Employing latent profile analysis to identify student motivational profiles

by

Pauline Huisman

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Tuesday July 8, 2021 at 15:45 AM.

Student number:	4933095	
Project duration:	April 22, 2021 – July 8, 2021	
Thesis committee:	Dr. A. J. Cabo,	TU Delft, supervisor
	L. Y. J. Wong,	TU Delft, supervisor
	Prof. dr. ir. C. Vuik,	TU Delft

Abstract

Latent profile analysis is a statistical modeling approach used to identify hidden subpopulations (i.e., latent profiles) within a population. These latent profiles are identified based on values of observed continuous variables, also known as profile indicators. While LPA is getting more popular in education sciences and psychology to group people based on similar characteristics, very little is known about the mathematical formulation. In this thesis, the mathematical foundations of LPA is introduced and explained. This leads to a discussion on the assumptions for the model.

After investigating the mathematical foundations of LPA, we applied LPA to identify different profiles of motivation in a student population at Delft University of Technology. We used a set of survey data measuring four types of motivation (i.e., profile indicators). Results of the analysis showed that there are four different student motivational profiles, each consisting of a different combination of the four types of motivation.

Preface

This research has been conducted under supervision of Dr. A. J. Cabo and L.Y.J. Wong on behalf of the department of Statistics of the faculty EEMCS at the University of Technology Delft.

I would like to express my special thanks to my supervisors Dr. A. J. Cabo and L.Y.J. Wong for their guidance during this project. They have always reassured me and were always open to questions. I also would like to thank Prof. dr. ir. C. Vuik for taking a seat in my thesis committee.

Pauline Huisman
Delft, July 2021

Contents

1 Introduction

The PRogramme of Innovation in Mathematics Education (PRIME) is developing mathematics courses for engineering programmes at the TU Delft. In order to measure effects of these innovations, statistical research is conducted in PRIME. Part of the research in PRIME is about the motivation of students to study for a mathematics course. This thesis is about this motivational part of the research in PRIME. We aim to identify different profiles of motivation in student populations.

Identifying hidden profiles within a population can be done in multiple ways by using different types of cluster analyses. There are many types of cluster analysis and they often have the same goal: to correctly classify similar cases into one of the subgroups. Individuals in a cluster tend to be more similar to each other than unrelated randomly selected individuals (Fox, 2016). Some examples of cluster analyses are k-means clustering, latent class analysis (LCA) and latent profile analysis (LPA). In section 2.1 the methods k-means clustering and LCA are shortly explained. This thesis will be about the clustering method LPA, hence we aim to identify different profiles of motivation using LPA.

Latent profile analysis (LPA) is a statistical modeling approach used to identify hidden profiles within a population. These latent profiles are identified based on values of observed continuous variables, also known as profile indicators. Specifically, LPA models unobserved population heterogeneity by grouping individuals into latent profiles based on similarities in their response variable scores. (Peugh and Fan, 2013). The individuals are assigned to a profile by using the probability of membership. They are assigned to the profile for which their probability is the highest. This is why LPA is a model-based (person-centered) approach.

LPA is a type of latent variable mixture model. Latent variable modeling refers to multiple statistical procedures that use one or more (unobserved) latent variables to investigate relationships between a larger set of observed variables. The latent variable refers to the latent categorical variable of cluster membership, hence this cannot be measured directly from the data. The term mixture in the latent variable mixture model refers to the fact that the data is being sampled from a population composed of a mix of probability distributions. From this mixture of probability distributions, each probability distribution belongs to a profile, and each profile distribution is characterized by its own unique set of parameters consisting of means and covariances. In LPA it is assumed that the distributions in the mixture are normally distributed. (Pastor et al., 2007)

Researchers have increasingly used LPA in recent years in different fields, for example in criminology, education, marketing and psychology (Tein et al., 2013). Since LPA is mostly used in these fields, very little is known about the mathematical formulation of LPA. Therefore in this thesis we will give an overview of the mathematics that is used in LPA to form latent profiles.

Thesis outline

In this research, statistics is applied to educational research. Before we apply the statistical method LPA on actual data sets, it is important for us to have a good understanding on how LPA works and the mathematics involved to identify latent profiles. By doing so, we will be able to better interpret the results from the analysis and discuss the implications for educational research. In chapter 2, the important mathematical model equation used in LPA is given and this equation is investigated. With this equation the latent profiles are eventually created by estimating the unique model parameters (e.g., means and covariances) which characterize the profile distributions. Furthermore, the equation provides estimates for the probabilities of an individual belonging to a profile.

Chapter 3 presents the research on motivational profiles in PRIME by first explaining what PRIME is and giving a brief explanation of the different types of motivations according to the fields of education and psychology. In this chapter we will also bring up our main research question: *'What are the student profiles that can be distinguished based on students' motivation and are these profiles significantly different?'*. To answer this question we will apply LPA to the PRIME data set. To determine which model of the LPA gives the best model fit we will use several fit indices. When the best model is determined and the individuals are assigned to the profiles, we will check if these profiles are significantly different by applying a statistical method.

We will also look at some previous studies on LPA. To compare our results with the results of the previous studies we have a small research question. This sub research question is *'Are the identified profiles comparable to those found in previous studies?'*.

2

Latent Profile Analysis

In this chapter we will first explain a few examples of clustering methods. Then we will take a close look into the mathematics that is used to create the latent profiles. One equation, the LPA model equation, is central to the mathematical formulation of LPA. This LPA model equation is used to find estimates for the unique parameters per profile. To find these estimates from the LPA model equation we have to use the Expectation-Maximization (EM) algorithm.
In addition, certain specifications are defined to create the models, namely the number of profiles and how the variables are related to one another (e.g., whether the variances and covariances are allowed to vary between the profiles). To choose which specification gives the best model fit, we use multiple fit indices.

2.1. Clustering methods

As mentioned before in the introduction, there are several types of clustering methods. We will shortly explain three examples of the clustering methods, k-means cluster analysis, latent class analysis (LCA) and latent profile analysis (LPA). We will eventually use LPA in this thesis, so therefore we will also compare the k-means cluster analysis and LCA with LPA.

2.1.1. K-means cluster analysis

One of these modeling techniques is the k-means cluster analysis. This analysis is similar to LPA since it also groups participants into categories based on response variable score similarities. But there are some clear differences between the two. At first, there is some difference in the latent profile membership determinations. In the k-means cluster analysis, each individual is assigned to one and only one cluster. An indicator shows if an individual is (1) or is not (0) member of a cluster. This is seen as hard clustering. On the other hand, latent profile analysis is a form of soft clustering. In LPA, each point is assigned to all the clusters with different probabilities. So the latent profile membership is estimated as a probability conditional on a participant's response variable scores. The individuals are assigned to the profile for which their probability is highest. (Peugh and Fan, 2013)
Another big difference between LPA and k-means cluster analysis contains the unique set of the parameters means and covariances for each profile. Unlike k-means cluster analysis, in LPA is it allowed to vary the variances and covariances across all the latent profiles.

2.1.2. Latent class analysis (LCA)

The clustering method that is most similar to LPA is latent class analysis (LCA). Where LPA tries to revover hidden groups based on the means of continuous observed variables, LCA does the same for categorical variables. Another difference is that in LCA there is no assumption that the variables are distributed in any particular way. In LCA you assume that within each class, the observed variables are unrelated to each other (Oberski, 2016).

2.1.3. Latent profile analysis (LPA)

LPA is a statistical modeling approach used to identify hidden profiles within a population. These latent profiles are identified based on values of continuous observed variables, also known as the profile indicators. The individuals are assigned to a profile by using the probability of membership. They are assigned to the profile for which their probability is the highest. This is why LPA is a model-based (person-centered) approach.

LPA is under the mathematicians mostly known as gaussian mixture model (Oberski, 2016). The name latent profile analysis is more commonly used in the social sciences, therefore we will use LPA in this thesis. This research is about this clustering method, the latent profile analysis.

2.2. Model description

The data that is analysed in LPA is sampled from a population that consists of a mix of distributions. Each of these distributions belongs to a profile. So it is the goal of LPA to identify which individuals belong to which profile distribution. Each profile distribution is characterized by its own unique set of parameters consisting of means and covariances. This set of parameters is unknown, as is the probability of belonging to a profile (profile membership). LPA will estimate these unknowns in order to from the latent profiles. To estimate these parameters some assumptions are made, see Spurk et al., 2020.

2.2.1. Assumptions

1. Within each latent profile the continuous indicators are normally distributed.

2. Unobserved heterogeneity. Unobserved population heterogeneity occurs when the variables that cause the heterogeneity can not be observed directly from the data. In this case, the subpopulations are latent and must be derived from the data.

3. The data of each individual is independent of the other individuals. We also assume that the profiles are independent of one another. This means that every profile consists of different people.

To illustrate how LPA works mathematically, Pastor et al., 2007 provide the following concrete example. In this example you want to identify profiles in a population based on a single factor. Suppose also that the population consists of two different profiles of persons, so this means in the latent profile analysis two different types of distributions. By the first assumption in 2.2.1, these two distributions are assumed to be normal. In general the number of profiles in not known beforehand.

