Maynooth University
National University of Ireland Maynooth

MSc. Data Science and Analytics Thesis

Regularised Model-based Clustering of Educational Engagement Measures

Author: Saad Siddiqui

Student Number: 23250068

Supervisor: Dr. Keefe Murphy

A thesis submitted in fulfilment of the requirements for the degree of Masters in Data Science and Analytics 2023-2024

to the

Department of Mathematics & Statistics

Maynooth University

9th August 2024

Acknowledgements

First and foremost, I want to thank Allah for guiding and blessing me throughout this journey. Your guidance has been a source of strength and inspiration.

I would like to express my deepest gratitude to my supervisor, Dr. Keefe Murphy. Your unwavering support, patience, and invaluable guidance have been the cornerstones of this journey. Your belief in my abilities and your constant encouragement have made all the difference. Thank you for being such an incredible mentor and for always being there for me.

I would also like to extend my sincere appreciation to Maynooth University. The knowledge, resources, and community you have provided have been instrumental in my academic journey. Thank you for creating an environment that nurtures learning and growth.

To my uncle and aunt, Dr. Sadiq and Dr. Seher, words cannot express how grateful I am for your endless support. Your encouragement and belief in my dreams have been a driving force behind my achievements. I am truly indebted to you both for everything you have done to support my education and my goals. Thank you from the bottom of my heart.

A special thanks to my amazing parents, Mama and Baba, for always rooting for me and providing unwavering support. Your love and encouragement have been my constant source of strength. I couldn't have done this without you.

To my best friend, Sufyan, thank you for always being there for me. Your friendship, support, and late-night study sessions have been invaluable. You made this journey so much more enjoyable and manageable. Thank you for being my rock.

Abstract

This thesis employs model-based clustering using finite Gaussian mixture models to analyse student engagement and academic performance data from a primary school in northern Spain. In addition to using the constrained covariance matrix parameterisations afforded by the family of Gaussian parsimonious clustering models, the simultaneous employment of prior distributions for regularisation and incorporation of a noise component to capture outliers were found to lead to a superior model compared to previous applications of latent profile analysis to these data. Three distinct clusters of student engagement patterns were found, corresponding to low, medium, and high engagement. Bootstrapping was used to quantify the uncertainty in the estimated engagement profiles and the entropy and average posterior probability measures were used to assess the quality of the uncovered partition of students. The interpretability of the resulting clusters, as well as their demonstrated relationship to covariates related to academic achievement, demographics, and self-regulation, provide valuable insights for educators and policymakers. Indeed, our findings could be used to inform tailored interventions and guide the provision of educational supports.

Keywords: Educational engagement, finite Gaussian mixture models, latent profile analysis, regularising prior distributions, uniform noise component.

Contents

List of tables

List of figures

1 Introduction

In educational research, traditional variable-centered methodologies, such as correlation, regression, and comparisons of means (e.g., t-tests), are commonly utilised. These methods typically depend on a representative sample drawn from the population to calculate or compare central tendency measures, such as the mean or median. These measures are assumed to accurately reflect the characteristics of the entire population, thereby enabling generalisations. However, this approach overlooks the inherent individual differences prevalent across all domains of human behavior and function. Learners demonstrate considerable variability in their behaviors, attitudes, and dispositions, seldom conforming to a singular, common pattern or average behavior. Consequently, employing an "average" to represent the diverse nature of learners constitutes an oversimplification.

Given the significant heterogeneity among learners, there has been an increasing interest in methodologies that capture individual differences, focusing on the patterns and variations among students. These methodologies are known as person-centered methods. Person-centered methods can be broadly categorised into two types: heuristic, dissimilarity-based clustering algorithms (e.g., agglomerative hierarchical clustering and partitional clustering algorithms such as k-means) and model-based clustering (MBC) approaches (e.g., finite Gaussian mixture models). This thesis focuses on the MBC paradigm, emphasising that unlike variable-centered methods, person-centered methods aim to capture heterogeneity by identifying latent (unobserved or hidden) patterns within the data. These patterns form subgroups or "clusters" that are assumed to be homogeneous. The MBC framework represents a probabilistic approach to statistical unsupervised learning, aiming to discover clusters of observations within a dataset. The MBC paradigm is termed model-based because it is grounded in a generative probabilistic model, in contrast to heuristic clustering algorithms that rely on dissimilarity criteria.

In this thesis, we analyse a dataset concerning school engagement, academic achievement, and self-regulated learning pertaining to students from a primary school in northern Spain (Estévez et al. 2021). These data were previously modelled using one of the person-centered methods; namely, latent profile analysis (LPA). While the term "latent profile analysis" is commonly used in educational and social sciences literature, the term "finite Gaussian mixture models" is more prevalent in the statistical literature. Indeed, the term LPA is exactly equivalent to an unconstrained finite Gaussian mixture model. In light of this, we extend the previous analysis by incorporating constraints and utilising Gaussian parsimonious clustering models as implemented in the popular package `mclust` Scrucca et al. (2016) for the statistical computing environment `R` (R Core Team 2024). Although mixture models and the `mclust` package have garnered attention in social sciences, their adoption in educational research has been relatively slow. Furthermore, their application in learning analytics research has been limited.

A critical aspect of this research involves examining the parameters of GPCM models to identify the best-fitting model. This involves using the Bayesian Information Criterion (BIC) to select the optimal number of components and the appropriate covariance type. The EM (Expectation-Maximisation) algorithm plays a crucial role in estimating the parameters of the mixture model, by iteratively maximising the

complete log-likelihood. Subsequently, we explore some advanced functionalities of the `mclust` package, by adding a uniform noise component to capture outliers whose engagement patterns deviate from the more defined profiles and by incorporating prior distributions to achieve regularisation, such that the EM algorithm conducts maximum *a posteriori* estimation rather than maximum likelihood estimation

Through this process, the research aims to produce estimates of the parameters and posterior probabilities of component membership, thereby achieving a robust and accurate clustering partition capable of characterising the latent engagement profiles in an interpretable manner. Specifically, we construct clusters using educational engagement measures and then subsequently relate the uncovered latent profiles to covariates related to academic achievement and self-regularisation. Our findings contribute to the existing literature on student engagement by offering a detailed characterisation of engagement profiles and their association with academic performance and other factors. By identifying groups of students with similar engagement patterns, educators can tailor their strategies to address the specific needs of each group, ultimately enhancing educational outcomes. Furthermore, this research highlights the utility of model-based clustering as a valuable tool in educational research, promoting its application in future studies.

The remainder of this thesis is organised as follows. The background to the data along with some exploratory data analysis is provided in Section 2. The theory underpinning finite Gaussian mixture modelling is described in Section 3, along with some of the advanced modelling extensions we explored. Preliminary clustering results are presented in Section 4 using standard applications of Gaussian parsimonious clustering models, along with more robust results of a richer analysis based on advanced clustering methodologies which exploit add-on features of `mclust`. A detailed appraisal of the profiles uncovered by the optimal model are then provided in Section 5, along with results attempting to relate the corresponding clusters to covariates. Finally, the thesis concludes with a brief discussion of limitations and future research ideas in Section 6.

2 Background to the data

In the realm of learning analytics, understanding the factors which contribute to student success is of paramount importance. A recent study conducted with a group of $n = 717$ primary school students from northern Spain aimed to characterise school engagement and explore its association with academic performance and self-regulation (Estévez et al. 2021). The study Fredricks et al. (2005) employed the School Engagement Measure (SEM) to assess the students' behavioral, cognitive, and emotional engagement, while their self-regulation was evaluated using the self-regulation strategy Inventory–Self-Report (Cleary 2006). Academic achievement was measured based on the students' self-reported grades in Spanish and Mathematics, which were rated on a Likert scale of 1 to 5.

Existing research has highlighted the critical role of self-regulation in influencing academic performance (Azevedo 2015). Self-regulation encompasses various cognitive, metacognitive, motivational, and emotional aspects that are crucial for students' academic success. Studies have shown that adolescents and

young adults who exhibit stronger self-regulation tend to perform better academically, as they are better equipped to manage their learning, stay motivated, and regulate their emotions.

The primary objective of this study was to identify distinct profiles or patterns of student engagement and then investigate how these profiles relate to academic performance and self-regulation. By examining the relationship between engagement profiles, self-regulation, and academic achievement, the current study aims to shed light on the individual differences that contribute to student outcomes, which can inform educational practices and identify intervention strategies aimed at supporting the success of students with varying engagement patterns (Olivier et al. 2020).

The data set for this study is available on a remote GitHub repository[1], allowing for further exploration and analysis by the educational research community.

2.1 Exploratory data analysis

We begin by describing the engagement variables that will be clustered and subsequently discuss the additional self-regulation and academic achievement variables which we treat as covariates which may or may not be associated with the uncovered clusters.

2.1.1 Engagement variables

We have selected three key engagement variables for analysis: behavioral, cognitive, and emotional engagement. Table 1 presents the relevant summary statistics for these variables. It is important to note that each variable is measured on a continuous scale ranging from 1 (indicating 'never engaged') to 5 (indicating 'always engaged'). The summary statistics reveal that, on average, behavioral engagement is the highest, while emotional engagement exhibits the greatest variability within the sample. To further support our exploratory data analysis of these three dimensions of engagement, we present boxplots in Figure 2 and a generalised pairs plot in Figure 1.

The data characteristics further enhance our understanding of student engagement. Table 1 highlights the number of unique engagement scores for each dimension, with behavioral engagement having only 17 unique values, cognitive engagement having only 30, and emotional engagement having only 22. These number of unique values are worth noting as they have implications for our later clustering analysis, specifically in relation to the need to incorporate prior distributions to impose regularisation.

[1] https://github.com/lamethods/data/tree/main/3_engSRLach

Table 1: The number of unique values (of a sample of $n = 717$ students in total), mean, standard deviation (SD), minimum (Min), median, and maximum (Max) for the three engagement variables: behavioral, cognitive, and emotional. Each variable is measured on a continuous scale from 1 (never engaged) to 5 (always engaged).

Variable	No. Unique	Mean	SD	Min	Median	Max
Behavioural	17	4.17	0.63	1	4.25	5
Cognitive	30	2.92	0.77	1	2.92	5
Emotional	22	3.61	0.91	1	3.61	5

We now discuss the findings of both Figure 2 and Figure 1 to provide a comprehensive overview of how these engagement aspects vary among students, highlighting key statistical relationships and features.

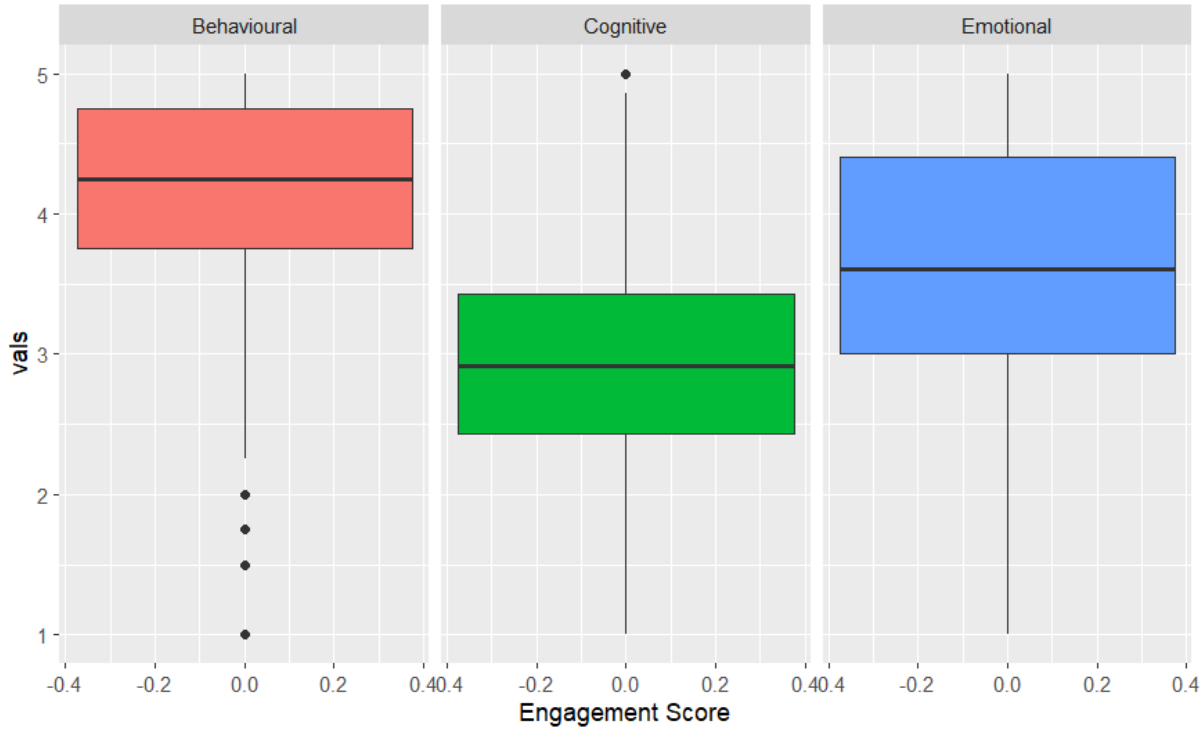


Figure 1: Box plot of the behavioural, cognitive, and emotional engagement scores, indicating the spread of scores along the Likert scale.

Behavioral engagement stands out with the highest average engagement score among the three dimensions, with a mean of 4.17. The boxplot also indicates the presence of a small number of outlying students with exceptionally low behavioral engagement scores. Indeed, in Figure 1, the associated density plot in the top-left panel reveals a right-skewed distribution, suggesting that while most students exhibit high behavioral engagement, there are a few with significantly lower scores.

Cognitive engagement, in contrast, has the lowest mean score of 2.92. The variability in cognitive engagement is moderate, with a standard deviation of 0.77, indicating a relatively narrow spread of scores around the mean. The density plot for cognitive engagement shows a near-normal distribution with slight skewness, implying that the scores are fairly balanced, with few extreme values.

Emotional engagement is characterised by the highest variability among the three dimensions, with a standard deviation of 0.91. This high variability indicates significant differences in emotional engagement among students. The mean score for emotional engagement is 3.61, which is higher than cognitive engagement but lower than behavioral engagement. The density plot for emotional engagement shows a left-skewed distribution, suggesting a larger number of students with high emotional engagement scores, but also a notable number of outliers with lower scores.

