D) plot

B. Cluster Analysis

1) Lasso K-Means with 3 clusters : Lasso K-Means with 2 clusters demonstrated relatively good results in clustering into 2 distinct clusters, showing the difference among these clusters. Therefore, Lasso K-Means with 3 clusters is taken in order to have an understanding of the nature of 3 clusters assignments.

"Fig. 18" demonstrates cluster distributions and coefficient matrices in 3D plot using MMDS, where each color corresponds to the cluster assignment. It can be observed that the creation of a third cluster is achieved by dividing two main clusters discussed in "Section IV-C2". Additionally, it is noticeable that there is a slight fluctuation of individuals between clusters, depicted as grey triangles around the center in "Fig. 18 [Coefficient matrices]"

"Fig. 19" is boxplots of each feature mean values of individuals within 3 clusters, thereby we can observe how

Fig. 18. K-Means with Lasso coefficients: Clusters distribution of final 3 clusters & Representation of coefficient matrices on three-dimensional(3D) plot

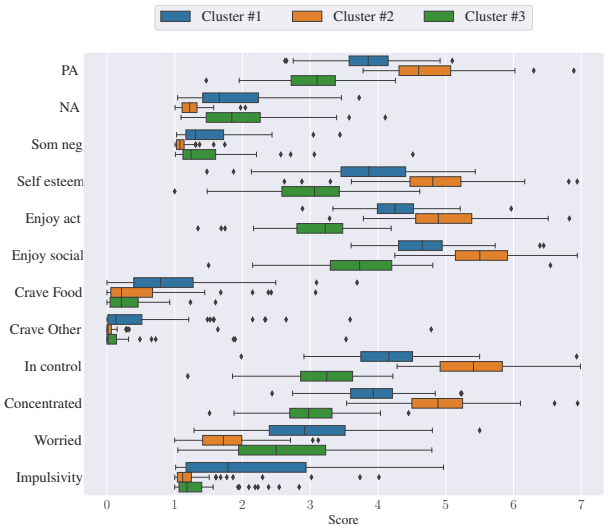

Fig. 19. Boxplots demonstrates each feature mean value of individuals within 3 clusters

third cluster collected outliers of two main clusters in six features: Positive affects (PA), Self esteem, Enjoy action & social, In control and Concentrated, leaving individuals with average mean values in these features for the first cluster.

Generally, by removing individuals of the first cluster, the second and third clusters become more separated from each other in terms of positive and negative signals of individuals. Moreover, individuals from the second cluster show less overlap in mean features (except for Impulsivity, Crave Other & Food) with individuals in the third cluster. Therefore, this can be indicated as an improvement in clustering assignment, suggesting that the first cluster can be considered as a middle cluster between the second and third clusters. However, there is still can be individuals in the first cluster, who more likely to relate to the different clusters, thereby making the clusters imperfect in their assignments.