As mentioned before, LPA provides estimates for the probability of belonging to a profile and for the unique set of parameters consisting of the means and covariances per profile. Say the continuous single indicator of cluster membership for person i is y_i, with a population of N individuals ($i = 1, \dots, N$). And assume that parameters μ_1 and σ_1^2 could be estimated for profile 1, parameters μ_2 and σ_2^2 could be estimated for profile 2 and that the probabilities of membership, which also can be seen as the weights given to each profile, are π_1 for belonging to profile 1 and π_2 for profile 2. Then the model for this example is represented using the following equation:

$$f(y_i|\theta) = \pi_1 f_1(y_i|\mu_1, \sigma_1^2) + \pi_2 f_2(y_i|\mu_2, \sigma_2^2), \qquad i = 1, \dots, N \tag{2.1}$$

From this equation we see that the distribution f of the cluster indicator given the model parameters $\theta = (\pi_1, \mu_1, \sigma_1^2, \pi_2, \mu_2, \sigma_2^2)$, is a linear combination of two different probability distributions (f_1 and f_2) with their own unique set of parameters.

This was a concrete example of LPA of the univariate model, since there was only one single indicator. When there is more than one continuous indicator we have the case of a multivariate model. In this case the multiple indicator variables for person i are contained in the vector $\mathbf{y}_i$.
Another extension of the univariate model to get the multivariate model is that now the distribution for each profile k is characterized by a mean vector μ_k and covariance matrix Σ_k.
When the population consists of K different distributions, all assumed to be multivariate normal, and

you have multiple continuous indicator variables, then the multivariate representation of equation (2.1) is given in definition 2.2.1 in equation (2.2).

Definition 2.2.1. (Pastor et al., 2007 and Morgan et al., 2016) Given that

- $\theta_k = (\mu_k, \Sigma_k)$ is the vector containing the mean (μ_k) and the covariance matrix (Σ_k) for each profile k;
- ξ is the vector containing all the parameters in $\theta_1, \dots, \theta_K$;
- π_k is the probability of belonging to latent profile k, it can also be seen as the weight of the mixture model. The weights are non-negative and must satisfy: $\sum_{k=1}^{K} \pi_k = 1$;
- K represents the total number of underlying profiles;
- $\psi = (\pi_1, \dots, \pi_K, \xi^T)^T$ is the vector containing all the unknown parameters in this model;
- $\mathbf{y}_i$ represents the observed variables of person i;
- f_k is the normal density function associated to profile k,

the LPA model equation is represented by the following equation:

$$f(\mathbf{y}_i|\psi) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\theta_k), \qquad i = 1, \dots, N \tag{2.2}$$

This model shows that the multivariate distribution of multiple cluster indicators, contained in vector $\mathbf{y}_i$ for person i, given the model parameters ψ is a weighted mixture of K separate multivariate distributions.

2.3. Model estimation

As mentioned in section 2.2 LPA provides estimates for the unknown parameters ψ of the LPA model equation: the weights π_k and the means μ_k and covariances Σ_k of each latent profile k. With these estimates, the latent profiles are formed. To find these estimates, we will use the maximum likelihood. Therefore the log-likelihood derivatives need to be computed.

Proposition 2.3.1. Suppose there are N observations $\mathbf{y}_j$. Then the log-likelihood for ψ is given by:

$$\log L(\psi) = \sum_{j=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_j|\theta_k) \right\} \tag{2.3}$$

Proof. Using McLachlan and Peel, 2000 and the LPA model equation (2.2) the log-likelihood for ψ is calculated as follows:

$$\begin{aligned} \log L(\psi) &= \log \prod_{j=1}^{N} f(\mathbf{y}_j|\psi) \\ &= \sum_{j=1}^{N} \log f(\mathbf{y}_j|\psi) \\ &= \sum_{j=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_j|\theta_k) \right\} \end{aligned}$$

□

Since there is a summation inside the logarithm, it will be complex to calculate the derivative of this expression and then solving for the parameters. We can use an iterative method to achieve this purpose. It is called the Expectation-Maximization (EM) algorithm.

2.3.1. Expectation-Maximization algorithm

General EM-algorithm

The Expectation-Maximization algorithm is an approach to the iterative computation of maximum likelihood estimates. The algorithm is used when the target data is only partially observed.

Definition 2.3.1. (Bijma et al., 2017 and Bishop, 2006)
Given the set of all observed data X, the set of all latent variables Z, the set of all model parameters θ and a joint distribution $p(X, Z|\theta)$, the general steps for the EM-algorithm are the following:

1. Initialise $\tilde{\theta}_0$.
2. **E-step:** Given $\tilde{\theta}_i$, determine the function
$$Q(\theta, \tilde{\theta}_i) = \mathbb{E}_{\tilde{\theta}_i}[\log p(X, Z|\theta)] \tag{2.4}$$
3. **M-step:** Define $\tilde{\theta}_{i+1}$ as the point where this function takes on its maximum:
$$\tilde{\theta}_{i+1} := \max_{\theta} Q(\theta, \tilde{\theta}_i) \tag{2.5}$$

The E- and M-steps are alternated repeatedly until the likelihood values $p_{\tilde{\theta}_i}(X)$ converges to the maximum of the likelihood. Then $\tilde{\theta}_i$ will converge to the maximum likelihood estimator.

EM-algorithm for LPA model equation

Based on McLachlan and Peel, 2000, Bishop, 2006 and Murphy, 2012 we will apply the EM-algorithm to the LPA model equation to estimate the model parameters π_k, μ_k and Σ_k.

In this calculation some component-label vectors, $\mathbf{z}_1, \dots, \mathbf{z}_n$, are used to indicate if an observed variable belongs to a certain profile. We take $\mathbf{z}$ as a latent variable where $\mathbf{z}_j$ is a k-dimensional vector with

$$z_{jk} = \begin{cases} 1, & \text{if } \mathbf{y}_j \text{ belongs to the } k\text{th profile} \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

Proposition 2.3.2. Let $\gamma(z_{jk})$ be the probability that a data point $\mathbf{y}_j$ belongs to profile k. The estimates for the parameters π_k, μ_k and Σ_k found by the EM-algorithm are:

$$\hat{\pi}_k = \frac{\sum_{j=1}^N \gamma(z_{jk})}{N} \tag{2.7}$$
$$\hat{\mu}_k = \frac{\sum_{j=1}^N \gamma(z_{jk})\mathbf{y}_j}{\sum_{j=1}^N \gamma(z_{jk})} \tag{2.8}$$
$$\hat{\Sigma}_k = \frac{\sum_{j=1}^N \gamma(z_{jk})(\mathbf{y}_j - \mu_k)(\mathbf{y}_j - \mu_k)^T}{\sum_{j=1}^N \gamma(z_{jk})} \tag{2.9}$$

Proof. We will perform the steps of the EM-algorithm stated in definition 2.3.1. In our approach the parameters for the model are $\theta = \{\pi, \mu, \Sigma\}$.
Within the formulation of the mixture problem in the EM-framework, the observed-data matrix consisting of the vectors
$$\mathbf{Y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_N^T)^T$$
is viewed as being incomplete, and the associated component-label vectors, $\mathbf{z}_1, \dots, \mathbf{z}_N$, are not available. Using the component-label vectors defined in (2.6) we obtain the complete data-vector

$$\mathbf{Y}_c = (\mathbf{y}^T, \mathbf{z}^T)^T,$$

where
$$\mathbf{Z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_N^T)^T$$

So now we can try to maximize the likelihood for the complete data set $\{\mathbf{Y}, \mathbf{Z}\}$. The likelihood function is given by

$$p(\mathbf{Y}, \mathbf{Z}|\theta) = \prod_{j=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{jk}} (f_k(\mathbf{y}_j|\theta_k))^{z_{jk}} \tag{2.10}$$

Hence the log-likelihood function is given by

$$\log p(\mathbf{Y}, \mathbf{Z}|\theta) = \sum_{j=1}^{N} \sum_{k=1}^{K} z_{jk} \left[\log \pi_k + \log f_k(\mathbf{y}_j|\theta_k)\right] \tag{2.11}$$

The first step of the EM-algorithm is to initialise the values for the parameters π, μ and Σ. Let θ^* be the value specified initially for θ.
In the E-step we have to evaluate

$$\begin{aligned} Q(\theta^*, \theta) &= \mathbb{E}\left[\log p(\mathbf{Y}, \mathbf{Z}|\theta^*)\right] \\ &= \sum_{Z} p(\mathbf{Z}|\mathbf{Y}, \theta) \log p(\mathbf{Y}, \mathbf{Z}|\theta^*) \end{aligned} \tag{2.12}$$

To evaluate this function we need to know what the probability is that a data point $\mathbf{y}_i$ belongs to profile k. Using Bayes' rule we get the following:

$$\begin{aligned} p(\mathbf{Z}|\mathbf{Y}, \theta) = p(z_{jk} = 1|\mathbf{y}_j) &= \frac{p(\mathbf{y}_j|z_{jk} = 1)p(z_{jk} = 1)}{\sum_{a=1}^{K} p(\mathbf{y}_j|z_{ja} = 1)p(z_{ja} = 1)} \\ &= \frac{\pi_k f_k(\mathbf{y}_j)}{\sum_{a=1}^{K} \pi_a f_a(\mathbf{y}_j)} \\ &= \gamma(z_{jk}) \end{aligned} \tag{2.13}$$

So combining equations (2.11), (2.12) and (2.13), and using the fact that the latent variable z will only be 1 once everytime the summation is evaluated (see (2.6)), yields

$$\begin{aligned} Q(\theta^*, \theta) &= \sum_{Z} \gamma(z_{jk}) \log p(Y, Z|\theta^*) \\ &= \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma(z_{jk}) \left[\log \pi_k + \log f_k(\mathbf{y}_j|\theta_k)\right] \end{aligned} \tag{2.14}$$