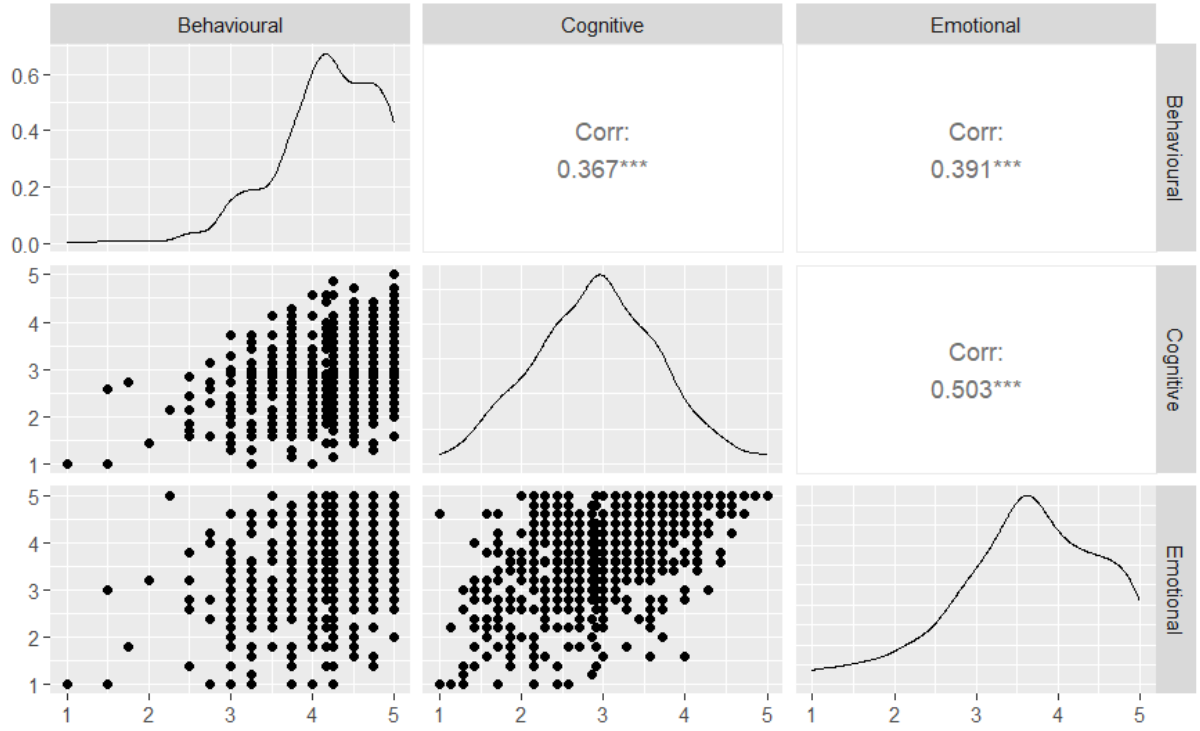


Figure 2: Generalised pairs plot of the behavioural, cognitive, and emotional engagement variables, showing pairwise scatterplots and correlation information on the off-diagonal panels and kernel density estimates along the diagonal panels.

The correlation matrix provied as part of Figure 2 gives additional insights into the relationships between the three engagement dimensions. There is a moderately strong positive correlation of 0.367 between behavioral and cognitive engagement, suggesting that students who are behaviorally engaged tend also to be cognitively engaged. The correlation between behavioral and emotional engagement is similarly moderately strong at 0.391, indicating a positive relationship between these two dimensions. The strongest correlation, at 0.503, is observed between cognitive and emotional engagement, implying that cognitive and emotional aspects of engagement are more closely related compared to other pairs. It is notable that all three sample correlations are positive.

Overall, the exploratory data analysis provides a detailed understanding of the engagement levels among students, highlighting critical areas for further investigation and potential intervention to improve student engagement across different dimensions. Indeed, several key insights emerge from this analysis. Behavioral engagement stands out as the highest on average, indicating that students are more consistently engaged behaviorally compared to cognitively and emotionally. The high variability in emotional engagement

suggests diverse emotional responses among students, which could be influenced by personal factors or external circumstances. The moderately strong correlations among all three dimensions underscore the interconnected nature of student engagement, where enhancements in one dimension could positively affect the others. Additionally, the analysis points out a large number of duplicated engagement scores, which may indicate a common pattern or similar engagement levels among many students.

2.1.2 Self-regulation and academic performance

We now present a series of plots to explore the relationship between self-regulation and academic performance among students. Understanding these relationships is crucial, as self-regulation is a significant predictor of academic success. The following plots will help us visualise key patterns and statistical relationships between these variables, providing insights into how different aspects of self-regulation impact academic outcomes.

Self-regulation is measured through various dimensions, including goal setting, time management, and self-monitoring. Academic performance, on the other hand, is typically assessed through grades, test scores, and overall GPA. By analysing these dimensions in conjunction, we aim to identify trends and correlations that might inform educational strategies and interventions.

The bar plots in Figure 3 and Figure 4 illustrate the proportion of students achieving different grades in Mathematics and Spanish, stratified by gender. These visualisations provide insights into the performance distribution across males and females in these subjects.

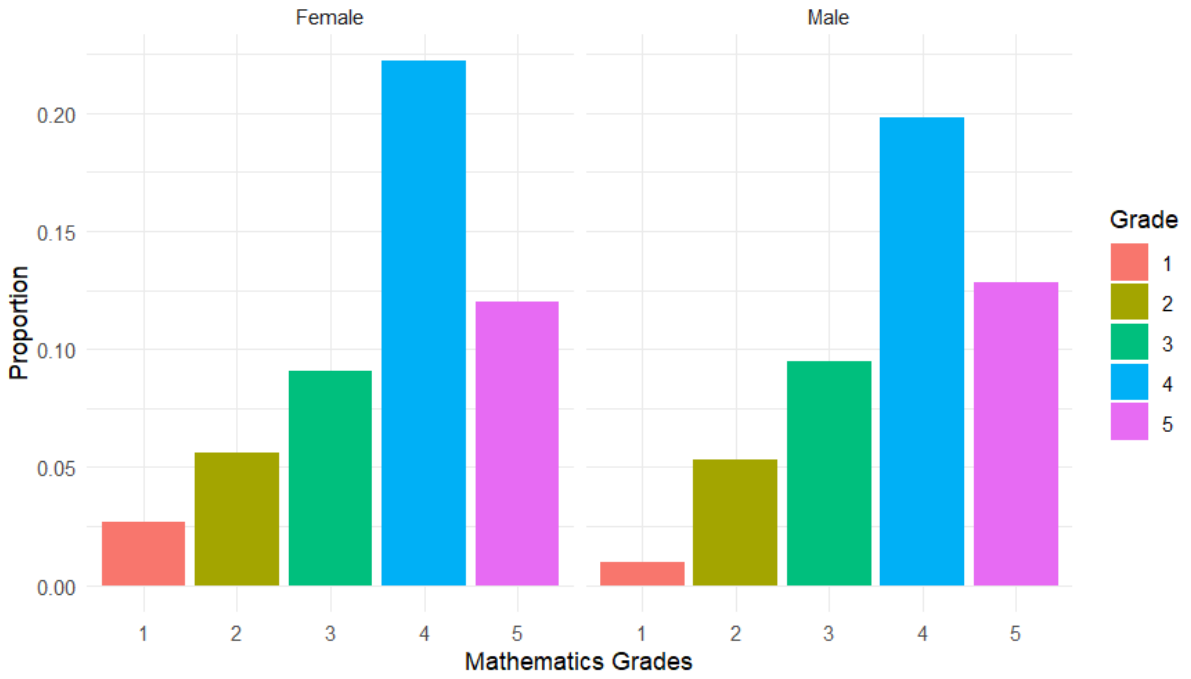


Figure 3: Bar plot illustrating the distribution of Mathematics grades among male and female students, highlighting the differences in performance across genders.

In the Mathematics bar plot in Figure 3, we observe that a significant proportion of female students achieved a grade of 4, making it the most common grade among females, followed by grade 5. This indicates a concentration of female students in the higher performance tiers, suggesting strong mathematical capabilities within this group. Conversely, male students also predominantly achieved a grade of 4, but unlike the females, the second most common grade for males was 5, closely followed by grade 3. This indicates that both genders have a strong showing in grade 4, but females tend to perform slightly better overall in Mathematics, as evidenced by the higher proportion achieving grade 5.

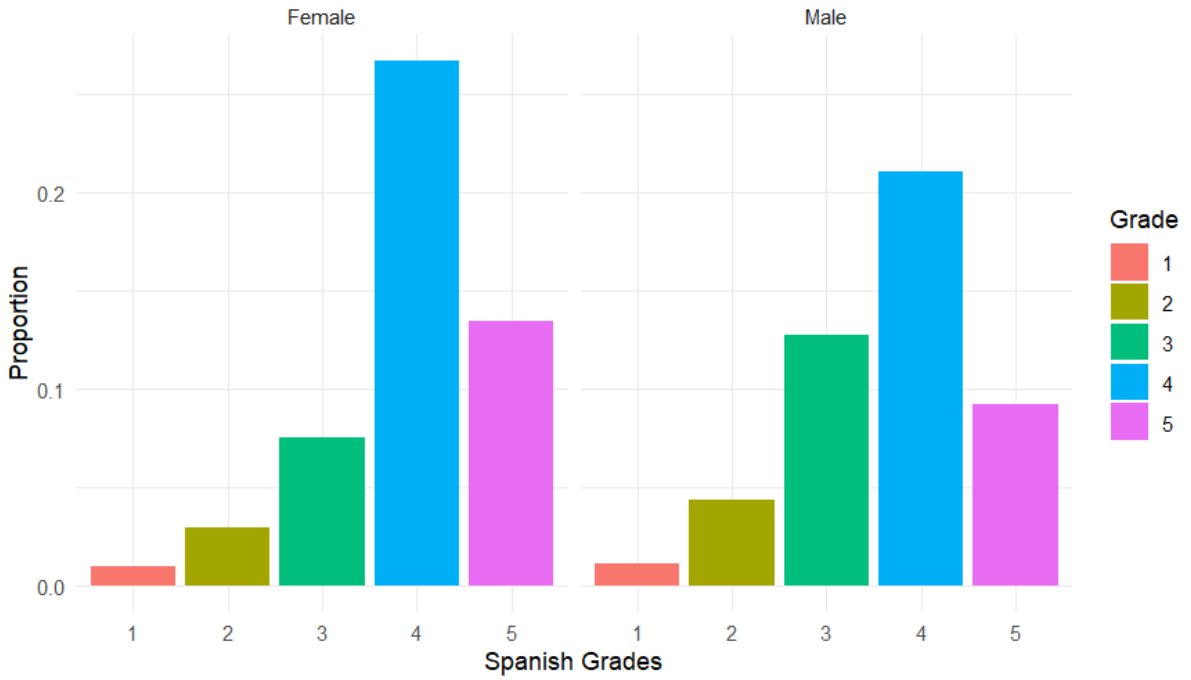


Figure 4: Bar plot illustrating the distribution of Spanish grades among male and female students, highlighting the differences in performance across genders.

In the Spanish grade distribution plot in Figure 4, the performance distribution shows a similar pattern. The majority of female students achieved a grade of 4, followed by grade 5, indicating high proficiency in Spanish among females. This distribution mirrors the pattern observed in Mathematics, reinforcing the trend of females excelling in academic performance. On the other hand, male students showed a significant proportion achieving grade 4, followed by grades 3 and 5. The proportion of males achieving grade 3 was higher in Spanish than in Mathematics, suggesting that males exhibit more variability in their Spanish performance compared to Mathematics.

In both subjects, the predominant grades among females were 4 and 5, highlighting their academic strengths. For males, the performance was more varied, with grades 3 and 4 being the most common in Spanish, and grades 4 and 5 being the most common in Mathematics. This analysis indicates that while females tend to cluster at the higher end of the performance spectrum, males display a wider range of performance outcomes in both Mathematics and Spanish. These insights could inform targeted educational interventions to support students based on their specific needs and performance trends.

The findings presented in the correlation heatmap offer valuable insights into the factors influencing academic performance. These findings can inform the development of targeted academic support programs and interventions aimed at enhancing self-regulation skills and creating optimal learning environments.

The Kendall correlation plot in Figure 5 illustrates the pairwise relationships between various variables, including Mathematics grades, Spanish grades, Environment management, Time management, Maladaptive behavior, and Information management. Each cell in the plot represents the Kendall correlation coefficient between two variables, with the color intensity indicating the strength and direction of the correlation. Positive correlations are displayed in shades of blue, while negative correlations are shown in shades of red, with the scale on the right providing a reference for interpreting the correlation coefficients.

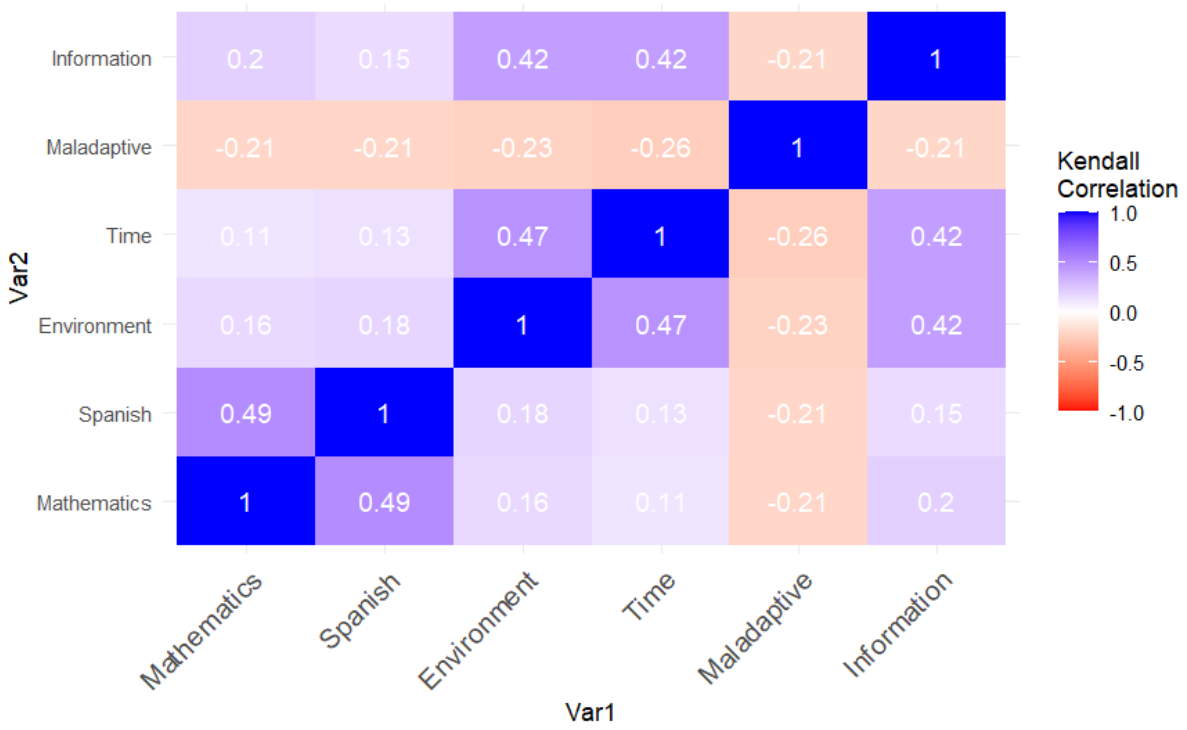


Figure 5: Kendall correlation amongst academic grades in Spanish and Mathematics and the self-regulation variables.