In the M-step Q is being optimized with respect to π_k, μ_k and Σ_k. Q also needs to take the restriction into account that all π values should sum up to one. So we want to find the maximum of Q subjected to the equality constraint $\sum_{k=1}^{K} \pi_k = 1$. To achieve this, we will use a Lagrange multiplier λ. Hence:

$$Q(\theta^*, \theta) = \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma(z_{jk}) \left[\log \pi_k + \log f_k(\mathbf{y}_j|\theta_k)\right] - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{2.15}$$

Maximizing this equation gives us the parameter estimates.
Estimation of π_k.
First we maximize (2.15) with respect to π_k.

$$\frac{\partial Q(\theta^*, \theta)}{\partial \pi_k} = \sum_{j=1}^{N} \frac{\gamma(z_{jk})}{\pi_k} - \lambda \tag{2.16}$$

Setting this equal to zero gives:

$$\sum_{j=1}^{N} \frac{\gamma(z_{jk})}{\pi_k} - \lambda = 0$$

$$\Leftrightarrow \sum_{j=1}^{N} \gamma(z_{jk}) = \pi_k \lambda \tag{2.17}$$

$$\Leftrightarrow \sum_{k=1}^{K} \sum_{j=1}^{N} \gamma(z_{jk}) = \sum_{k=1}^{K} \pi_k \lambda \tag{2.18}$$

We know that the weights π_k must sum up to one. Since $\gamma(z_{jk})$ is a probability distribution, these probabilities over k must also sum up to one. Then from (2.18) it follows that $\lambda = N$. Using this result and (2.17) we get the following estimation for π_k:

$$\hat{\pi}_k = \frac{\sum_{j=1}^{N} \gamma(z_{jk})}{N} \tag{2.19}$$

Estimation of μ_k
To find the estimate of μ_k, we have to maximize $Q(\theta^*, \theta)$ in (2.14) with respect to μ_k. To do this maximization we use the assumption that the profiles are normally distributed with f_k as the profile-specific normal density function with profile-specific mean vector μ_k and covariance matrix Σ_k.
From this assumption and (2.14) it follows that:

$$\begin{aligned} Q(\theta^*, \theta) &= \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma(z_{jk}) \left[\log \pi_k + \log f_k(\mathbf{y}_j | \theta_k)\right] \\ &= \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma(z_{jk}) \left[\log \pi_k + \log \left[\frac{\exp(-\frac{1}{2}(\mathbf{y}_j - \mu_k)^T \Sigma_k^{-1} (\mathbf{y}_j - \mu_k))}{\sqrt{(2\pi)^N |\Sigma_k|}}\right]\right] \\ &= \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma(z_{jk}) \left[\log \pi_k - \frac{1}{2}\left[\log(|\Sigma_k|) + (\mathbf{y}_j - \mu_k)^T \Sigma_k^{-1} (\mathbf{y}_j - \mu_k) + N \log(2\pi)\right]\right] \end{aligned} \tag{2.20}$$

We want to calculate the derivative of (2.20) with respect to μ_k. Hereby we will use the fact of taking vector derivatives ($\mathbf{x}$ is a vector, $\mathbf{B}$ is a constant matrix):

$$\frac{d}{d\mathbf{x}} \mathbf{x}^T \mathbf{B} \mathbf{x} = 2\mathbf{B}\mathbf{x}.$$

This calculation of the derivative results in:

$$\frac{\partial Q(\theta^*, \theta)}{\partial \mu_k} = \Sigma_k^{-1} \sum_{j=1}^{N} \gamma(z_{jk})(\mathbf{y}_j - \mu_k) \tag{2.21}$$

Setting this equal to zero gives the following estimate for μ_k:

$$\Sigma_k^{-1} \sum_{j=1}^{N} \gamma(z_{jk})(\mathbf{y}_j - \mu_k) = 0$$

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \gamma(z_{jk}) \mathbf{y}_j}{\sum_{j=1}^{N} \gamma(z_{jk})} \tag{2.22}$$

Estimation of Σ_k To find the estimate of Σ_k, we have to calculate the derivative of (2.20) with respect to Σ_k. We can make this calculation easier, by using the trace-trick which reorders the scalar inner product $\mathbf{x}^T\mathbf{A}\mathbf{x}$ using the trace function (Murphy, 2012):

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = Tr(\mathbf{x}^T\mathbf{A}\mathbf{x}) = Tr(\mathbf{x}\mathbf{x}^T\mathbf{A}) = Tr(\mathbf{A}\mathbf{x}\mathbf{x}^T)$$

Let $\Sigma_k^{-1} = \Lambda$. We rewrite $Q(\theta^*, \theta)$ of (2.20) using this trace-trick:

$$\begin{aligned} Q(\theta^*, \theta) = & \sum_{k=1}^{K}\sum_{j=1}^{N}\left(\gamma(z_{jk})\log \pi_k - \frac{N}{2}\gamma(z_{jk})\log(2\pi)\right) - \frac{1}{2}\sum_{j=1}^{N}\gamma(z_{jk})\log|\Lambda| \\ & - \frac{1}{2}Tr\left\{\left(\sum_{j=1}^{N}\gamma(z_{jk})(\mathbf{y}_j - \mu_k)(\mathbf{y}_j - \mu_k)^T\right)\Lambda\right\} \end{aligned} \tag{2.23}$$

We take a derivative of this expression with respect to Λ. Hereby we use the fact that taking a derivative of the trace of a matrix goes as follows, with $\mathbf{A}$ and $\mathbf{B}$ matrices (Murphy, 2012):

$$\frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{B}\mathbf{A}) = \mathbf{B}^T.$$

Taking the derivative of (2.23) and setting it equal to zero yields

$$\frac{\partial Q(\theta^*, \theta)}{\partial \Lambda} = -\frac{1}{2}\left(\sum_{j=1}^{N}\gamma(z_{jk})\right)\Lambda^{-T} - \sum_{j=1}^{N}\gamma(z_{jk})(\mathbf{y}_j - \mu_k)(\mathbf{y}_j - \mu_k)^T = 0 \tag{2.24}$$

$$\Lambda^{-T} = \Lambda^{-1} = \hat{\Sigma}_k = \frac{\sum_{j=1}^{N}\gamma(z_{jk})(\mathbf{y}_j - \mu_k)(\mathbf{y}_j - \mu_k)^T}{\sum_{j=1}^{N}\gamma(z_{jk})} \tag{2.25}$$

□

2.3.2. Creating the latent profiles

After each profile distribution is characterized by the estimates from the EM-algorithm, the persons have to be classified. Each individual is assigned to the profile with the highest probability of membership. So in order to classify a given person, these probabilities of membership have to be calculated.
From the proof of proposition 2.3.2 (equation (2.13)) the probability of membership is derived.

Corollary 2.3.1. The probability of membership of individual i to profile k is

$$p_{ik} = \mathbb{P}(\text{individual } i \in \text{group } k|\mathbf{y}_j; \hat{\psi}) = \frac{\hat{\pi}_k f_k(\mathbf{y}_i|\hat{\theta}_k)}{\sum_{a=1}^{K}\hat{\pi}_a f_a(\mathbf{y}_i|\hat{\theta}_a)} \tag{2.26}$$

With this the latent profiles are formed and the persons are classified. To select the best types of profiles, you can use some test statistics. This is explained in section 2.5.

2.4. Model specification

There are several specifications in terms of to what extent each indicator varies and how the profile indicators relate to one another. The model can thus be specified in terms of whether and how the variances and covariances are estimated, and so there are different parametrizations of the covariance matrix Σ_k. In all the models the means are freely estimated in the different profiles.

Following Pastor et al., 2007 and Johnson, 2021 we are going to consider four different types of models. Here we use that there are r profile indicators.

1. **Equal variances and covariances fixed to 0**
   In this model the variances are equal across profiles. But the variances are allowed to differ across indicators within a profile, this is indicated by the different subscripts for the variances in the covariance matrix (2.27). In this matrix the covariances are fixed to 0, so this means that the indicators are uncorrelated to one another both within and across the clusters. In other words, in this specification the amount of variation around the mean for a specific variable is the same in each profile.

$$\Sigma_k = \begin{bmatrix} \sigma_1^2 & & & \\ 0 & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \sigma_r^2 \end{bmatrix} \tag{2.27}$$

2. **Varying variances and covariances fixed to 0**
   In this model the variances are allowed to differ both within and across profiles, this is indicated by the additional subscript k in the covariance matrix (2.28).
   If we look at this model in terms of variations, then we could say that in this specification the amount of variation of a profile indicator can be different in each profile.

$$\Sigma_k = \begin{bmatrix} \sigma_{1k}^2 & & & \\ 0 & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \sigma_{rk}^2 \end{bmatrix} \tag{2.28}$$

3. **Equal variances and equal covariances**
   In this model both the variances and covariances are constrained to be equal across profiles. The covariances are allowed to be freely estimated within a profile.

$$\Sigma_k = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1} & \sigma_{r2} & \dots & \sigma_r^2 \end{bmatrix} \tag{2.29}$$

4. **Varying variances and varying covariances**
   In this model both the variances and covariances are allowed to vary both within and across profiles.

$$\Sigma_k = \begin{bmatrix} \sigma_{1k}^2 & & & \\ \sigma_{21k} & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1k} & \sigma_{r2} & \dots & \sigma_{rk}^2 \end{bmatrix} \tag{2.30}$$

The first and second models are much alike. In the first model the variances are equal and in the second the variances are varying. If we compare these two models then we can conclude that the simpler model (so equal variances and covariances fixed to 0) is less computationally intensive, because there are less parameters that need to be calculated, since in the first model only one overall parameter for the variances in a profile indicator have to be calculated and in the second model one parameter for each profile has to be calculated. But on the other hand, the first model is less realistic, since in this model the variances are equal in every profile.