A key observation from the plot is the strong positive correlation (0.49) between Mathematics and Spanish grades, indicating that students who perform well in Mathematics also tend to excel in Spanish. The correlation between Mathematics and Environment management is relatively weak (0.16), suggesting minimal relationship between Mathematics performance and students' ability to manage their learning environment. Similarly, the weak correlation (0.11) between Mathematics and Time management indicates no significant relationship between Mathematics grades and students' time management skills. However, a negative correlation (-0.21) is observed between Mathematics and Maladaptive behavior, implying that students who exhibit fewer maladaptive behaviors tend to achieve higher grades in Mathematics. Additionally, a moderate correlation (0.20) exists between Mathematics and Information management, suggesting a slight advantage in Mathematics performance for students who manage information effectively.

Spanish grades show a moderate positive correlation with Information management (0.15) and weak correlations with Environment management (0.18) and Time management (0.13), similar to Mathematics. The correlation between Spanish grades and Maladaptive behavior is also negative (-0.21), echoing the pattern observed with Mathematics grades. There is a notable moderate positive correlation (0.47) between Environment and Time management, indicating that students proficient in managing their environment are also likely to excel in time management. Conversely, the negative correlation (-0.23) between Environment management and Maladaptive behavior suggests that effective environment management is associated with fewer maladaptive behaviors.

The negative correlation (-0.26) between Time management and Maladaptive behavior implies that better time management skills are associated with fewer maladaptive behaviors. Additionally, a moderate positive correlation (0.42) between Time management and Information management indicates that students proficient in managing their time also tend to manage information effectively. Finally, the negative correlation (-0.21) between Maladaptive behavior and Information management suggests that students exhibiting fewer maladaptive behaviors are better at managing information.

In conclusion, the Kendall correlation plot provides valuable insights into the relationships between different academic and self-regulation variables. The strong positive correlation between Mathematics and Spanish grades highlights the potential interdependence of language and numerical skills. The negative correlations between maladaptive behavior and academic performance, as well as self-regulation skills, underscore the importance of addressing maladaptive behaviors to enhance student outcomes. Understanding these relationships can inform targeted interventions and support strategies to improve student performance across various domains.

3 Model-based clustering via finite Gaussian mixture models

Cluster analysis has traditionally advanced through the invention and empirical testing of ad hoc methods, often isolated from formal statistical procedures. However, in recent years, basing cluster analysis on probabilistic models has proven beneficial for understanding when existing methods are effective and for developing new methods. This probabilistic foundation has enabled more systematic and theoretically grounded approaches to clustering, enhancing both the understanding and development of clustering techniques (McLachlan 2011).

Model-based clustering (MBC) is a sophisticated statistical technique used to identify groups or clusters within data by assuming that the data are generated from a finite mixture of probability distributions. Each cluster corresponds to a distinct component of this mixture, and the goal of MBC is to estimate the parameters of these components to best describe the underlying data structure. Specifically, finite Gaussian mixture models (GMMs) are commonly used for continuous and normally distributed variables, with cluster membership represented by a latent categorical variable. Thus, GMMs are considered a type of latent profile analysis model.

Finite mixture models (FMMs) provide the statistical framework for MBC, allowing complex data to be modeled by combining simpler distributions. An FMM assumes that observed data are generated from a finite mixture of underlying distributions, each corresponding to a distinct subgroup or cluster. Gaussian mixture models are a popular variant of FMMs, where each underlying distribution is a (multivariate) Gaussian distribution. This implies that data within each cluster are normally distributed, but with potentially different means and covariance matrices. The appeal of GMMs lies in the fact that mixtures of Gaussians can accurately approximate any continuous density, making them a versatile tool for a wide range of applications (Fraley and Raftery 2002).

To estimate the parameters of a GMM and the associated latent variable for cluster membership, a likelihood-based approach is typically employed. The likelihood function quantifies the probability of observing the data given the parameter values and the latent variable. Maximum likelihood estimation (MLE) is a common method for estimating parameters and the latent variable that maximises the likelihood function. Model selection, which involves comparing models with different numbers of clusters and parameterizations, is used to determine the optimal number of clusters, ensuring the best fit for the data (McLachlan and Rathnayake 2014).

In summary, model-based clustering from the perspective of latent variable modeling assumes that data are generated from a probabilistic model with a specific number of clusters. A likelihood-based approach is used to estimate model parameters and the latent variable representing cluster assignment for each observation, guiding the selection of the optimal number of clusters. GMMs are a common framework for MBC, assuming that data in each cluster follow a Gaussian distribution, providing a robust and flexible method for clustering analysis in many applications (McLachlan 2011; McLachlan and Rathnayake 2014).

3.1 Definition and model fitting

At its core, model-based clustering treats the entire dataset as being derived from a mixture model with G components. Each component corresponds typically to a distinct cluster and is characterised by its own probability distribution. The primary task in MBC is to estimate the parameters of these distributions, including the mixing proportions, mean vectors, and covariance matrices.

The mixture model can be mathematically represented as:

$$f(\mathbf{y}_i; \Psi) = \sum_{g=1}^{G} \pi_g \phi_p(\mathbf{y}_i; \mu_g, \Sigma_g),$$

where $\mathbf{y}_i$ denotes the engagement scores for student i, ϕ_p represents a p-dimensional multivariate normal (MVN) distribution, and $\Psi = \{\pi_1, \ldots, \pi_{G-1}, \mu_1, \ldots, \mu_G, \Sigma_1, \ldots, \Sigma_G\}$ denotes the set of parameters of the mixture model. Here, π_g, μ_g, and Σ_g are the cluster-specific mixing proportion, mean vector, and covariance parameters, respectively. The mixing proportions are typically $0 < \pi_g < 1$ and are constrained to sum to 1. While it is possible in `mclust` to constrain the mixing proportions to be equal to $1/G$, they are typically estimated subject to these constraints by default.

Clustering is typically viewed as an incomplete data problem or a latent variable problem, where the each observation's cluster membership is unobserved or latent. This perspective makes the problem suitable for inference using the Expectation-Maximisation (EM) algorithm in an MLE setting (Dempster et al. 1977) or Bayesian inference (Bensmail et al. 1997), which elegantly handles latent variables.

Since the formalisation of the EM algorithm by Dempster et al. (1977), it has become the predominant method for inference in model-based clustering (McLachlan et al. 2019). In this context, the data are considered as (y_i, z_i) for $i = 1, \ldots, n$, where y_i denotes the observed data on d variables, and $z_i = (z_{i,1}, \ldots, z_{i,G})$ represents the unobserved portion of the data. Specifically, $z_{i,g}$ is defined as:

$$z_{i,g} = \begin{cases} 1 & \text{if } i \text{ belongs to cluster } g \\ 0 & \text{otherwise.} \end{cases}$$

Each $z_1, \ldots, z_n$ is assumed to be independent and identically distributed, following a multinomial distribution with G categories, with probabilities $\pi_1, \ldots, \pi_G$. Instead of directly maximising the observed data likelihood function, the EM algorithm maximises the complete-data log-likelihood, given by:

$$\ell_c = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \left[\log(\pi_g) + \log(\phi_p(\mathbf{y}_i; \mu_g, \Sigma_g)) \right],$$

The EM algorithm is iterative, consisting of an expectation step (E-step) and a maximisation step (M-step). In the E-step, the conditional expectation of the complete data log-likelihood function, given the observed data and the current parameter estimates, is computed. In the M-step, the expected complete data log-likelihood function from the E-step is maximised with respect to the model parameters. These E- and M-steps are iterated until convergence, achieving at least a local maximum of the observed data likelihood function under mild regularity conditions (Dempster et al. 1977). Convergence of the EM algorithm can be assessed by monitoring the change in log-likelihood and/or parameter estimates between successive iterations or by using Aitken's acceleration-based stopping criterion (McLachlan and Krishnan 2007). Initial values for the EM algorithm are crucial; (Melnykov and Melnykov 2012) suggest useful initialisation strategies. The `mclust` package specifically employs model-based agglomerative hierarchical clustering (Scrucca et al. 2016).

In practice, for the complete data log-likelihood function where $f_g(\cdot \mid \theta_g)$ is the multivariate Gaussian distribution, the E-step involves computing the conditional expectation of $z_{i,g}$ for $i = 1, \ldots, n$ and $g = 1, \ldots, G$ given the current parameter estimates and the data y. Given the expected values $\hat{z}$, the M-step maximises the expected complete data log-likelihood function with respect to the mixing proportions and mean parameters, for which closed-form solutions are available; closed-form solutions for the covariance matrices are available for certain parameterisations.

Ultimately, the EM algorithm produces maximum likelihood estimates of the parameters and posterior probabilities of component membership upon convergence, with the latter denoted as $\hat{z}_{ig}$. Upon convergence, $\hat{z}_{i,g}$, the conditional expectation of $z_{i,g}$, represents the estimated conditional probability that

observation i belongs to cluster g. Thus, a hard classification of cluster membership for each observation is obtained by assigning each observation to the cluster $g\prime$ for which $\hat{z}_{i,g\prime} = \max_g \hat{z}_{i,g}$, and the uncertainty in that cluster membership is quantified by $(1 - \max_g \hat{z}_{i,g})$ for observation i (Bensmail et al. 1997).

3.2 Gaussian parsimonious clustering models

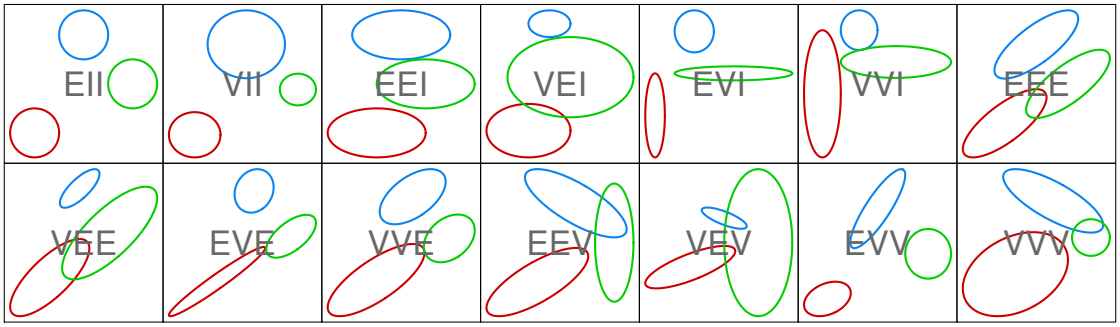
Among various mixture models, Gaussian parsimonious (GPCMs) are particularly popular in model-based clustering, especially in applications such as the `mclust` package in `R`. GPCMs assume that each component of the mixture follows an elliptical multivariate normal distribution. This assumption simplifies the clustering process and allows for parsimonious parameterisation of the covariance matrices. The covariance matrix of each component can be parameterised using an eigen-decomposition of the form:

$$\Sigma_g = \lambda_g D_g A_g D_g^\top,$$

where λ_g is a scalar controlling the volume of the ellipsoid, A_g is a diagonal matrix specifying the shape of the density contours with $\det(A_g) = 1$, and D_g is an orthogonal matrix determining the orientation of the ellipsoid.

Constraining different combinations of elements of the above eigendecomposition across clusters affords great flexibility in modelling the geometric characteristics of the clusters. A model where all elements are allowed to vary (referred to as a VVV model, for varying volume, varying shape, and varying orientation) is referred to as a fully constrained model. Such a model is what LPA typically refers to. However, such implementations do not allow for specifying which features (orientation, size, shape) should be common or different among clusters. The above reparameterisation of the covariance matrix allows for partial feature commonality across clusters, moving between heteroscedasticity and homoscedasticity. In contrast to the VVV model, the EVE model for example allows equal volumes and orientations, but varying shapes. All 14 models in the GPCM family are illustrated in Figure 6. This figure is a reproduction of a similar figure in Murphy and Murphy (2020), with permission from the authors.

Figure 6: Ellipses of isodensity for each of the 14 Gaussian models obtained by eigen-decomposition in the case of three groups in two dimensions.



3.3 Model selection and estimation

One of the significant challenges in MBC is determining the number of components G. The value of G is typically unknown and must be estimated from the data. Additionally, there are multiple possible choices for the component distributions in GPCMs, which adds another layer of complexity to the clustering process. As different models may vary based on their assumptions about the shape, volume, and orientation of the covariance matrices, and selecting the right model is crucial for accurate clustering.

Selecting the optimal model and the number of components is typically achieved using the Bayesian Information Criterion (BIC), due to its balance between goodness of fit and model complexity. A higher BIC value indicates a better model, as it suggests a better trade-off between the model's accuracy in describing the data and its simplicity. This approach helps prevent overfitting by penalising models with unnecessary complexity, thereby ensuring that the selected model is both statistically robust and interpretable.

The final clustering partition is obtained by assigning each observation to the cluster with the highest posterior probability, $\hat{z}_{jg}$, obtained at the convergence of the model with the highest BIC. The BIC for model m is calculated as:

$$\text{BIC}_m = 2\hat{\ell} - \kappa_m \log(N),$$

where $\hat{\ell}$ is the maximised log-likelihood, κ_m is the number of parameters estimated by the model, and N is the number of observations. Thus, this criterion evaluates the fit of a GMM to a given set of data by considering both the likelihood of the data given the model and the complexity of the model itself, represented by the number of parameters to be estimated.

However, despite the appeal of the BIC being based on an underlying statistical modelling framework, it is not the only criterion available. While there is no universally accepted standard for determining the optimal model in Gaussian mixture modeling, researchers can follow established guidelines to make informed decisions. The process involves examining fit indices such as the BIC and assessing model interpretability, and ensuring alignment with theoretical expectations.

A recommended approach by Scrucca et al. (2016) is to start with a one-cluster solution for each model, which serves as a baseline for comparison. Subsequently, the number of clusters is incrementally increased, and each new solution is evaluated to determine if the additional cluster provides a statistically and conceptually superior model. This iterative process allows researchers to identify the point at which adding more clusters no longer significantly improves the model. By adhering to these guidelines, researchers can systematically and rigorously identify the most suitable model for their data, leveraging both statistical criteria and theoretical considerations to inform their choices.

3.4 Adding regularising prior distributions

Given the relative lack of unique values for the engagement variables, a superior BIC could be achieved by incorporating prior distributions to achieve regularisation. Often, BIC plots without priors shows

a number of jagged peaks, with some BIC values missing for some models due to failure in the EM computations caused by singularity and/or shrinking components (Fraley et al. 2012). It has been suggested that including a prior distribution over the mixture parameters is an effective way to avoid singularities and degeneracies in maximum likelihood estimation. Furthermore, this can help to prevent overfitting and improve model performance. In situations such as here where the variables of interest are discrete or take on only a few unique values relative to the overall sample size, including a prior distribution can help to regularise the estimation process of the model. The EM algorithm can still be used for model fitting, but maximum likelihood estimates (MLEs) are replaced by maximum *a posteriori* (MAP) estimates. A slightly modified version of BIC can be used for model selection, with the maximised log-likelihood replaced by the log-likelihood evaluated at the MAP or posterior mode. This can be easily achieved within `mclust`, however only 10 of the 14 model types are available for use in conjunction with priors.

Thus, to mitigate against singularities and degeneracies in maximum likelihood estimation due to the non-uniqueness of the engagement scores, we incorporate a prior distribution strategy. The `mclust` package in `R` provides a robust framework for implementing these priors and we know describe the default priors implemented in that package.

For multivariate data, we use a normal prior on the mean (conditional on the covariance matrix):

$$\mu \mid \Sigma \sim N(\mu_P, \Sigma/\kappa_P) \propto |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{\kappa_P}{2}\text{tr}\left((\mu-\mu_P)^\top\Sigma^{-1}(\mu-\mu_P)\right)\right],$$

where the prior mean μ_P is typically the overall sample mean of the observed data and κ_P is a hyper-parameter which controls the degree of shrinkage of the component-specific means to this overall mean, and $\text{tr}(\cdot)$ denotes the trace operator, i.e., the sum of the diagonal elements. By default, $\kappa_P = 0.01$.

An inverse Wishart prior is assumed for the covariance matrix:

$$\Sigma \sim \text{inverseWishart}(\nu_P, \Lambda_P) \propto |\Sigma|^{-\frac{\nu_P+d+1}{2}} \exp\left[-\frac{1}{2}\text{tr}\left(\Sigma^{-1}\Lambda_P^{-1}\right)\right],$$

where ν_P is the degrees of freedom hyperparameter (equal to $d+2$ by default, where d is the dimension of the data) and Λ_P is a scale matrix hyperameter, equal to $S/G^{2/d}$ by default, where S is the sample covariance matrix and G is the number of mixture components, as before.

3.5 Adding the noise component to the model

Mixture models, particularly mixtures of multivariate Gaussian distributions, are extensively used in statistical modeling. The Gaussian mixture model assumes that within each group, data have an elliptical scatter, which can be limiting. Non-elliptical groups are often modeled by multiple components, leading to over-fitting and ambiguous clustering rules, resulting in higher misclassification rates. Additionally, GMMs may struggle to accommodate clusters with heavy tails or outliers, posing further challenges in accurate modeling. Furthermore, issues such as variations in component volume and shape can lead

to unbounded mixture likelihoods. To ensure meaningful solutions, we focus on models where the log-likelihood converges to a finite value. For models with unconstrained covariance matrices (Σ), it is crucial that each cluster contains at least $p + 1$ units (where p is the number of dimensions) to prevent computational singularities. In practice, high values of G (number of components) often lead to spurious solutions. These may manifest as models with empty components that effectively reduce the number of components or as degenerate components with too few observations, sometimes even singletons. An excellent review of issues of degeneracy and spurious solutions is provided by García-Escudero et al. (2018).

To address the concerns, we avail of a a practical framework for non-Gaussian clustering which has been proposed by Banfield and Raftery (1993), by incorporating a uniformly-distributed noise component. In model-based clustering, a noise component is often added to address the presence of outliers in the data. Outliers can significantly distort the clustering solution by affecting the parameter estimates of the model. Including a noise component helps in enhancing the robustness of the clustering model by accounting for observations that do not fit well into any of the defined clusters or allowing the main clusters to be modeled more accurately without the influence of atypical data points and providing a clearer interpretation of the clusters by distinguishing between genuine clusters and noise (Hennig and Coretto 2008).

Adding a noise component typically involves defining a separate class for outliers. This class is modeled with a uniform distribution, which reflects the assumption that outliers are uniformly distributed in the feature space and do not belong to any specific cluster. In `mclust`, the noise component is referred to as component "0", an additional mixing proportion π_0 is accounted for, along with the hypervolume V of the uniform distribution enclosing the data being automatically estimated. Ultimately, the model can be rewritten as

$$f(\mathbf{y}_i; \Psi) = \frac{\pi_0}{V} + \sum_{g=1}^{G} \pi_g \phi_p(\mathbf{y}_i; \mu_g, \Sigma_g),$$

and constrained parameterisations of Σ_g and the aforementioned prior distributions can still be used in conjunction with this expanded model definition.

Practically speaking, incorporating a noise component in `mclust` requires an initial guess as to which observations are non-Gaussian outliers. This involves using a measure to calculate distance to identify potential outliers. The Mahalanobis distance is effective for multivariate data as it accounts for the correlations between variables. We leverage the fact that squared Mahalanobis distances follow a chi-squared distribution to initially designate observations with probability greater than 0.9 of being noise as noise.

Overall, by addressing the conditioning of covariances and incorporating a noise component, the `mclust` package offers robust solutions for mixture models, even when faced with challenging data distributions. These techniques ensure finite log-likelihood convergence and prevent the formation of spurious solutions, thereby enhancing the reliability and interpretability of the fitted models.

3.6 Performance assessment measures in model-based clustering

While the BIC is used as a tool in model selection to choose the number of components and select among the 14 GPCM covariance parameterisations, it is only a relative measure. It informs as to whether one model is better than another, but says nothing about the inherent quality of the "best" model. To quantify the clustering performance in this thesis, we rely on two measures frequently employed in latent profile analysis.

3.6.1 Entropy

Entropy is a measure of uncertainty or randomness in a dataset. In the context of model-based clustering, entropy is used to evaluate the uncertainty associated with the assignment of observations to clusters. High entropy indicates that observations are assigned to clusters with high uncertainty, implying that the clusters are not well-defined or that there is significant overlap between them. Conversely, low entropy suggests that observations are assigned to clusters with high certainty, indicating well-defined and distinct clusters (Jung and Wickrama 2008).

The entropy indicating the clear delineation of clusters can be assessed to obtain a more robust and appropriate basis for the comparison of the models. In LPA, a slightly different definition of normalised entropy is used. Mathematically, the observation-specific contribution E_i to the overall entropy E for a clustering model can be defined as:

$$E_i = 1 + \frac{\sum_{g=1}^{G} \hat{z}_{ig} \log(\hat{z}_{ig})}{\log(G)},$$

where all quantities are as previously defined. Subsequently, these contributions can be averaged such that $E = \sum_{i=1}^{n} E_i/n$. An overall entropy value close to 1 is ideal, while values above 0.6 are generally considered acceptable, although there is no agreed upon optimal cutoff for entropy.

3.6.2 Average posterior probabilities

Average posterior probabilities are used to assess the reliability and quality of the cluster assignments. The posterior probability $\hat{z}_{ig}$ represents the probability that observation i belongs to cluster g given the data and the model. These probabilities are computed for each observation and each cluster (Biernacki et al. 2000). The average posterior probability for cluster g is the mean of the posterior probabilities for all observations assigned to that cluster according to the MAP procedure:

$$\text{Average Posterior Probability for Cluster } g = \frac{1}{n_g} \sum_{i \in C_g} \hat{z}_{ig}$$

where C_g is the set of observations assigned to cluster g and n_g is the size of this set.

High average posterior probabilities indicate that the observations are assigned to their respective clusters with high confidence, reflecting well-defined clusters (Fraley and Raftery 2002). Low average posterior

probabilities suggest that the assignments are made with less confidence, indicating potential issues with the clustering solution. A rule of thumb is that a cutoff of 0.8 for average posterior probabilities has been suggested to indicate acceptably high assignment certainty and well-separated clusters.

In summary, entropy and average posterior probabilities are key metrics in model-based clustering for evaluating the quality and reliability of the cluster assignments. High entropy and high average posterior probabilities are desirable, as they indicate well-defined and distinct clusters with high confidence in the assignment of observations to clusters.

3.7 Bootstrapping

The bootstrap is a versatile and powerful technique for approximating the sampling distribution of a statistic of interest. This approach involves generating a large number of bootstrap samples from the empirical distribution. This can be achieved by resampling with replacement from the observed data, known as the nonparametric bootstrap, or from a parametric distribution where unknown parameters are replaced by their estimates, referred to as the parametric bootstrap. A Bayesian variant of the bootstrap, as introduced by Rubin (1981), involves resampling with weights for each observation drawn from a uniform Dirichlet distribution. A related method is the weighted likelihood bootstrap (Newton and Raftery 1994), which repeatedly fits a statistical model using weighted maximum likelihood, with weights derived similarly to the Bayesian bootstrap.

Consider $\hat{\theta}$ as the estimate of a set of GMM parameters θ for a given model M, which is determined by the chosen covariance parameterisation and number of mixture components. The bootstrap distribution for these parameters is obtained through the following steps:

- Drawing a bootstrap sample of size n using one of the resampling techniques mentioned above to form the bootstrap sample $(x_1^\star, \ldots, x_n^\star)$.
- Fitting the GMM M to obtain the bootstrap estimates $\hat{\theta}^\star$.
- Repeating the previous steps a large number of times B.

The resulting bootstrap distribution for the parameters of interest, $\hat{\Psi}_1^\star, \hat{\Psi}_2^\star, \ldots, \hat{\Psi}_B^\star$, can then be used to compute bootstrap standard errors (as the square root of the diagonal elements of the bootstrap covariance matrix) or bootstrap percentile confidence intervals. Further details on this process can be found in Efron (1992).

Practically, bootstrap resampling in the context of model-based clustering can be implemented using the `MclustBootstrap()` function in the `mclust` package in `R`. This function accepts the fitted model object returned from functions such as `Mclust()` or `mclustBIC()`. Additionally, it includes the optional argument `type`, which specifies the type of bootstrap samples to draw (`"bs"` for nonparametric bootstrap, `"pb"` for parametric bootstrap, and `"wlbs"` for weighted likelihood bootstrap), and the argument `nboot`, which sets the number of bootstrap samples B. For reliable confidence intervals, at least 999 samples are recommended.

In this thesis, however, we only use bootstrapping procedures to interrogate the results of the final optimal model. This final optimal model, as will be shown in Section 4 incorporates both prior distributions and a noise component. As the implementation of `MclustBootstrap()` does not account for either priors or noise components at the time of writing, some manual modification of the source code was required. This was only feasible for the `"bs"` method; consequently we only present results of a nonparametric bootstrap procedure.

4 Clustering results

We begin by describing the results of using `mclust` with its default arguments and then present our richer analyses incorporating prior distributions and a noise component.

4.1 Preliminary clustering results

In the initial phase of model-based clustering, we aimed to identify the most appropriate Gaussian parsimonious clustering model based on the BIC. Utilising the `mclust` package in `R`, we evaluated a range of models characterised by different numbers of clusters (from 1 to 9) and all 14 available covariance parameterisations. The BIC was employed as the model selection criterion, which considers both the covariance structure and the number of mixture components in the model. The results of this evaluation are depicted in Figure 7 and summarised in Table 2.

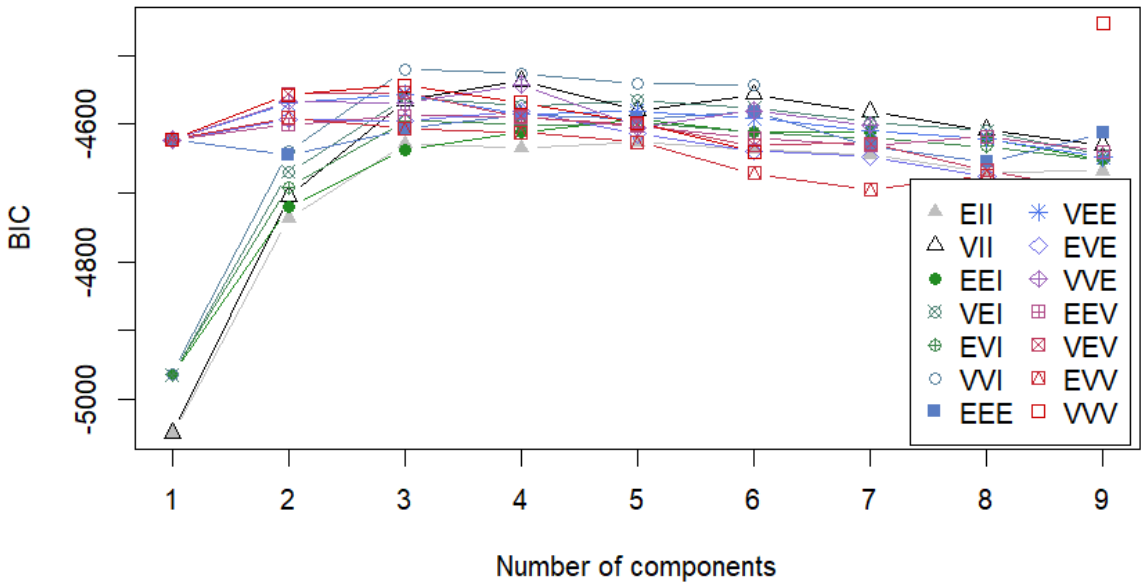


Figure 7: BIC values for all 14 available covariance parameterisations using the default settings in `mclust`.

Table 2: Best BIC values using default `mclust` settings.

	VVV,9	VVI,3	VVI,4
BIC	-4453.309	-4520.169	-4525.757
BIC diff	0.000	-66.86	-72.45

Among all the evaluated models, the VVV model with 9 components was identified as the best solution based on the BIC values. The VVV model allows for variable volume, variable shape, and variable orientation of the covariance matrices, providing a high degree of flexibility in capturing the underlying data structure. The specific BIC values for the best three covariance parameterisations are presented in Table 2. These include VVV, VVI, and VVI models with 9, 3, and 4 components, respectively. The VVV model with 9 components had the lowest BIC value of -4453.339, indicating the best fit among all tested models. The differences in BIC values between the VVV model and the next best models (both VVI) were substantial, reinforcing the selection of the VVV model with 9 components as the optimal clustering solution.

However, despite the favourable BIC outcome of the VVV model, which corresponds to the standard LPA model, the solution with 9 clusters may not be the best choice for the dataset. The primary reason for this is that the selection of 9 clusters can indicate overfitting. Overfitting occurs when the model captures not only the underlying data structure but also the noise and minor variations within the dataset. This can lead to overly complex models that do not generalise well to new, unseen data.

Additionally, from a practical perspective, a solution with 9 clusters might not be easily interpretable or meaningful. An optimal clustering solution should balance statistical fit with interpretability and practical relevance. Therefore, while the VVV model with 9 components presents the lowest BIC, it is crucial to consider other factors, such as model simplicity, cluster interpretability, and theoretical conformance, before concluding the optimal number of clusters for the dataset. Further analysis and validation are necessary to ensure that the chosen model provides a robust and meaningful segmentation of the data. On the evidence of Table 2, a VVI model with 3 clusters seems particularly worthy of further exploration.

The plot in Figure 8 highlights the inadequacy of the nine-cluster solution. Despite the high number of clusters, there is considerable overlap between clusters, indicating that the model may be overfitting the data. The presence of many clusters fails to provide clear and distinct separation among the data points, suggesting that this clustering solution does not capture the underlying structure of the dataset effectively. The observed overlap and lack of clear boundaries between clusters point to the necessity of re-evaluating the clustering model, potentially reducing the number of clusters or adjusting the clustering methodology to achieve a more optimal solution.