In the first two models the profile indicators are not allowed to relate to another. In the third and fourth model the variables can relate to another, due to the unfixed covariances. The third model is the simplest of the two because there the covariances are equal across the profiles, so this means that the covariance between two profile indicators must be the same in every profile. In the fourth model this doesn't have to be the case, because here the covariance can be freely estimated within each profile, so this means that when for example two profile indicators are statistically significant in one profile, this does not have to be the case in another profile.

2.5. Evaluating model fit

We want to find out the optimal number of profiles in the data. This typically involves estimating models with incremental numbers of latent profiles. Then using statistical indicators the best fitted model can be found. Multiple statistical indicators have to be used to find the optimal number of latent profiles, because every statistical fit indicator uses different values to indicate the best fit and also not every fit index indicates the same model to have the best fit.
There are a lot of different statistical indicators available. We will discuss the most common statistical indicators used in latent profile analysis; AIC, BIC, entropy, BLRT and ICL (Tein et al., 2013).

Information-based criteria (IC) indices are based on the log likelihood function of a fitted model, where the ICs impose a penalty on the number of parameters and/or sample size (Nylund et al., 2007). We will describe two IC indices, namely the AIC and BIC.

Definition 2.5.1. (Fox, 2016)
The Akaike information criterion (AIC) is deined as

$$\text{AIC} = -2\log_e L(\hat{\theta}) + 2s$$

where $\log_e L(\hat{\theta})$ is the maximized log-likelihood under the model, θ is the parameter vector for the model and s is the number of estimated parameters.

Definition 2.5.2. (Fox, 2016)
The Bayesian information criterion (BIC) is defined as

$$\text{BIC} = -2\log_e L(\hat{\theta}) + s\log_e n$$

where $\log_e L(\hat{\theta})$ is the maximized log-likelihood under the model, θ is the parameter vector for the model, s is the number of estimated parameters and n is the number of observations.

The AIC and BIC will take each of the models with different numbers of latent profiles and rank them from best to worst. For both the AIC and BIC it holds that the model with the lowest values is considered as the best fitting model.

Definition 2.5.3. (Tein et al., 2013 and Pastor et al., 2007)
The entropy statistic E is

$$E = 1 - \frac{\sum_{j=1}^{N}\sum_{k=1}^{K}(-p_{jk}\log(p_{jk}))}{N\log(K)} \tag{2.31}$$

where p_{jk} is the probability of belonging to a cluster:

$$p_{jk} = \frac{\pi_k f_k(\mathbf{y}_j|\mu_k, \Sigma_k)}{\sum_{a=1}^{K}\pi_a f_a(\mathbf{y}_j|\mu_a, \Sigma_a)} \tag{2.32}$$

Higher entropy values represents a better fit, it indicates more precision in group membership classification. Values larger than 0.80 indicate that the latent profiles are very distinguishable from each other.

Definition 2.5.4. (Nylund et al., 2007 and Tein et al., 2013)
The bootstrap likelihood ratio test (BLRT) compares a $k-1$-profile model with a k-profile model. The BLRT uses a bootstrap resampling method to approximate the p-value, which indicates if the null hypothesis that the $(k-1)$-profile provides a significantly better fit than the k-profile model should be rejected in favor of the alternative hypothesis that the k-profile model provides a better fit than the $k-1$-profile model. A statistically significant p-value $(p < 0.05)$ indicates that the null hypothesis should be rejected, so then it will indicate a significant improvement in the model fit. This test can only be used for models that use the same parametrization.

Definition 2.5.5. (Scrucca et al., 2016)
The integrated complete-data likelihood (ICL) is defined as

$$\text{ICL} = \text{BIC} + 2\sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik}\log(\gamma_{ik})$$

where γ_{ik} is the conditional probability that $\mathbf{y}_i$ belongs to the kth mixture component, and $z_{ik} = 1$ if the ith datapoint is assigned to cluster k and 0 otherwise (like is stated in section 2.3.1). Models with the lowest ICL gives the best fitted model.

3 PRIME research

In this chapter we will look into an application in PRIME. We will identify different profiles of motivation in a student population at Delft University of Technology. For this we will use LPA. In LPA the profiles are identified based on values of continuous observed variables identified as profile indicators. So before we can really apply LPA on the dataset, we have to prepare the dataset by making the observed variables continuous. This is done by factor analysis. Then we will use R to apply LPA to the prepared data. When the profiles are created we will ensure that the profiles are clearly differentiated using a statistical method.

3.1. PRIME

The PRogramme of Innovation in Mathematics Education (PRIME) (“PRIME”, n.d.) redesigns mathematics courses for the engineering programmes of the TU Delft. Using a blended learning cycle design PRIME aims to:

- enhance teaching and learning,
- improve connection between mathematics and engineering,
- increase student active participation and motivation.

The third aim is most relevant for our current research where we examine the motivational profiles of students. To increase student’s motivation it is helpful to identify clusters of the students based on their motivation. In this thesis we will make this overview by dividing the students into different profiles based on their motivations. The profiles can’t be measured directly, so they are latent variables. Therefore we will use latent profile analysis to identify those latent profiles. To do this research the following research question is formulated:
Main research question:
What are the student profiles that can be distinguished based on students’ motivation and are these profiles significantly different?

3.1.1. Motivations and Motivational Profiles

In this research, motivation is considered form the perspective of self-determination theory (Deci and Ryan, 2000). According to the self-determination theory, students’ motivation can be differentiated by the autonomy that is experienced. Therefore, an important distinction is made between autonomous and controlled motivation.

Controlled motivation is about acting with a sense of pressure. Controlled motivation consist of two subtypes: *introjected* and *external motivation*. Introjected motivation is experienced when feelings of pressure come from yourself, so for example shame. External motivation is experienced when feelings of pressure come from an external source such as demands from an authority figure.

Autonomous motivation concerns the sense of volition and the psychological freedom. There are also two types of autonomous motivation: *identified* and *intrinsic* motivation. With identified motivation, the activity is personally important or valuable. With intrinsic motivation, students study out of individual interest or the satisfaction the task or activity brings them.

Previous research shows that autonomous and controlled motivation can be experienced by the same student (Wijnia and Baars, 2021). Therefore, it is important to examine whether both types of motivations benefit learning or whether one type of motivation is more beneficial for learning than the other. Studies that used a person-centred approach have identified two to six motivational profiles. Among the different motivational profiles identified across studies, four types of motivational profiles were most common. The labels attached to the four motivational profiles were:

1. Good-quality: high levels of autonomous motivation and low levels of controlled motivation
2. Poor-quality: low levels of autonomous motivation and high levels of controlled motivation
3. High-quantity: high levels of both autonomous and controlled motivations
4. Low-quantity: low levels of both autonomous and controlled motivation.

Besides the four most common motivational profiles, studies also identified other motivational profiles, such as moderately autonomous motivation and moderately unmotivated where the differences between autonomous and controlled motivation were less extreme. Differences in the number and types of motivational profiles in the studies could be due to the differences in the studies in which the motivational profiles were investigated. One of the differences is the educational context and learning environment. For example, students may experience higher levels of autonomy in a blended learning environment than in a traditional on-campus learning environment. Another difference is the level of specificity in which motivation is being measured. Moderate profiles were more commonly identified in studies that used a finer-grained representation of motivation (i.e., intrinsic, identified, introjected, and extrinsic) compared to a higher order dimension (i.e., autonomous and controlled).
Therefore, it is of interest to examine how the number and types of motivational profiles identified in our study will compare to previous studies. The types of motivational profiles are for example, students studying math in a blended learning environment and motivation measured at a finer-grain. Hence, a small research question of my research is:
Are the identified profiles comparable to those found in previous studies?

3.1.2. Prior research and hypothesis

There has been prior research on motivational profiles in education using latent profile analysis. Wijnia and Baars, 2021 made a review of all prior research that has been done. This investigation of prior research identified between *two to six motivational profiles*. We will use this observation in our research. This investigation also concluded that among the different motivational profiles identified across studies, the mode of number of profiles is four. This leads to the following hypothesis for our research:
The number of student profiles that can be distinguished based on students' motivation of the PRIME dataset is four.

3.1.3. Data

PRIME has collected the answers on a survey from students of an engineering programme at the TU Delft. This survey was about the motivation to study for a mathematics course. It is a validated survey adapted from the Journal of Educational Psychology (Vansteenkiste, 2009). It consisted of 16 statements, which where of answers to the question 'Why are you studying in general? I am studying...'. Each statement belongs to one of the four types of motivations: external, introjected, identified and intrinsic. In section 3.1.1 these types of motivation are explained. There are 4 statements per type of motivation. Here are examples of one statement per type of motivation:

- **External motivation:**
  Because that's what others (parents, friends, etc.) force me to do.
- **Introjected motivation:**
  Because I want others to think I'm a good student.

- **Identified motivation:**
  Because I want to learn new things.
- **Intrinsic motivation:**
  Because I enjoy doing it.

All the 16 statements of the survey can be found in Appendix A.

Students indicated how important each of the listed motives is for them to study using a 5-point Likert scale, where 1 indicates completely not important, and 5 indicates very important. Data of 187 students can be used.

3.2. Preparing survey data for analysis

The data consists of ordinal data, namely the 16 survey items per person. There are 4 survey items per latent variable (type of motivation). To use this data with the fact that four survey items relate to a specific type of motivation we have to convert the 16 survey items into 4 composite values. This can be done in several ways. We have investigated three different ways and checked which one will work in our research. These different ways are using the total score for each subscale, using the median and using weighted factor loadings. We will eventually use the weighted factor loadings in our research.

3.2.1. Using total score

An approach to convert the ordinal data into some composite value is to use the total score. In this approach the four items associated with each factor are used to form a sum subscale total, in order to represent each of the four types of motivations. This approach has been used in multiple latent profile analysis researches (Pastor et al., 2007). Hence, the subscale scores are on a scale of 4-20.
With this approach there can't be any weight assigned to a factor, hence this approach treats every answer of a statement as contributing to the composite score equally. But this is not the case, since the order in the numbers have some important meaning. Therefore this approach will not be as good as the other one and will not be used in this research.

3.2.2. Using median

Another approach is to convert the ordinal data into a composite value using the median. With this approach attention is payed to the fact that every answer of the statements does not have to be treated equally, since here the median is taken from each of the four items associated with each factor. Since in this approach only the median is used, it is quite a simple way to handle this. It doesn't really give a fine understanding of the problem. We have performed the latent profile analysis using this approach, but this approach will not be used in the main analysis of this research. If you are interested in the latent profile analysis when the ordinal data is converted into a composite scores using the median, we refer you to appendix B.

3.2.3. Using weighted factor loadings

When converting ordinal data into composite value it is important to note that in the points from the 5-point scale the order matters (e.g., 1 = 'completely not important' and 5 = 'very important'). An approach which we can use to achieve this is by using factor analysis to generate the composite scores. This approach is used in our research and explained in the remainder of this section.

From Yong and Pearce, 2013 the following is known about factor analysis.
Factor analysis is a statistical approach that uses the concept that measurable and observable variables can be reduced to fewer latent variables that are unobservable and share a common variance. The goal of factor analysis is to summarize data so that relationships and patterns can be easily interpreted and understood. Factor analysis is commonly used when you want to discover the number of factors that have an effect on variables and to analyze which variables 'belong' to another. Factor analysis is useful for studies that involve items from questionnaires.
In factor analysis the factor loading for a variable is a measure of how much the variable contributes to the factor. How these factor loadings are estimated is out of scope of this thesis, because we will use

the R-function `factanal()` which is available in R-Studio to perform factor analysis. For the mathematical model of factor analysis and more information about this analysis, see Yong and Pearce, 2013 and Lawley and Maxwell, 1962.

From Starkweather, 2012 the general steps for generating composite scores using weighted factor loadings are the following:

1. Recode ordinal responses to numeric responses
2. Apply a factor analysis model which shows the calculated correlation structure of the variables
3. Save the factor scores and factor loadings
4. Rescale the factor scores using the factor loadings, the weighted mean and the weighted standard deviation of the original data. In this way the composite scores refer to the same labels (e.g. 'completely not important') as the original scores. Here, the factor loadings are the weights for the weighted mean and weighted standard deviation calculation.

   Definition 3.2.1. The weighted mean of data $x_1, x_2, \ldots, x_n$ using the set of weights $w_1, w_2, \ldots, w_n$ is defined as,

   $$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

   Definition 3.2.2. The weighted standard deviation of data $x_1, x_2, \ldots, x_n$ using the set of weights $w_1, w_2, \ldots, w_n$ is defined as,

   $$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} w_i}{(\sum_{i=1}^{n} w_i)^2 - \sum_{i=1}^{n} (w_i)^2} \sum_{i=1}^{n} w_i (x_i - \hat{\mu})^2}$$

After performing these steps we create a dataframe consisting of the rescaled factor scores as the composite scores for each section of the questionnaire. The R-code of preparing the data can be found in appendix C.

3.3. Formulating the model in R

3.3.1. TidyLPA

Now that we have transformed the data into ready to use data, we can formulate the model in R. There are several packages available which you can use to carry out latent profile analysis in R. The package that we are going to use is tidyLPA. Using tidyLPA gives you the functionality to specify different models that determine whether and how the parameters means, variances and covariances are estimated. TidyLPA is also able to specify and compare different solutions for the estimate of the number of profiles (Rosenberg et al., 2018).

In tidyLPA both the input and output are a dataframe. This can then easily be used to create plots. TidyLPA also uses the "pipe" operator, `%>%` to compose functions. By *piping* we pass the results of one function to the next. These functions that the results can be passed on to are stated below (Rosenberg, 2020).

- `select()`:
  Picks variables based on their names from the dataset.
- `scale()`:
  Scale the data.
- `estimate_profiles()`:
  Estimates latent profiles. In this command you can specify the number of profiles and the type of model that you want to use.

- `plot_profiles()`:
  Plot the latent variables with the variable means and variances. In this profile plot there is being payed attention to the visualization of classification uncertainty. This is done by showing:
  1. Bars which show a confidence interval for the class mean point
  2. Boxes which show the standard deviations within each profile
  3. Raw data, where the posterior class probability is used to weight the transparency of the data. In this way each datapoint is most clearly visible for the profile where it is most likely to belong to.
- `compare_solutions()`:
  Compares the fit of several estimated models. These models have varying numbers of profiles and model specifications, so this function helps select the optimal number of classes and model specification.
- `mutate()`:
  Adds new variables to a dataframe that are functions of existing variables.
- `get_data()`:
  Get data from objects generated by tidyLPA. This function returns the original data frame, with variables that are outcomes of the profiles included. So in this data frame among others the variables used to create the profiles and the profiles assignments are given. This function is often used when you want to use the estimated profiles in subsequent analyses.

In tidyLPA you first have to mention the name of the dataframe followed by the names of the variables used to create the profiles, using the `select()` function. Then you also specify the number of profiles and the type of model that you want to use, this is done by the `estimate_profiles` function. The four models explained in section 2.4 can be specified in tidyLPA.

3.3.2. Mclust

TidyLPA is a ”wrapper” to the mclust package. This means that tidyLPA uses mclust functions to carry out LPA, hence it provides ”wrappers” to these functions that make them easier to use (Rosenberg, 2021). From Scrucca et al., 2016 mclust is a popular R-package for Gaussian mixture modelling for model-based clustering, classification and density estimation. It provides functions to estimate the parameters by the EM-algorithm to form the latent profiles. There are also functions that combine model-based hierarchical clustering, EM for mixture estimation and fit indices for model selection. In mclust the default fit index for selecting a model is the BIC. Mclust gives an extensive strategy for clustering, density estimation and discriminant analysis. It is also possible to perform single E- and M-steps of the EM-algorithm. Some additional functions are also available to display and visualise fitted models along with clustering, classification, and density estimation results. There are also functions available where you can visualise the fit indices for several models with multiple number of profiles.

In section 2.3.1 the EM-algorithm was explained. There it is stated that you have to have to initialise values for the parameters with which you can start the EM-algorithm. These initial parameters can be collected by first performing another clustering method. By Scrucca et al., 2016, mclust uses the partitions obtained from model-based hierarchical agglomerative clustering (MBHAC) method to collect these initial parameters.

Definition 3.3.1. (Scrucca and Raftery, 2015)
Model-based hierarchical agglomerative clustering (MBHAC) is an approach where k clusters are obtained from a large number of smaller clusters by recursively merging the two clusters that have the smallest dissimilarity in a model-based sense, i.e. the dissimilarity used for agglomeration is derived from a probabilistic model. The dissimilarity based on a Gaussian mixture model is equal to the decrease in the classification likelihood resulting by merging of two clusters.

3.4. Results from tidyLPA

We will perform the latent profile analysis using the dataframe of the composite values that is created in section 3.2.3. The names of these composite scores that we will use in R are:

- External motivation: `ext.composite.score`
- Introjected motivation: `ijt.composite.score`
- Identified motivation: `idt.composite.score`
- Intrinsic motivation: `int.composite.score`

We want to estimate all the possible models, such that we can then use the fit indices to select the best model of all the models. We will estimate the models with the number of profiles between 2 to 6, because this is what has been investigated in prior research (see section 3.1.2) and also the four types of models from section 2.4, this is all specified using the `estimate_profiles` function of R. In the tidyLPA package the models from section 2.4 have different numbers:

- Model 1 in R: Equal variances and covariances fixed to 0.
- Model 2 in R: Varying variances and covariances fixed to 0.
- Model 3 in R: Equal variances and equal covariances.
- Model 6 in R: Varying variances and varying covariances.

The R-code of LPA using tidyLPA can be found in appendix C.