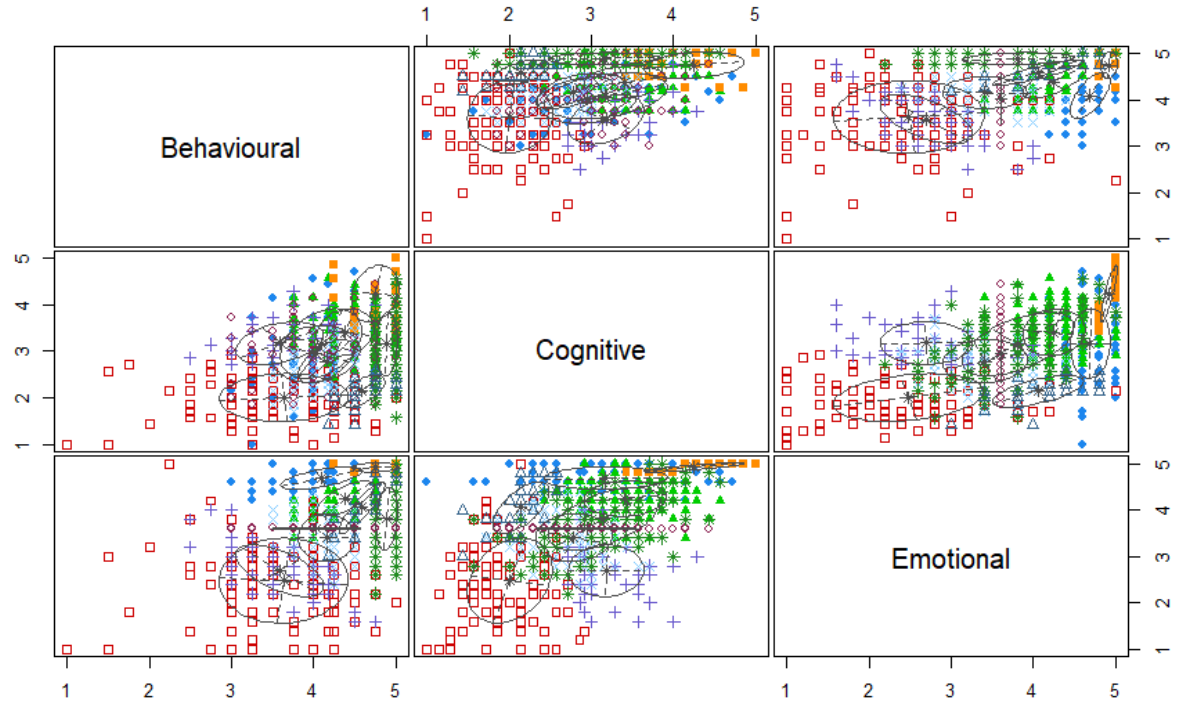


Figure 8: The pairwise scatterplot matrix illustrates the clustering results for the dataset using a model with nine clusters. Each panel represents the scatterplot for a pair of variables, categorized into three dimensions: Behavioral, Cognitive, and Emotional. The points are colored and shaped based on the cluster assignments.

That being said, the findings from this preliminary clustering analysis provide a solid foundation for further investigation. As this is the maximum of the range considered, there is evidence that this model overfits the number of clusters and represents a so-called 'spurious solution'. The identified clusters can be analysed in more detail to understand their characteristics and the relationships between different clusters.

4.2 Advanced cluster modelling

In the initial analysis above, a default model was employed without prior adjustments or inclusion of a noise component, resulting in a nine-cluster solution which did not yield optimal clustering performance due to significant overlap among clusters and a lack of clear separation. To enhance the model's performance and address overfitting issues, prior distributions were incorporated in this analysis. The inclusion of priors helps to regularise the model, aiming to find a more realistic clustering solution by preventing the model from fitting noise in the data. Regularisation is also motivated by the relative scarcity of distinct values for the engagement scores. Initially, we incorporate the default prior settings in `mclust`, which correspond to the priors previously described, using the function `defaultPrior()`.

The updated BIC scores in Figure 9 and Table 3 reflect this adjustment, guiding the selection of a clustering solution with fewer components that balances model complexity and fit to the data, and avoiding spurious solutions across the entire range of the number of components.

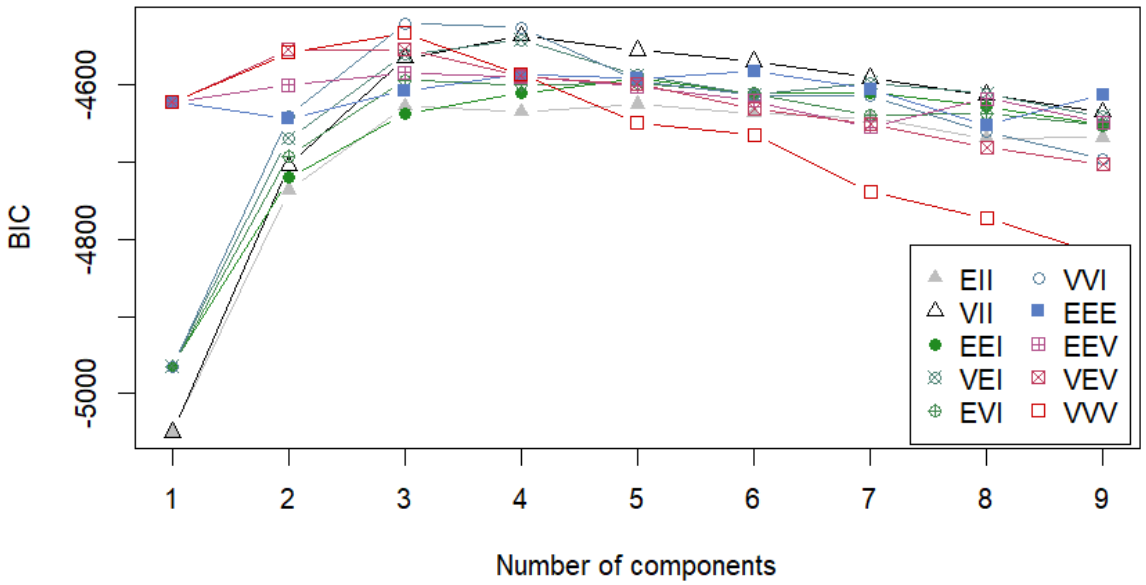


Figure 9: This plot illustrates the BIC scores for various clustering models applied to the dataset using default priors. The x-axis represents the number of clusters, while the y-axis shows the BIC scores. Each line corresponds to a different covariance structure for the Gaussian Mixture Models (GMMs), indicated by different shapes and colors.

Table 3: Best BIC values with priors.

	VVI,3	VVI,4	VVV,3
BIC	-4521.213	-4526.906	-4533.572
BIC diff	0.000	-5.69	-12.36

After implementing these refinements, we evaluated the BIC values to identify the best models. Table 3 summarises the BIC values for the top-performing models with different priors. The results indicate that the VVI model with 3 clusters yielded the best BIC value of -4521.213, followed by the VVI model with 4 clusters and the VVV model with 3 clusters. These findings suggest that the VVI model configurations tend to perform well with the given data and the applied priors. Such a model is denoted as (VVI, 3) and features diagonal covariance matrices of differing volume and shape, with axis-aligned orientation. Within each cluster, the variables are independent. The resulting clusters are shown in Figure 10.

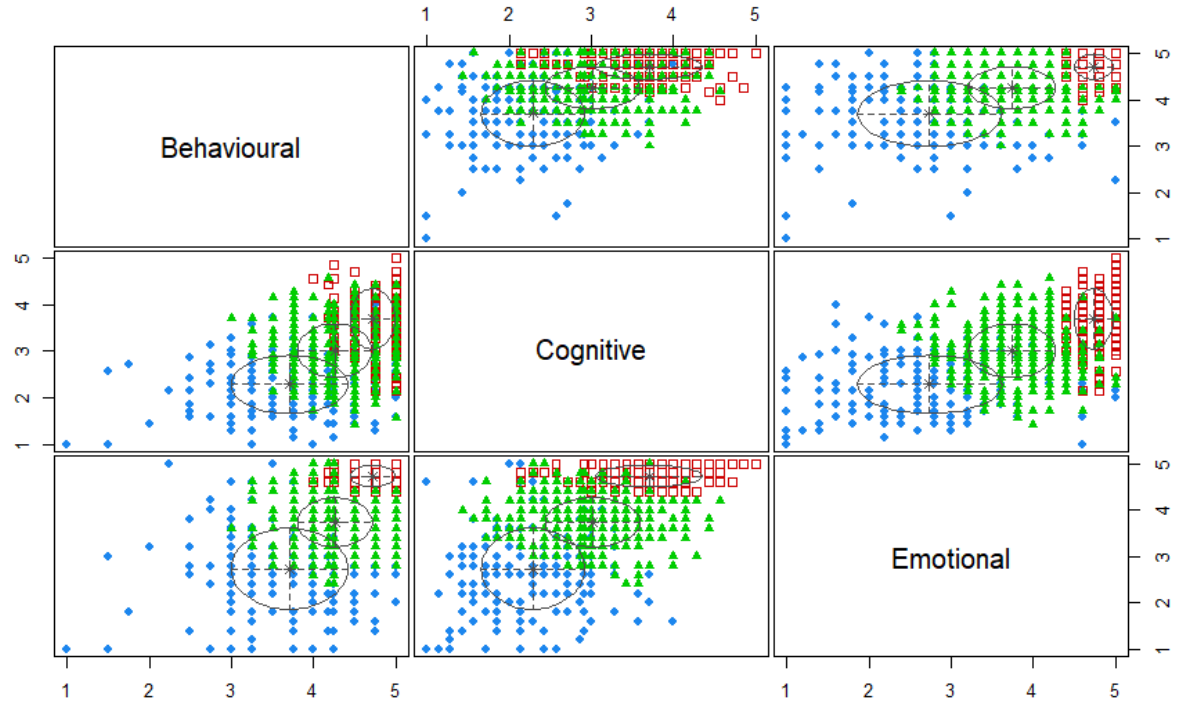


Figure 10: This figure presents enhanced pairwise scatterplots of engagement data across Behavioral, Cognitive, and Emotional dimensions after using 'mclust' with the default prior settings. Different colors and shapes represent distinct clusters within each dimension, facilitating a clearer interpretation of how clusters overlap or separate across different engagement metrics.

In order to ensure that the best model has been found, we consider three alternatives. In each case, we employ the same prior distributions from before. Firstly, we attempt to find a more parsimonious model by enforcing the constraint that the mixing proportions be equal to $1/G$ for all clusters. Secondly, we modify some of the hyperparameters of the priors. Finally, we add a uniform noise component to capture outlying observations which depart from normality, so as to remove their effect on their more defined clusters. We narrow the range of candidate models to $G = 2, \ldots, 5$, as there appears to be little support for numbers of clusters outside this range according to the BIC values already computed above.

4.2.1 Constraining the mixing proportions

We begin by constraining the mixing proportions, such that they are equal across clusters. Specifically, we set them to $\frac{1}{G}$, where G represents the number of clusters, and are no longer estimated. This aims to enforce a more parsimonious model by assuming that each cluster has an equal contribution to the overall data distribution. By doing so, we simplify the model and potentially enhance its interpretability and robustness. However, in doing so, no superior model could be found. In fact, the top-performing model under these constraints is a 4-component VEI model, with a BIC value of 6432.85. As this BIC value is positive, in contrast to the BIC values reported earlier, this is indicative of another spurious solution. The best non-spurious solution in terms of BIC is a (VVI, 5) model, however its BIC of -4541.91 is inferior to the previously estimated (VVI, 3) model with priors and unequal mixing proportions.

4.2.2 Modifying the hyperparameters of the priors

In the second alternative, we adjust some of the hyperparameters of the priors used in the model. This modification involves fine-tuning the parameters that govern the prior distributions of the model's parameters, while letting the distributions themselves be unchanged. The objective is to better capture the underlying data structure by allowing the priors to more accurately reflect the data's characteristics. This step ensures that the priors are neither too informative nor too vague, striking a balance that facilitates effective clustering.

To address the limitations identified, we proceed by modifying the hyperparameter values to enhance model performance and stability. We consider three strategies:

- In the first strategy, we assume a diagonal scale parameter for the prior on the variance. Given that the best model to date is a VVI model, this adjustment aims to reflect the conditional independence of the variables *a priori*, while avoiding the complexities associated with full covariance matrices.
- In the second strategy, we remove the shrinkage prior on the mean but assume the default prior for the variance. This approach focuses on reducing the influence of the prior on the mean vector, allowing the data to more strongly dictate the location parameters of the clusters. This could potentially lead to more data-driven and accurate mean estimates for the clusters.
- In the third strategy, we combine both modifications by removing the prior on the mean and assuming a diagonal scale parameter for the prior on the variance. This comprehensive approach aims to minimise the influence of priors on both the mean vector and covariance matrices, thus relying more heavily on the observed data to determine the cluster parameters. This strategy may enhance the model's flexibility and robustness.

We omit the details of the BIC scores under each strategy and note only that the second strategy yielded the model with the overall best BIC score. It is again a (VVI, 3) model under which the default prior on the covariances is used, but the shrinkage hyperparameter κ_P in the prior on the means is set to zero. This yields a BIC of -4520.89, which compares favourably to the BIC of -4521.213 for the earlier (VVI, 3) model with the default priors for both the means and covariances.

4.2.3 Adding a uniform noise component

Finally, we incorporate a uniform noise component to account for outlying observations that deviate from normality. These outliers can significantly affect the clustering results by distorting the defined clusters. By introducing a noise component, we aim to isolate the influence of these outliers, allowing the model to focus on the more structured parts of the data. This addition helps in mitigating the impact of anomalies and enhances the overall clustering quality.

Doing so requires an initial guess of which observations are outliers, which we obtain, as previously described, using the Mahalanobis distance given the multivariate nature of the data. The observations initially designated as outliers as a result are highlighted in the form of red triangles in Figure 11.