This estimation in R using tidyLPA gives the different types of models that are formed. These are every type of the 4 different models. And for each of these models the number of profiles from 2 to 6 are being estimated. For each type of model the fit indices are given:

```
    tidyLPA analysis using mclust:

Model Classes AIC     BIC     Entropy prob_min prob_max n_min n_max BLRT_p
1     2       1438.90 1480.48 0.83    0.85     0.98     0.17  0.83  0.01
1     3       1423.55 1481.13 0.71    0.79     0.92     0.14  0.59  0.01
1     4       1405.92 1479.49 0.67    0.75     0.87     0.12  0.33  0.01
1     5       1387.50 1477.06 0.76    0.79     0.95     0.02  0.38  0.01
1     6       1385.17 1490.72 0.77    0.79     0.91     0.04  0.40  0.17
2     2       1424.73 1479.10 0.75    0.92     0.93     0.33  0.67  0.01
2     3
2     4
2     5
2     6
3     2       1384.68 1445.45 0.69    0.90     0.92     0.45  0.55  0.01
3     3       1365.65 1442.42 0.71    0.77     0.92     0.20  0.52  0.01
3     4       1363.75 1456.50 0.69    0.68     0.93     0.09  0.46  0.18
3     5       1373.85 1482.60 0.57    0.00     0.93     0.00  0.43  0.99
3     6       1325.19 1449.93 0.87    0.85     0.98     0.01  0.48  0.01
6     2       1355.11 1447.87 0.81    0.88     0.96     0.22  0.78  0.01
6     3       1329.48 1470.22 0.72    0.84     0.93     0.20  0.45  0.01
6     4       1286.76 1475.47 0.92    0.95     1.00     0.10  0.64  0.01
6     5       1293.12 1529.81 0.88    0.86     0.99     0.10  0.38  0.44
6     6
```

From the results it is noticeable that model 2 can't be estimated for 3 to 6 profiles, and model 6 can't be estimated for 6 profiles. The reason these models can not be estimated is that the EM algorithm fails to converge. This happens due to singularity in the covariance matrix estimate (Scrucca et al., 2016). What this singularity exactly means and why it follows from this that some models can't converge is not investigated in this research, we focus on the models that are estimated. In the discussion (chapter 5) we will give a recommendation how to continue with this non-convergence models.
The other models do give some results. So we will continue the analysis with these models. To find the best model we are using the fit indices explained in section 2.5. The probabilities minimum and

maximum (prob_min and prob_max) are the minimum and maximum of the average latent class probabilities for most likely class membership, by assigned class. So since the individuals are assigned to the profiles they have the highest probability of belonging to, these prob_min and prob_max should be as high as possible, since this reflects greater classification certainty.

Comparing the fit indices with each other gives that for model 6 (varying variances and varying covariances) with four profiles has the best model fit given that the AIC (1286.76) is the lowest, the entropy (0.92) is the highest, the probabilities minimum (0.95) and maximum (1.00) are the highest and the BLRT p-value (0.01) is statistically significant.

In section 3.3.1 it is explained that using the function compare_solutions we can look for the optimal model. The results of this function are:

```
    Compare tidyLPA solutions:

 Model Classes BIC      Warnings
 1     2       1480.481
 1     3       1481.127
 1     4       1479.490
 1     5       1477.060
 1     6       1490.718
 2     2       1479.102
 2     3                Warning
 2     4                Warning
 2     5                Warning
 2     6                Warning
 3     2       1445.452
 3     3       1442.415
 3     4       1456.505
 3     5       1482.601 Warning
 3     6       1449.928 Warning
 6     2       1447.871
 6     3       1470.218
 6     4       1475.469
 6     5       1529.805
 6     6                Warning

Best model according to BIC is Model 3 with 3 classes.

An analytic hierarchy process, based on the fit indices AIC, AWE, BIC, CLC,
and KIC (Akogul & Erisoglu, 2017), suggests the best solution is Model 6
with 4 classes.
```

The output of this function gives the BIC values for each estimated model. It also gives some conclusion about the best fitted model based only on the BIC. This is model 3 with 3 profiles, since this model has the lowest BIC value. But the output of this function also gives that the best solution of the best fitted model is based on the multiple fit indices AIC, AWE, BIC, CLC and KIC (Akogul and Erisoglu, 2017). The function compare_solutions used all those fit indices to get the best fitted model, and the conclusion to which this function arrived is also model 6 with 4 profiles. So this again shows that multiple fit indices are necessary to obtain the best fitted model.

Using the plot_profiles function we can create the profile plot of the best fitted model (figure 3.1). In this profile plot the 4 different latent profiles are displayed using different colours and with the properties which are stated in the explanation of the plot_profiles function in section 3.3.1. This plot shows the means and variances of the four types of motivations for each profile. The plot is not really clear, since the values of the types of motivations do vary much. Therefore we also created a table with the means and standard deviations of the best fitted model, see table 3.1.

At first, the number of individuals per profile (n) is given in the table. From this we see that profile 1 contains the most individuals and profile 4 the least.

In this table also the means of the profiles and types of motivation can be compared. The standard deviation indicates if the values tend to be close to the mean. A high standard deviation indicates that the values are widely spread out and a low standard deviation indicates that the values are closely to the mean.

These standard deviations are illustrated in the profile plot (figure 3.1) as the boxes around the mean value. So when we look at the external motivation for example, then we would expect from table 3.1 that profile 1 should be widely spread (SD = 0.87) and profile 4 should tend to be close to the mean value (SD = 0.13). When we look at the profile plot (figure 3.1) these expectations indeed hold.

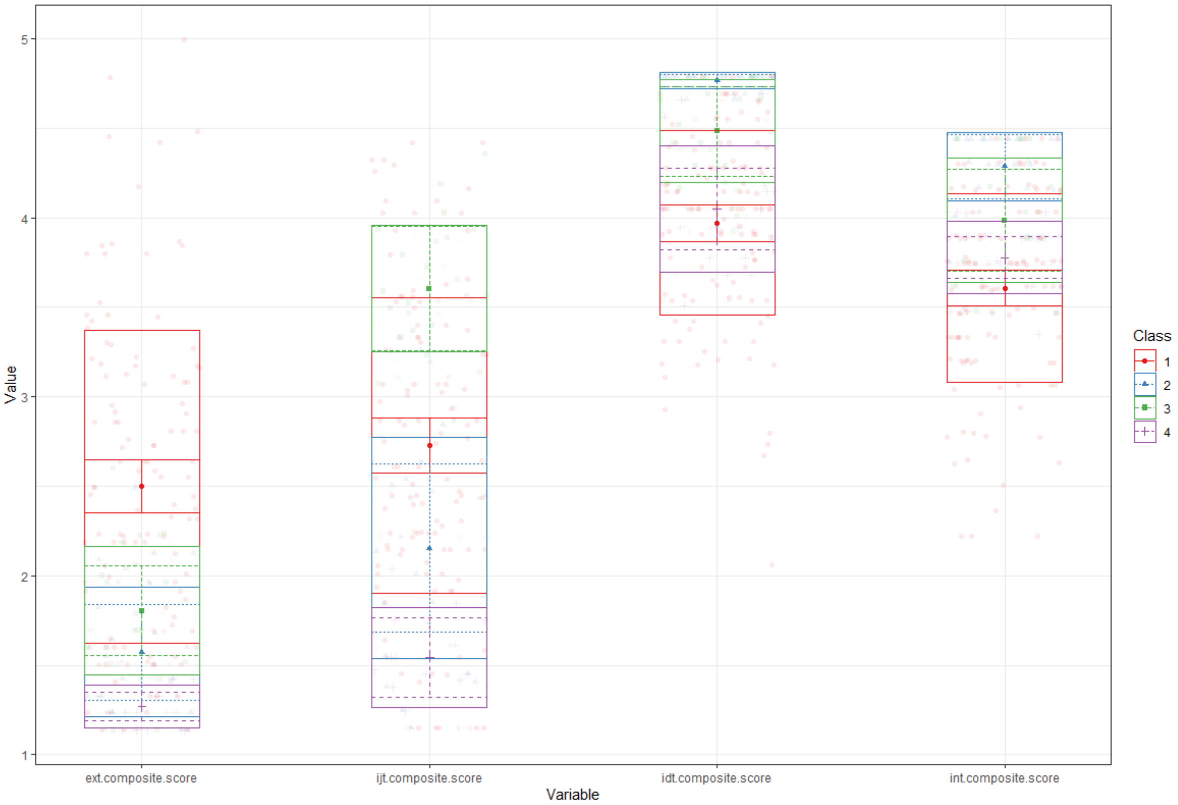


Figure 3.1: Profile plot of the best fitted model. The means per type of motivation per profile are indicated with the points. The boxes around the points indicate the variations of the motivation types per profile.

Means and standard deviations associated with the best fitted model

Profile	n	External motivation		Introjected motivation		Identified motivation		Intrinsic motivation	
		Means	SD	Means	SD	Means	SD	Means	SD
1	115	2.54	0.87	2.73	0.83	3.96	0.52	3.59	0.53
2	23	1.59	0.38	2.16	0.62	4.77	0.05	4.28	0.20
3	24	1.79	0.36	3.61	0.35	4.48	0.29	3.97	0.35
4	19	1.27	0.13	1.55	0.29	4.04	0.36	3.78	0.20

Table 3.1: Table for the means and standard deviations associated with the best fitted model.

To interpret the profile plot better, it is possible to scale the data before estimating the profiles. Using the `scale()` function in R the data is being scaled and hence standardized means are being created. In figure 3.2 the profile plot with the scaled data is shown. Scores below -1 indicate low scores and scores above 1 indicate high scores.