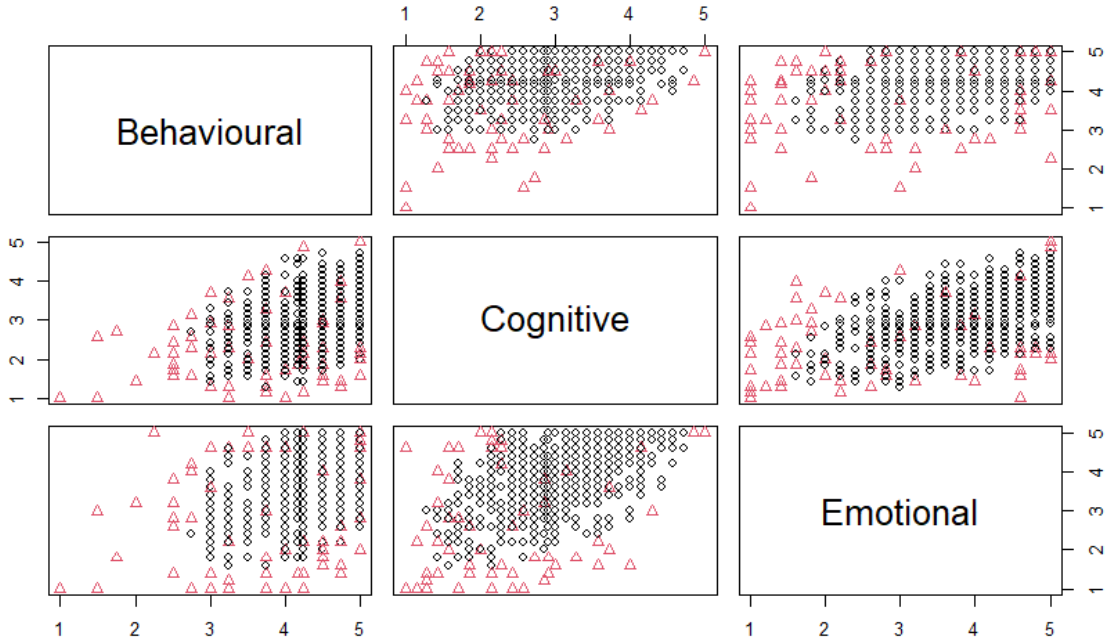


Figure 11: This pairwise scatterplot of the engagement data highlights observations initially designated as noise according to our procedure based on the Mahalanobis distance using red triangles.

In conjunction with the use of a noise component, we retain the same configuration of hyperparameters that were identified as optimal above, i.e., our model with a noise component again removes the prior on the means but retains the default prior on the covariances. Such a model is a (VVI, 3) model, where the number three refers only to the number of noise components, but now with the best BIC to date of -4507.66.

5 Exploring the latent profiles

After extensive evaluation and experimentation with various models, we have identified the final model that best represents our data based on the BIC. The model selection process involved exploring different GPCMs with varying configurations of priors and associated hyperparameters, as well as incorporating a uniform noise component to account for outliers.

The final model, which we denote as (VVI, 3), outperforms all previous models based on the BIC criterion and includes three clusters with diagonal covariance matrices of varying volumes and shapes, unequal mixing proportions, a regularising prior on only the covariance matrices, and an additional non-Gaussian component.

We now summarise the results of this optimal in greater detail and then investigate how the clusters uncovered by this model relate to the available covariates related to academic achievement, demographics, and self-regulation variables.

5.1 Summarising the final model

Table 4 displays that 356, 108, and 237 observations are in clusters 1, 2, and 3 out of the 717 students in the sample, while 16 of observations are designated as noise. Thus, cluster 1 is the largest, followed by cluster 3, while cluster 2 is relatively small and the noise component only captures a very small number of outlying students.

Table 4: Number of students in each cluster, including the noise component, expressed as percentages of 717 in parentheses.

Noise	Cluster 1	Cluster 2	Cluster 3
16 (2.23%)	356 (49.65%)	108 (15.06%)	237 (33.05%)

To enhance the understanding of these results and provide a measure of uncertainty for the estimated means of the latent profiles, it is important to incorporate bootstrap resampling. This can be effectively achieved using the `MclustBootstrap()` function. This approach will allow us to estimate the variability and confidence intervals of the mean engagement scores for each profile, thereby offering a more comprehensive and reliable interpretation of the data. The process of bootstrap resampling has been previously detailed and provides a robust method for assessing the stability of our parameter estimates.

The bootstrap technique involves repeatedly resampling the data with replacement and refitting the model to these bootstrap samples, thereby generating a distribution of the parameter estimates. This method offers insights into the variability and stability of the estimates.

Firgure 12 is visual representation of bootstrap distributions of the component means of the GMM. Each subplot in the figure represents the distribution of one of the component means obtained from the bootstrap samples. These distributions are visualized as histograms, with the original GMM estimate indicated by a dashed vertical line and the 95% confidence intervals represented by horizontal segments.The narrow width of the confidence intervals indicates high stability and precision in the parameter estimates, whereas wider intervals suggest greater variability and potential uncertainty.

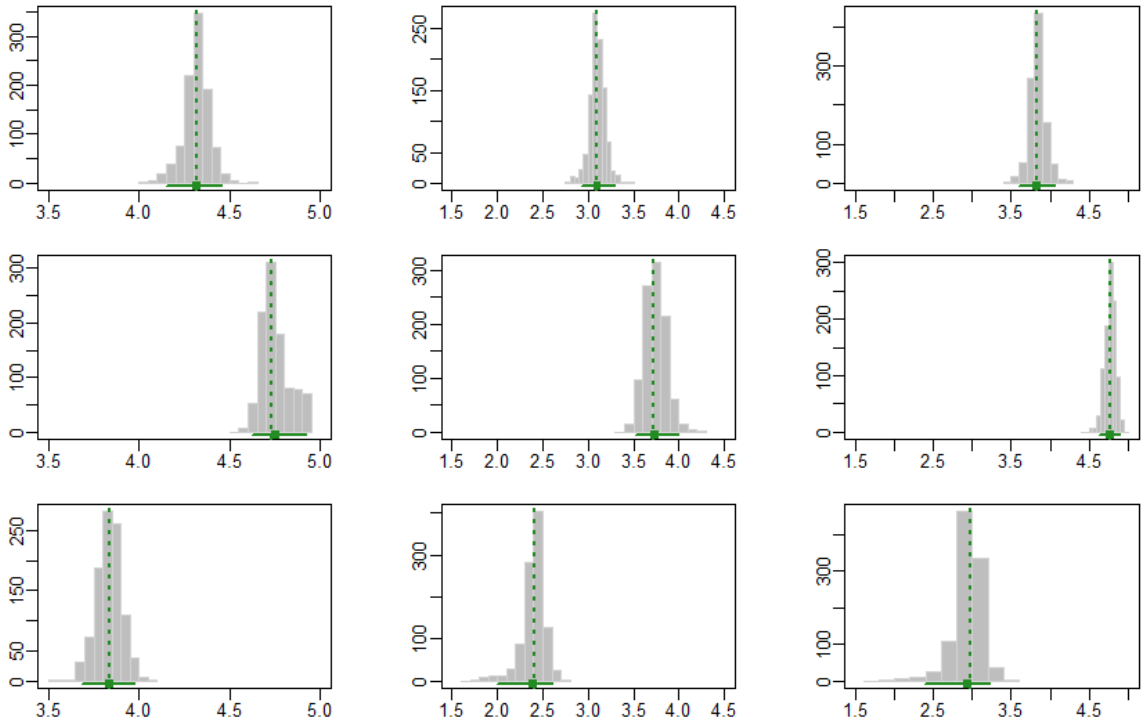


Figure 12: Each panel represents the bootstrap distribution of a component mean parameter. The three rows correspond to the three Gaussian clusters and the three columns correspond to the Behavioural, Cognitive, and Emotional engagement variables. The histograms represent the frequency of these values across the bootstrap samples. Green dashed vertical lines indicate the estimated mean parameter and the solid green horizontal lines indicate the 95% confidence intervals.

The plot in Figure 12 visualizes the bootstrap distribution of entropy for different clusters. By examining the spread of the bootstrap distributions using the `"bs"` method of the `MclustBootstrap()` function, we can evaluate the stability of the parameter estimates. Narrower distributions imply stable estimates, whereas wider distributions indicate variability. The presence of outliers can be detected if certain bootstrap samples consistently yield at extreme values. However, recall that the underlying model already also includes an additional uniform noise component for capturing outliers and has already had its number of components selected using BIC.

From the plot, we can observe that most clusters have a relatively low entropy, indicating that the cluster assignments are generally well-defined. However, some clusters exhibit a wider spread in entropy values, suggesting varying degrees of uncertainty in those cluster assignments. This visualization helps in assessing the stability and reliability of the clustering results, as clusters with consistently low entropy across bootstrap samples are considered more stable and robust.

In the analysis of latent profiles using Gaussian mixture models (GMM), it is crucial to accurately estimate the means of the observed variables within each profile and the mixing proportions that denote the prevalence of each profile in the population. The nonparametric bootstrap method offers a robust approach to derive these estimates along with their confidence intervals, enhancing the reliability of the findings.

Table 5: This table presents the estimated means of various behavioral and emotional variables for each latent profile, along with the lower and upper bounds of their 95% confidence intervals derived from the nonparametric bootstrap method.

Profile	Variable	Mean	Lower	Upper
1	Behavioural	4.32	4.16	4.46
1	Cognitive	3.10	2.93	3.30
1	Emotional	3.82	3.63	4.06
2	Behavioural	4.73	4.63	4.92
2	Cognitive	3.72	3.53	4.02
2	Emotional	4.77	4.62	4.90
3	Behavioural	3.84	3.70	3.96
3	Cognitive	2.41	2.02	2.60
3	Emotional	2.97	2.37	3.22

Table 5 presents the estimated means of various behavioral and emotional variables across three latent profiles. The inclusion of 95% confidence intervals for each mean provides an assessment of the precision of these estimates.The confidence intervals provide an understanding of the precision of these mean estimates, indicating the range within which the true mean values lie with 95% confidence. The distinct mean values for each profile underscore the heterogeneous nature of the population under study, suggesting that individuals within each profile exhibit unique behavioral and emotional characteristics.

Table 6: This table displays the mixing proportions of each latent profile, including the noise component, with their respective 95% confidence intervals derived from the nonparametric bootstrap method.

Profile	MixPro	Lower	Upper
1	0.467	0.349	0.624
2	0.143	0.074	0.207
3	0.343	0.160	0.484
0	0.047	0.015	0.085

Table 6 illustrates the mixing proportions for each latent profile, including a noise component, with their respective 95% confidence intervals. These proportions indicate the relative prevalence of each profile in the overall sample. The confidence intervals offer insight into the stability and reliability of these proportions, underscoring the robustness of the latent profile classification.

These results indicate that the smallest cluster (cluster 2) has the highest engagement scores for all three variables. Conversely, the largest cluster (cluster 3) has lower scores for all three engagement variables. Cluster 1 exhibits the lowest mean scores, particularly for the cognitive engagement attribute. For cluster 2, the behavioral and emotional engagement scores are comparable, whereas, for the other two profiles, the mean scores for these attributes are lower than those for behavioral engagement.

The parallel coordinate plot shown in Figure 13 summarises the mean vectors of the three Gaussian clusters, excluding the observations designated as outliers in the noise component. This plot provides a visual representation of the engagement profiles and their respective confidence intervals, represented as vertical bars.

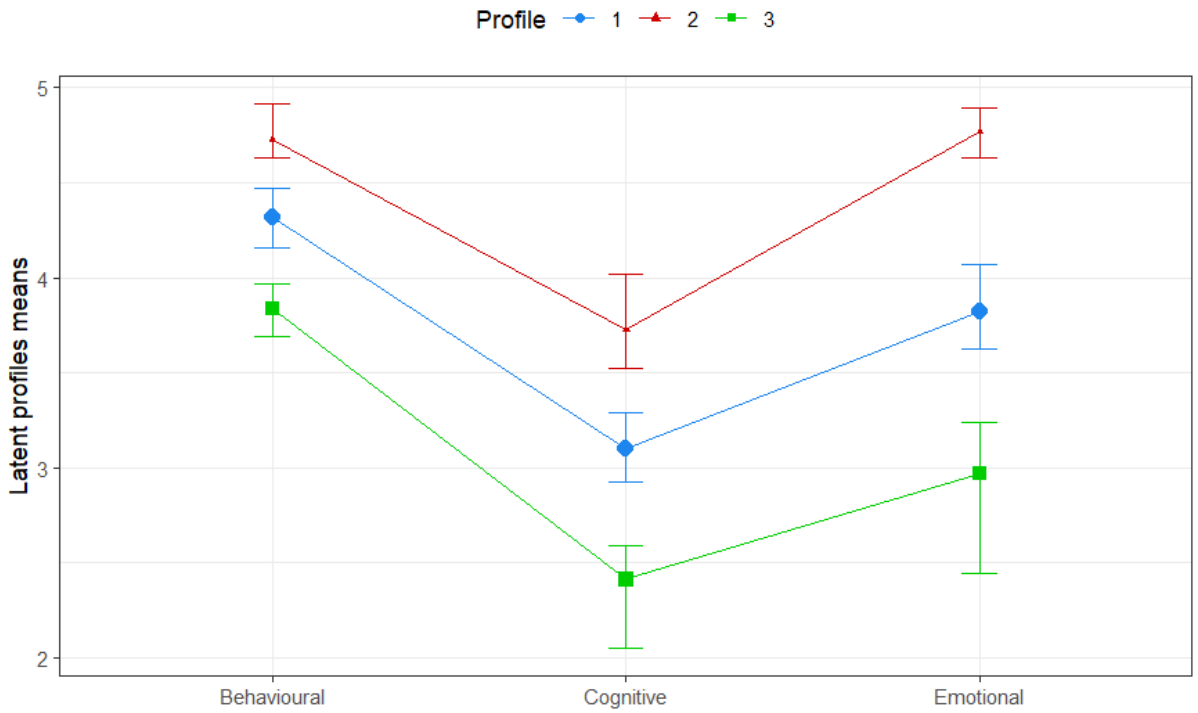


Figure 13: Parallel coordinate plot of the three profile means including the error bars that represent the confidence intervals for each mean value.

The analysis of the engagement profiles reveals distinct patterns across the three clusters. Cluster 2, the smallest cluster, demonstrates the highest engagement scores across behavioral, cognitive, and emotional dimensions. Conversely, the largest cluster, cluster 3, shows the lowest scores in these areas, with cluster 1's scores falling in between. Notably, cognitive engagement scores are the lowest across all clusters For cluster 2, behavioral and emotional engagement scores are relatively similar, whereas for clusters 1 and 3, emotional engagement scores are lower than behavioral engagement scores. Based on these observations, clusters 1, 3, and 2 can be characterised as "low," "medium," and "high" engagement profiles, respectively.

The parallel coordinate plot visually corroborates the findings from the bootstrap analysis. It highlights the distinct engagement levels across the three clusters and provides a clear visual differentiation between them. The addition of confidence intervals to the plot enhances the interpretability by illustrating the precision of the estimated means. These analyses collectively strengthen the understanding of the latent profiles in the data, providing a comprehensive view of the engagement patterns and the associated uncertainties. The figure also shows that none of the confidence intervals overlap which indicates that there is a statistically significant difference between clusters.

To have a comprehensive evaluation of the clustering model's stability and reliability further we would perform entropy and average posterior probability analyses post-bootstrapping.This helps in identifying areas of uncertainty and potential outliers, ensuring that the final clustering solution is both meaningful and dependable. By integrating these measures, we can confidently interpret and apply the clustering results, making informed decisions based on a validated and robust model.

Entropy measures the uncertainty associated with the assignment of observations to clusters. Lower entropy values indicate higher confidence in cluster assignments, while higher values suggest greater uncertainty. The Table 7 is related to entropy presenting a detailed examination of the clustering model, emphasizing the case-specific entropy contributions across different latent profiles.

Table 7: Entropy contributions by cluster, where 0 corresponds to the noise component.

Clusters	Mean	SD	Min	Max
0	0.69	0.17	0.50	0.99
1	0.65	0.16	0.15	0.91
2	0.69	0.18	0.19	0.93
3	0.66	0.16	0.25	0.93

The entropy formula is modified to treat the quantity G as the number of Gaussian clusters *plus* the extra uniform noise cluster. The overall entropy value of clusters is 0.66, which ensures that the solution is acceptable as 0.6 is a good rule of thumb for determining the acceptance of a solution according to the enropy measure.

Findings from the Figure 14 and Table 7 determines about each cluster, cluster 1 exhibits the widest spread of entropy values, indicating a higher degree of uncertainty in cluster assignments. The presence of cases with higher entropy suggests that some observations within this profile may not fit as well into the defined cluster structure. Cluster 2 shows a relatively narrow distribution of entropy values, with most cases concentrated towards the lower end. This indicates greater confidence in the cluster assignments, reflecting the distinctiveness and cohesiveness of this profile. Cluster 3 is similar to cluster 2, it also shows a lower range of entropy values, though slightly more spread out than cluster 2. This suggests a reasonable level of confidence in cluster assignments but with more variability compared to cluster 2.

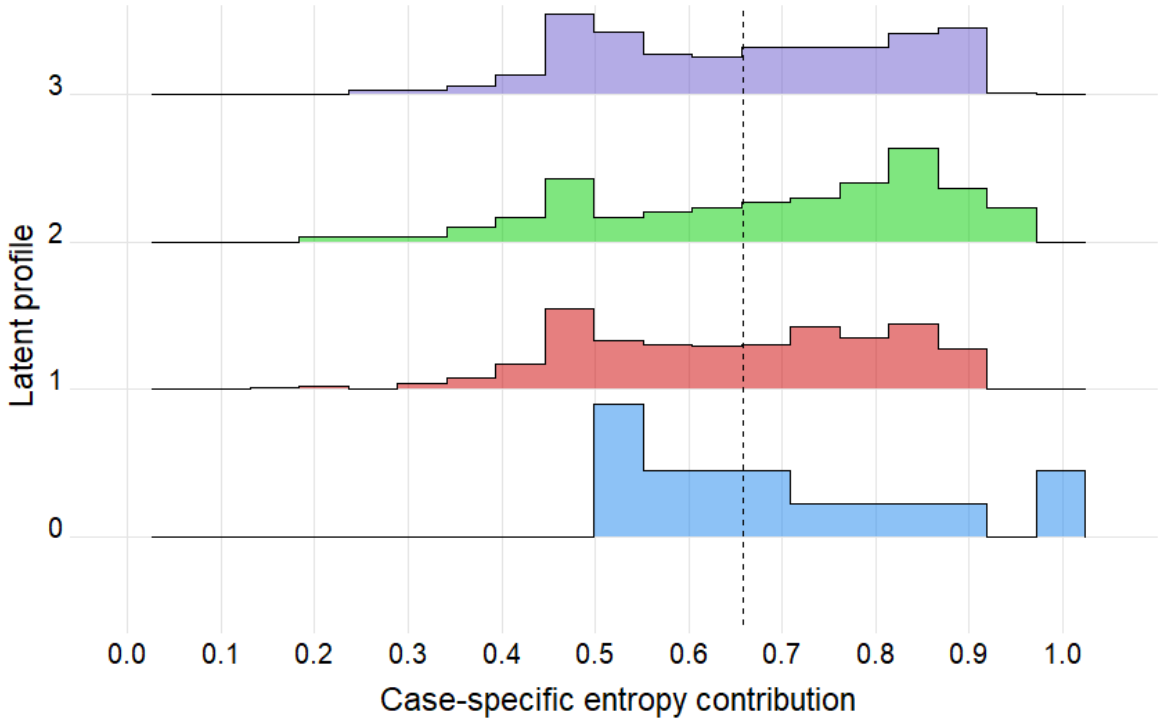


Figure 14: This plot illustrates the case-specific entropy contributions across different latent profiles, where 0 corresponds to the noise component. The dashed line represents the total entropy, providing a benchmark for comparing individual cluster contributions. Entropy measures the uncertainty in cluster assignments, with lower values indicating more definitive classifications.

The average posterior probability measures the likelihood that each observation belongs to its assigned cluster, with higher values indicating stronger cluster membership. Table 8 and Figure 15 are related to the average posterior probability that provide insights into the strength of the membership probabilities for each observation within the latent profiles.

Table 8: Average posterior probabilities by cluster, where 0 corresponds to the noise component.

Clusters	Mean	SD	Min	Max
0	0.79	0.16	0.54	0.99
1	0.81	0.14	0.46	0.98
2	0.84	0.14	0.43	0.98
3	0.80	0.15	0.47	0.98

These statistics are also indicative of a satisfactory clustering solution, given that 0.8 is typically recommended as a threshold for assessing a partition's quality when using the average posterior probability measure.

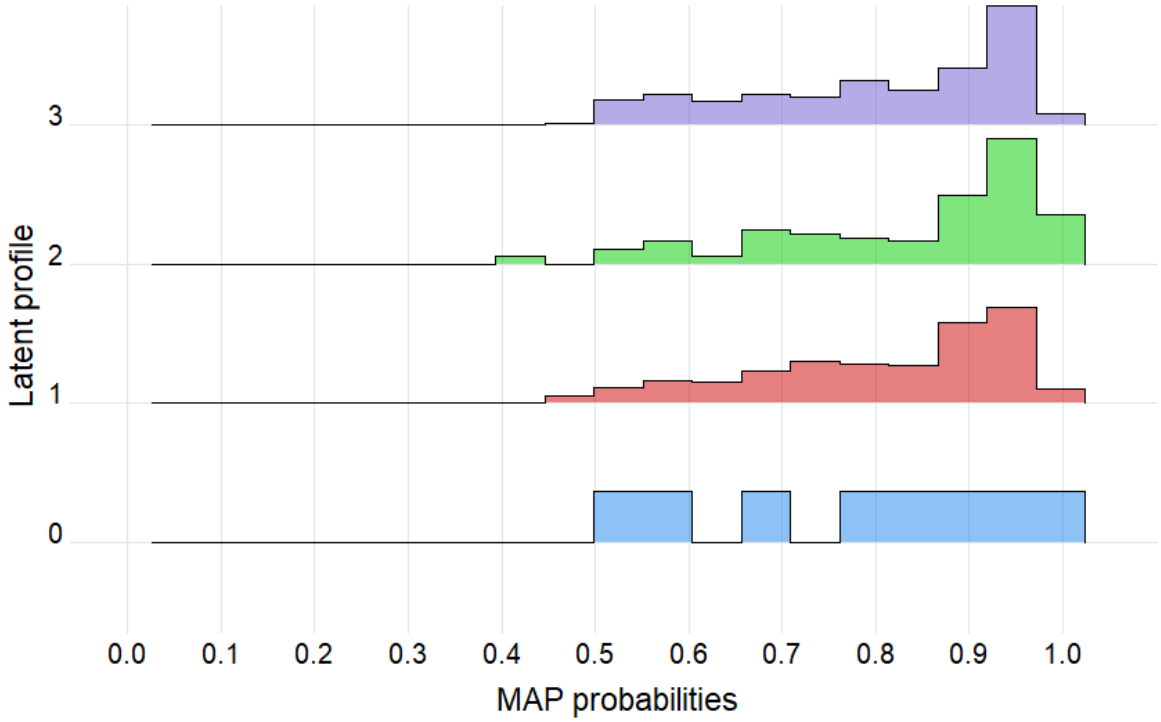


Figure 15: This plot visualises the distribution of average posterior probabilities for each latent profile, where 0 corresponds to the noise component. The density ridges are binned and scaled to highlight the probability concentration within each profile, illustrating the degree of membership certainty for the observations in each cluster.

As discussed before, for the purpose of these calculations, G is given by the total number of components *plus* the extra uniform noise cluster, i.e., accounting for both the Gaussian and non-Gaussian clusters. Table 8 accounts for the stability of the clusters and 0.8 is considered a good rule of thumb for average posterior probability and the overall value for our clusters is 0.81 which states our cluster is stable and acceptable.

This observation is particularly valuable for our findings as cluster 1 shows a broad distribution of posterior probabilities, with many cases having lower probabilities. This indicates weaker membership strength and suggests that some observations in this profile might not fit well within the defined cluster. Cluster 2 displays a higher concentration of posterior probabilities towards the upper end of the scale, signifying strong membership probabilities. This reinforces the distinctiveness and cohesiveness observed in the entropy analysis. Cluster 3's distribution is somewhat intermediate, with posterior probabilities spread across a wider range compared to cluster 2 but still indicating generally strong membership.

By integrating these analyses, we obtain a comprehensive understanding of the clustering model's performance. This approach helps to identify and address potential issues, thereby enhancing the overall validity and interpretability of the clustering results. This detailed examination supports the robustness of the clustering solution and ensures that the identified profiles are both distinct and meaningful, which is essential for the subsequent interpretation and application of the results.

5.2 Relating latent profiles to covariates

In this section, we explore the clusters identified through the model-based clustering processes described earlier. Using the optimal model, we tackle our secondary objective of examining how these clusters formed on engagement measures relate to covariates related to academic achievement, demographics, and self-regulation variables. Given the small number of observations assigned to the noise component, we focus our attention only on the three Gaussian VVI clusters throughout. By analysing the clusters' distribution across different performance metrics, we aim to uncover patterns and insights that highlight the distinct characteristics and behaviors of each cluster. This analysis will provide a deeper understanding of the heterogeneity among learners and how these subgroups differ in their academic achievements and engagement strategies.

The plot in Figure 16 presents the distribution of Mathematics grades across three distinct clusters, where each cluster shows the count of students achieving grades on a Likert scale from 1 to 5. Cluster 1 has a significant number of students with grades 3 and 5, indicating a varied performance with a notable concentration at the higher end of the scale. In Cluster 2, the majority of students scored 3 and 4, demonstrating a relatively consistent performance with fewer lower-end grades. Cluster 3 exhibits a higher number of students scoring 5, followed by grades 3 and 4, suggesting that this cluster tends to perform better in Mathematics compared to the others.

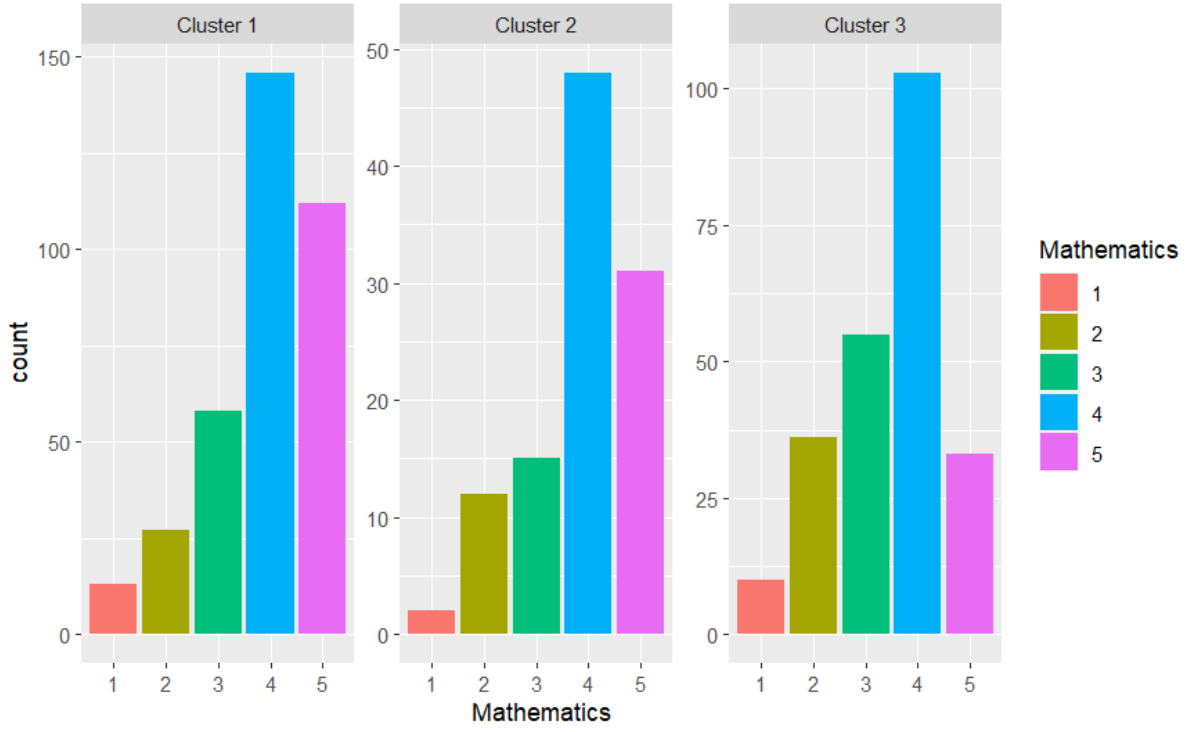


Figure 16: This figure illustrates the distribution of Mathematics grades across three clusters. Cluster 1 has a significant number of students achieving grades 3 and 5, showing varied performance. Cluster 2 has a majority scoring 3 and 4, indicating consistent performance with fewer lower-end grades. Cluster 3 shows a higher number of students achieving grade 5, suggesting this cluster tends to perform better in Mathematics.

Figure 17 illustrates the distribution of gender within each cluster. Cluster 1 has a balanced representation of male and female students, though slightly more females. Cluster 2 has a higher count of female students compared to male students, indicating a gender imbalance. Cluster 3 shows a more balanced gender distribution, similar to Cluster 1 but with fewer students overall. One student with an unknown gender has been discarded from this plot, but we note that they were assigned to cluster 3 by the model.

Figure 17: This figure shows the gender distribution within each cluster.

By interpreting these plots, we can gain insights into the distinct characteristics of each cluster, helping to tailor educational interventions to meet the diverse needs of students.

Figure 18 presents box plots of environment management scores across the clusters. Each box on the plot represents a cluster, with the horizontal line inside the box indicating the median score for that group. The box itself encompasses the middle 50% of the data, while the whiskers extend to capture the majority of the data points, excluding outliers. Students with unknown environment management scores have been discarded from this plot.

Cluster 2 consistently demonstrates the highest median environment management score, suggesting that, on average, this group exhibits superior performance in this area compared to the other clusters. In contrast, cluster 1 has the lowest median score. While the spread of scores, as indicated by the IQR, is relatively similar across all three clusters, there are some notable exceptions. Cluster 1, in particular, contains a few outliers, implying the presence of significantly higher or lower environment management scores within this group compared to the rest of the data.