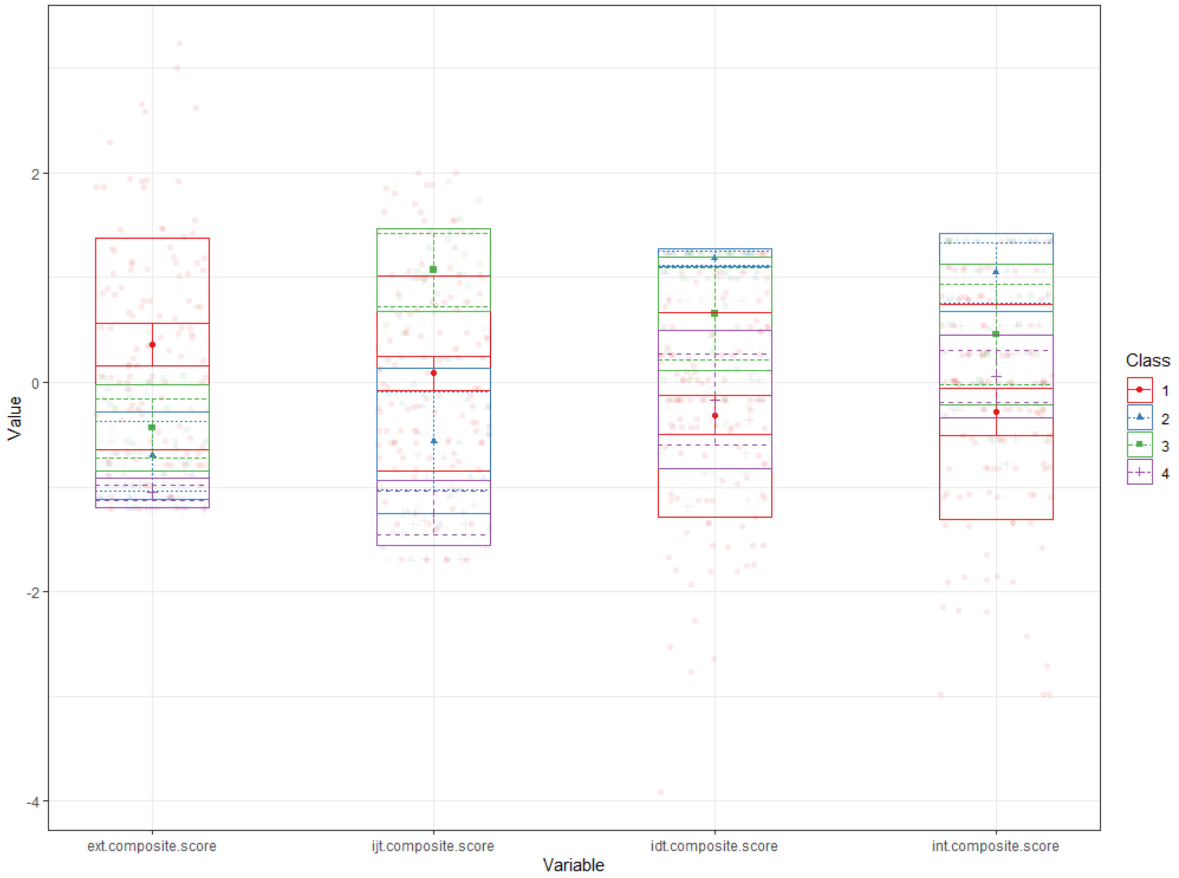


Figure 3.2: The profile plot with scaled data for the best fitted model. Here the data is scaled before estimating the profiles.

The profile plots in figures 3.1 and 3.2 and table 3.1 are used to name the created profiles. We will compare the mean scores on the four types of motivations between the profiles. Based on these comparisons we attach labels to the profiles. These labels are in line with the labels of previous research in section 3.1.1.

1. Profile 1 has the lowest scores on the identified and intrinsic motivations. And the highest score on external motivation. It scores moderately high on introjected motivation. Therefore the label attached to profile 1 is **poor quality**.

2. Profile 2 has the highest value on the identified and intrinsic motivations. It has moderate scores on external and introjected motivations. Therefore the label attached to profile 2 is **good quality**.

3. Profile 3 scores high values on the introjected, identified and intrinsic motivations. It has a moderate score on external motivation. Therefore the label attached to profile 3 is **high-quantity**.

4. Profile 4 scores low values on the external and introjected motivations. And has moderately low scores on the identified and intrinsic motivations. Therefore the label attached to profile 4 is **low-quantity**.

This means that our hypothesis in section 3.1.2 that the number of student profiles that can be distinguished based on students' motivation of the PRIME dataset is four holds.

3.4.1. Checking profiles

Now that the latent profiles are created we have to ensure that the profiles are clearly differentiated. So we have to test for differences between the four profiles. This is done by conducting a one-way ANOVA.

Definition 3.4.1. A one-way ANOVA is a statistical test that examines the relationship between a quantitative response variable and a factor. It tests the equality of two ore more population means by examining the variances in collected samples (Hesamian, 2016). The null hypothesis for the one-way ANOVA is that two or more means are equal. Hence, a significant result means that the two means are unequal.

From Hesamian, 2016 and Lee and Lee, 2018 the assumptions of ANOVA are:

1. Assumption of normality.
   The dependent variable is normally distributed in each group.
2. Assumption of homogeneity of variance.
   Each group has the same variance.
3. Assumption of independence.
   The groups should be independent. This means that the groups should be made up of different people.

To test for the difference between the four profiles we want to conduct four times one-way ANOVA using profile membership as the independent variable and the external, introjected, identified and intrinsic motivations as the dependent variables. But before we conduct the one-way ANOVA we have to check if the assumptions of the one-way ANOVA from definition 3.4.1 hold in our research:

1. Assumption of normality.
   From the first assumption of the Latent Profile Analysis in section 2.2 we know that the continuous indicators are normally distributed within each latent profile. These continuous indicators are the four types motivations (external, introjected, identified and intrinsic) and therefore this assumption holds.
2. Assumption of homogeneity of variance.
   The best fitted model from the LPA for which we want to perform the ANOVA is the model with varying variances and varying covariances. So this means that the variances are allowed to vary across the profiles. Therefore the variances in each group are not equal and this assumption does not hold.
3. Assumption of independence.
   The four profiles are independent from each other, because each profile consists of different people. Therefore this assumption holds also.

One-way ANOVA is performed only in cases where every assumption of definition 3.4.1 holds. Although, it is a robust statistic that can be used even when there is a deviation from the assumption of homogeneity of variance. When this is the case, the Games-Howell test can be applied (Lee and Lee, 2018). The Games-Howell test is applicable in cases where the homogeneity of variance assumption is violated.

In other researches the ANOVA was applied and in Wijnia and Baars, 2021 the Games-Howell test is also applied. These researches were researches from the social sciences, so here they didn't check whether the assumptions of the ANOVA were valid or whether the Games-Howell test could be applied. We will apply the Games-Howell test also, but we recommend that in future research this part could be more intensively investigated. Also about the mathematics behind the Games-Howell test could be more researched.

The Games-Howell test is performed in R. Here you have to use the ANOVA fits with the profile membership as the independent variable and the external, introjected, identified and intrinsic motivations as the dependent variables. To get these variables we use the `get_data()` function (see section 3.3.1). The R-code of the Games-Howell test can be found in appendix C. The results of the Games-Howell test for the external, introjected, identified and intrinsic motivation are stated in tables 3.2, 3.3, 3.4 and 3.5.

Results Games-Howell test for external motivation

	Mean	sd	n	Significant group
1	2.536	0.874	115	a
2	1.590	0.382	23	b
3	1.791	0.364	24	b
4	1.272	0.126	19	c

Table 3.2: Results of the Games-Howell test for external motivation. It shows the means and standard deviations per profile, the number of individuals per profile and the significance group indicators. Different letters indicate significantly different groups based on the external motivation.

Results Games-Howell test for introjected motivation

	Mean	sd	n	Significant group
1	2.734	0.825	115	a
2	2.158	0.622	23	b
3	3.607	0.349	24	c
4	1.550	0.292	19	d

Table 3.3: Results of the Games-Howell test for introjected motivation. It shows the means and standard deviations per profile, the number of individuals per profile and the significance group indicators. Different letters indicate significantly different groups based on the introjected motivation.

Results Games-Howell test for identified motivation

	Mean	sd	n	Significant group
1	3.955	0.516	115	a
2	4.765	0.047	23	b
3	4.481	0.289	24	c
4	4.043	0.362	19	a

Table 3.4: Results of the Games-Howell test for identified motivation. It shows the means and standard deviations per profile, the number of individuals per profile and the significance group indicators. Different letters indicate significantly different groups based on the identified motivation.

The values for the means, standard deviations and number of individuals n are equal to the values of this variables in table 3.1. The letters in the last column of tables 3.2, 3.3, 3.4 and 3.5 indicate significant differences between the profiles (p-value smaller than 0.05). When the letters are different it indicates that these profiles are significantly different from each other. From these results we conclude that the profiles 1-2, 1-3, 1-4, 2-4 and 3-4 are significantly different based on the external motivation (table 3.2), all profiles are significantly different based on the introjected motivation (table 3.3), the profiles 1-2, 1-3, 2-3, 2-4 and 3-4 are significantly different based on the identified motivation (table 3.4) and the profiles 1-2, 1-3, 1-4, 2-3 and 2-4 are signigicantly different based on the intrinsic motivation (table 3.5).
For the profiles where the letters are the same per type of motivation, we can not say anything about the significant difference between the profiles.

3.5. Results from mclust

In section 3.4 we saw that there were some models that can't be estimated. This is due to the non-convergence in the EM-algorithm. To check this we will perform the latent profile analysis using the mclust package, following the explanation of section 3.3.2 that the mclust package clearly uses the EM-algorithm and contains some additional functions for the EM-algorithm. We also want to make a clearer profile plot. For the R-code, see appendix D.

Mclust makes it possible to create plots of the fit indices, for example a BIC plot and ICL plot. The BIC plot for the models estimated is visible in figure 3.3.