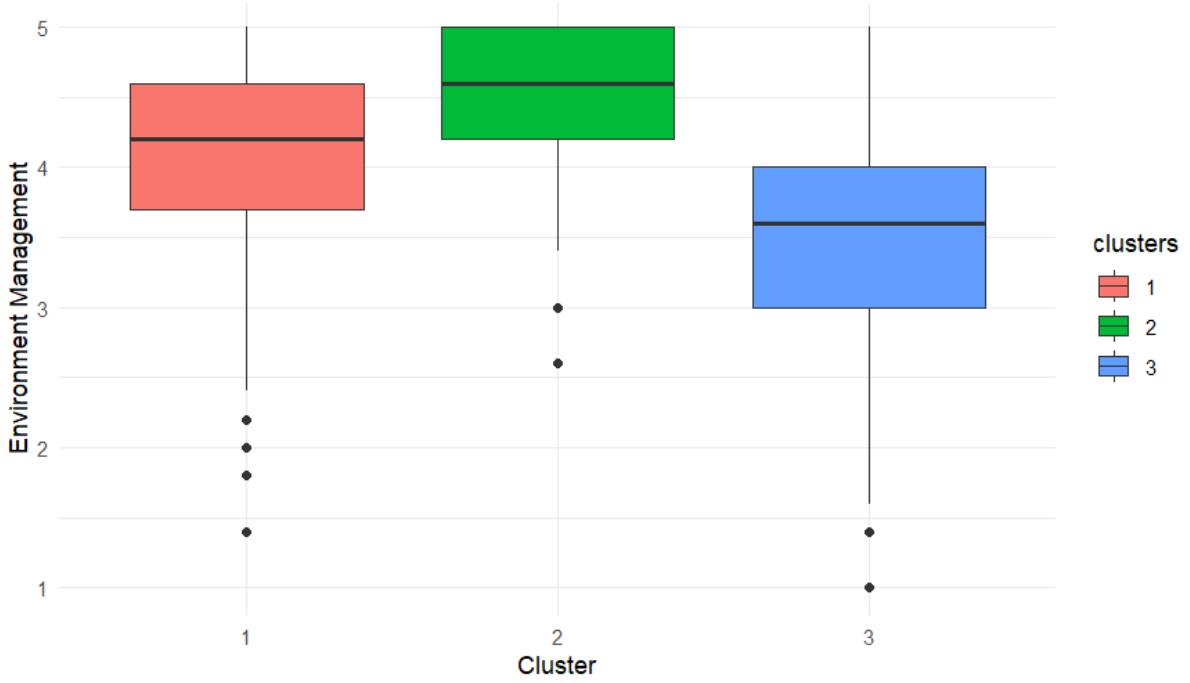


Figure 18: This figure presents box plots of environment management scores across the clusters.

Figure 19 is similar to Figure 18 but with time management in place of the environment management covariate. Again, students with unknown time management scores have been removed. This visualisation also reveals notable differences between the clusters. Cluster 2 exhibits the highest median time management score, suggesting that individuals in this group generally demonstrate better time management skills compared to those in Clusters 1 and 3. Additionally, the box for Cluster 3 is wider than the others, indicating a greater spread or variability in time management scores within that group.

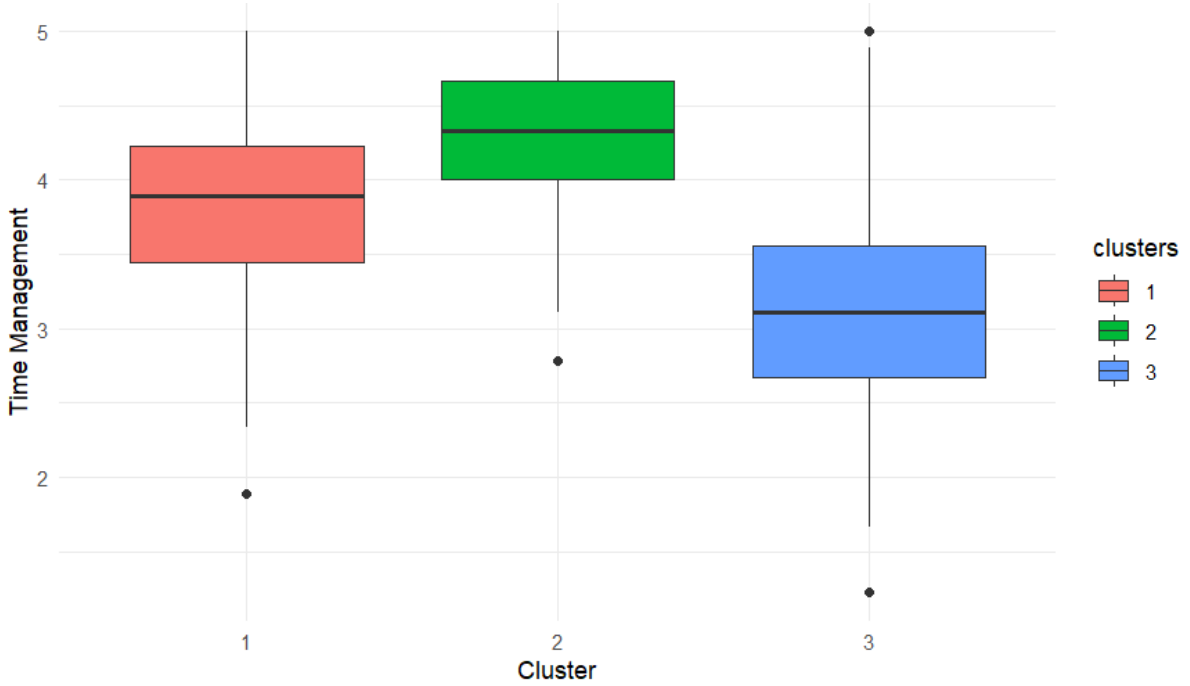


Figure 19: This figure presents box plots of time management scores across the clusters.

These findings underscore the importance of recognising and addressing the distinct profiles of each cluster. Cluster 3's strong performance in Mathematics and Cluster 2's superior time management skills highlight areas where these groups excel. Conversely, the lower time management scores and varied Mathematics performance in Cluster 1 indicate areas needing improvement. The gender distribution insights also suggest the need for gender-sensitive approaches in educational interventions.

By understanding these clusters' unique characteristics, educators and policymakers can design targeted strategies to enhance learning outcomes, promote balanced skill development, and ensure equitable educational opportunities for all students. This holistic approach will help maximise the potential of each student group, fostering an inclusive and effective learning environment.

6 Discussion

In this thesis, we have demonstrated that model-based clustering serves as an effective tool for analyzing and interpreting complex educational data. Our primary objective was to identify meaningful clusters of students based on their engagement measures and academic performance. Due to the nature of the data, which included a high number of duplicate values for the observed engagement measures, it was necessary to explore advanced features of the `mclust` library to derive a workable and interpretable solution.The model employed prior distributions for regularisation and incorporated a noise component and parsimonious constraints on the component-specific covariance matrices. By doing so, a model with superior BIC to the LPA model previously used to analyse these data was found.

The inclusion of a noise component played a pivotal role in achieving superior results. This component allowed us to account for inherent variability and measurement errors within the data, leading to more accurate and reliable clustering outcomes. This adjustment facilitated the differentiation between true cluster structures and random noise, ensuring that the identified clusters were robust and meaningful. It also helps to distinguish highly atypical students who may be problematic in a classroom setting.

Ultimately, we found three Gaussian clusters of engagement profiles, corresponding to low, medium, and high engagement. Interestingly, the highly-engaged cluster had the highest average engagement scores across all three engagement dimensions. Similarly, the cluster with low engagement had the lowest average behavioural, cognitive, and emotional engagement scores. Within each cluster, students tended to perform worst in terms of cognitive engagement.

The utilisation of a bootstrap procedure significantly enhanced the stability and reliability of our clustering results. By generating multiple bootstrap samples and performing clustering on each sample, we were able to assess the variability and consistency of the cluster solutions. This approach provided a more comprehensive understanding of the clustering structure and ensured that our results were not overly dependent on a single sample.

Entropy and average posterior probability emerged as critical metrics in evaluating the quality of our clustering solution. Entropy measures the uncertainty or randomness within the cluster assignments,

with higher entropy indicating more well-defined clusters. The average posterior probability indicates how confidently each data point is assigned to a cluster, with higher values suggesting greater confidence in the assignments. By optimising these metrics, we ensured that our final clustering solution was both stable and interpretable.

A secondary aim was to relate the clusters formed on the engagement measures to available covariates pertaining to demographics, academic achievement, and self-regulation. In light of this, the identified clusters can be characterised as follows. Cluster 1 includes a large mix of highly-engaged students, with many achieving high grades in Mathematics. The gender distribution is balanced, and students in this cluster generally have strong time management skills. Cluster 2 is smaller and primarily consists of students with average Mathematics grades, showing consistent academic performance. This cluster has more female students and the highest time management and environment management scores, indicating strong organisational skills. Cluster 3 contains students who perform well in Mathematics, but they have the lowest time management scores, suggesting variability in their organisational abilities. The gender distribution in this cluster is balanced, similar to Cluster 1, but with fewer students.

In summary, our findings underscore the utility of model-based clustering in uncovering distinct student profiles based on engagement and performance measures. These profiles can provide valuable insights for educators and policymakers to tailor interventions and support strategies to meet the diverse needs of students. By understanding the characteristics and needs of each cluster, educational institutions can develop targeted programs to enhance student engagement, improve academic performance, and foster a more supportive learning environment. The incorporation of noise components, prior distributions, bootstrap sampling, and the evaluation of entropy and average posterior probability were instrumental in achieving a stable and interpretable clustering solution.

6.1 Ideas for future work

In future research, there are several potential avenues to explore that could enhance and expand upon the current findings. One promising direction is to incorporate covariates directly into the model using the `MoEClust` package for `R` (Murphy and Murphy 2020). By including covariates in the modeling process, we can account for additional variables that may influence the clustering results allow the covariates to explicitly guide the construction of the clusters. This approach would allow for a richer understanding of the relationships between engagement measures, academic performance, and other influential factors.

Another potential avenue is to treat the covariates as response variables and jointly cluster all data using methods designed for model-based clustering of mixed-type data, using for example the `R` package `clustMD` (McParland and Gormley 2016). This method allows for simultaneous consideration of both categorical and continuous data, providing a more holistic view of the data structure. By jointly clustering engagement measures, academic performance, and the other variables presently treated instead as covariates, this approach can uncover deeper insights into how these different types of data interact and contribute to the overall clustering solution.

Lastly, future work could explore the use of mixtures of t-distributions instead of mixtures of Gaussians to account for outliers and non-normality, rather than relying on a noise component. The `R` package `teigen` (Andrews and McNicholas 2012) facilitates this approach, which is particularly useful for datasets with heavy tails or outliers. Mixtures of t-distributions are more robust to deviations from normality and can provide a better fit for data with significant outliers than Gaussian mixture models. The `teigen` package also includes a family of constraints on the covariance matrices, in the same spirit as GPCMs. This method could lead to more accurate and reliable clustering results by more appropriately modeling the underlying distribution of the data.

Exploring these directions in future research will not only validate and extend the current findings but also contribute to the development of more sophisticated and robust clustering methods for complex educational data.

Bibliography

Andrews, J. L., and McNicholas, P. D. (2012), "Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions," *Statistics and Computing*, 22, 1021–1029.

Azevedo, R. (2015), "Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues," *Educational Psychologist*, Taylor & Francis, 50, 84–94.

Banfield, J. D., and Raftery, A. E. (1993), "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, JSTOR, 803–821.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997), "Inference in model-based cluster analysis," *Statistics and Computing*, Springer, 7, 1–10.

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 22, 719–725.

Cleary, T. J. (2006), "The development and validation of the self-regulation strategy inventory–self-report," *Journal of School Psychology*, 44, 307–322.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, 39, 1–22.

Efron, B. (1992), "Bootstrap methods: Another look at the jackknife," in *Breakthroughs in Statistics: Methodology and distribution*, Springer, pp. 569–593.

Estévez, I., Rodríguez-Llorente, C., Piñeiro, I., González-Suárez, R., and Valle, A. (2021), "School engagement, academic achievement, and self-regulated learning," *Sustainability*, 13, 3011.

Fraley, C., and Raftery, A. E. (2002), "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, Taylor & Francis, 97, 611–631.

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012), *Mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation*, Department of Statistics, University of Washington.

Fredricks, J. A., Blumenfeld, P., Friedel, J., and Paris, A. (2005), "School engagement," in *What do children need to flourish? Conceptualizing and measuring indicators of positive development*, eds. K. A. Moore and L. H. Lippman, Boston, MA, U.S.A.: Springer, pp. 305–321. https://doi.org/10.1007/0-387-23823-9_19.

García-Escudero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., and Mayo-Iscar, A. (2018), "Eigenvalues and constraints in mixture modeling: Geometric and computational issues," *Advances in Data Analysis and Classification*, Springer, 12, 203–233.

Hennig, C., and Coretto, P. (2008), "The noise component in model-based cluster analysis," in *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation eV, Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*, Springer, pp. 127–138.

Jung, T., and Wickrama, K. A. (2008), "An introduction to latent class growth analysis and growth mixture modeling," *Social and Personality Psychology Compass*, Wiley Online Library, 2, 302–317.

McLachlan, G. J. (2011), "Commentary on Steinley and Brusco (2011): Recommendations and cautions." American Psychological Association.

McLachlan, G. J., and Krishnan, T. (2007), *The EM Algorithm and Extensions*, John Wiley & Sons.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019), "Finite mixture models," *Annual Review of Statistics and its Application*, Annual Reviews, 6, 355–378.

McLachlan, G. J., and Rathnayake, S. (2014), "On the number of components in a Gaussian mixture model," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, 4, 341–355.

McParland, D., and Gormley, I. C. (2016), "Model based clustering for mixed data: clustMD," *Advances in Data Analysis and Classification*, 10, 155–170.

Melnykov, V., and Melnykov, I. (2012), "Initializing the EM algorithm in gaussian mixture models with an unknown number of components," *Computational Statistics & Data Analysis*, Elsevier, 56, 1381–1395.

Murphy, K., and Murphy, T. B. (2020), "Gaussian parsimonious clustering models with covariates and a noise component," *Advances in Data Analysis and Classification*, 14, 293–325. https://doi.org/10.1007/s11634-019-00373-8.

Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society: Series B (Methodological)*, Oxford University Press, 56, 3–26.

Olivier, E., Galand, B., Hospel, V., and Dellisse, S. (2020), "Understanding behavioural engagement and achievement: The roles of teaching practices and student sense of competence and task value," *British Journal of Educational Psychology*, Wiley Online Library, 90, 887–909.

R Core Team (2024), "R: a language and environment for statistical computing," Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D. B. (1981), "The Bayesian bootstrap," *The Annals of Statistics*, JSTOR, 130–134.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016), "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models," *The R Journal*, NIH Public Access, 8, 289.