Results Games-Howell test for intrinsic motivation

	Mean	sd	n	Significant group
1	3.593	0.533	115	a
2	4.283	0.200	23	b
3	3.971	0.354	24	c
4	3.781	0.202	19	c

Table 3.5: Results of the Games-Howell test for intrinsic motivation. It shows the means and standard deviations per profile, the number of individuals per profile and the significance group indicators. Different letters indicate significantly different groups based on the intrinsic motivation.

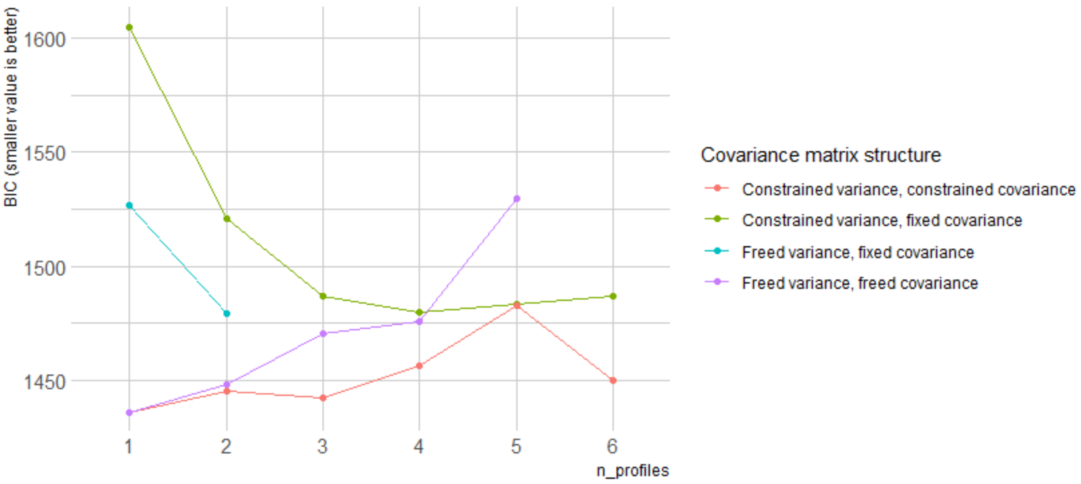


Figure 3.3: Plot for BIC model selection criteria for all the estimated models

In this BIC plot the different types of models can clearly be compared in terms of the BIC. From the figure it is clear that mclust also can't measure the models with varying variances and fixed covariances, similar to the results from tidyLPA, since the BIC plot does not give values for models 3,4,5 and 6 of the model with varying variances and fixed covariances. From figure 3.3 is is also clear that if we only look at the BIC fit index the models with equal variances and fixed covariances gives the worst fit (largest BIC value) for every amount of number of profiles except 5 profiles, then the varying variance and varying covariance gives the worst fit. The model with equal variances and equal covariances gives the best fitted models (smallest BIC value) across all the number of profiles.

When we look at the plot for ICL model selection criteria in figure 3.4 we see that the models with equal variances and fixed covariances still give the worst fitted models (largest ICL value) across all the number of profiles except for 5 number of profiles, then the model with equal variance and equal covariance is the worst. Now the best fitted model is not really clear across all the number of profiles. Hence multiple fit indices are needed to find the best fitted models.
Since the profile plot in figure 3.2 does not really visualise the latent profiles clearly, we want to make a clearer overview of the latent profiles that are formed in the best fitted model. Using mclust we obtain the clear plot in figure 3.5. At the end of section 3.4 we describe the latent profiles obtained. These profiles are now clearly visible in the figure.

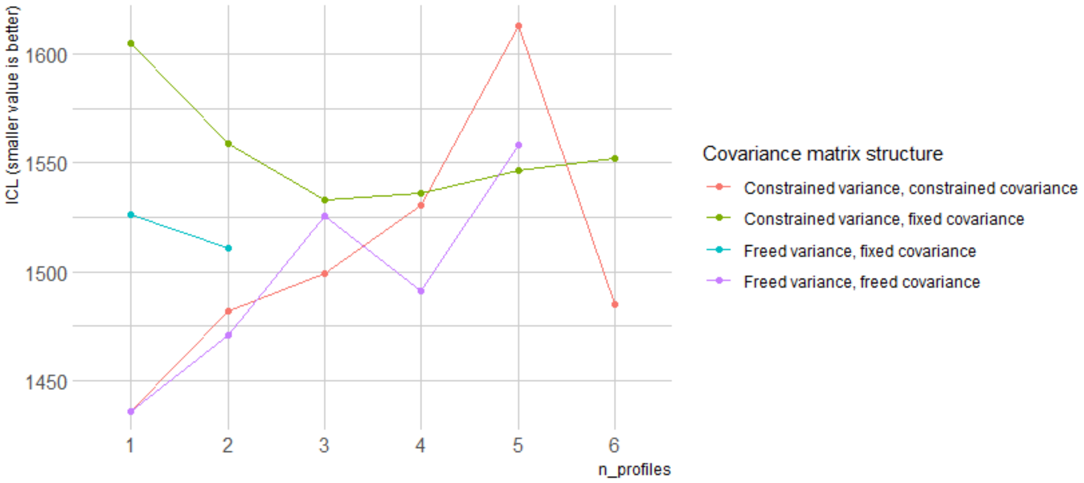


Figure 3.4: Plot for ICL model selection criteria for all the estimated models

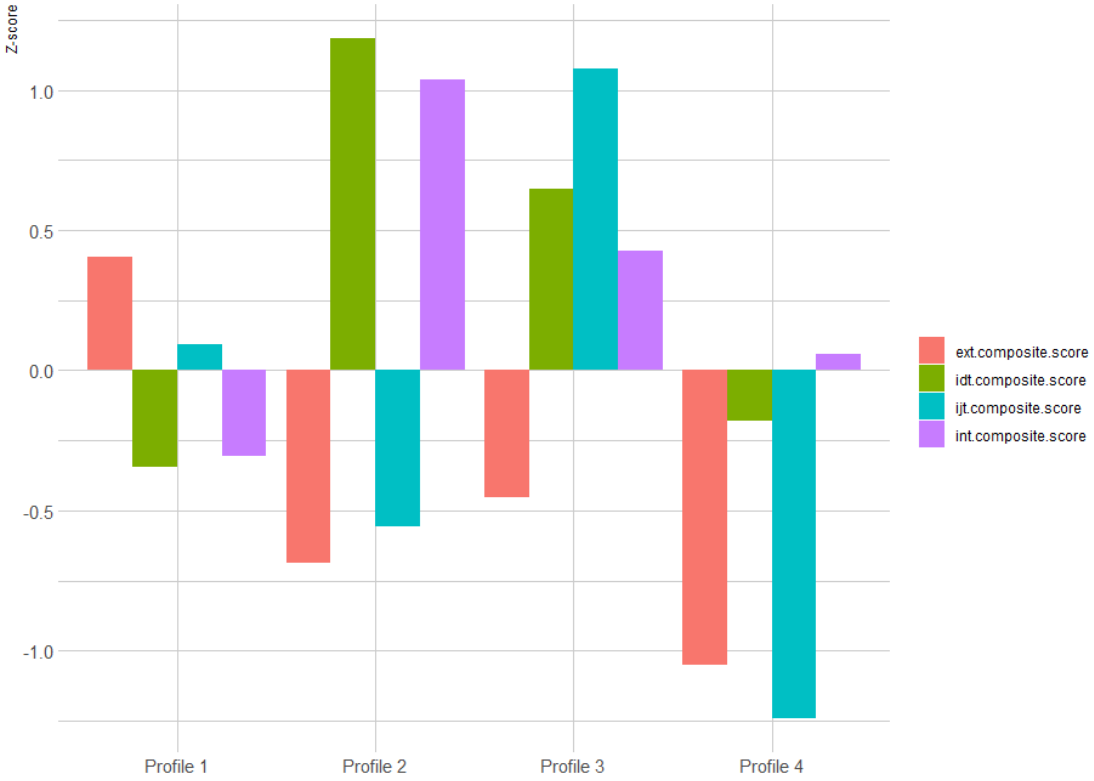


Figure 3.5: The profile plot from mclust for the best fitted model with the composite scores calculated by weighed factor loadings

4 Conclusion

In summary, this thesis has given a mathematical overview of latent profile analysis (LPA) and how to determine the best model fit. Furthermore we have illustrated how latent profile analysis can be applied to find motivational profiles of a student population in service mathematics education.

We have investigated how the latent profiles are created in LPA. This is done with the LPA model equation (equation 2.2). This equation consists among others of the unique model parameters means and covariances which characterize the profile distributions, and parameters for the probabilities of belonging to a profile. These parameters are estimated using maximum likelihood estimation via the EM algorithm.

Our main research question is *'What are the student profiles that can be distinguished based on students' motivation and are these profiles significantly different?'*. To answer this question we performed LPA on the data set of PRIME. LPA is performed in R using tidyLPA. During this analysis we have found that LPA only works when certain assumptions hold, namely the continuous profile indicators must be normally distributed within each latent profile and there must be unobserved population heterogeneity. We have also come to the conclusion that you have to convert your ordinal data into composite scores per profile indicator before you can use your data for LPA.

We have also found that LPA cannot estimate all models. Apparently the EM algorithm fails to converge for certain models. This is the case when there is singularity in the covariance matrix estimate. For a recommendation on how to continue the research for these models, see the discussion (chapter 5).

Using multiple fit indices we found that the best model fit is the model with varying variances and varying covariances with 4 profiles. By comparing the mean scores on the four types of motivations between the profiles we can label the four latent profiles as follows:

1. Poor quality: lowest scores on the identified and intrinsic motivations, highest score on external motivation and moderately high on introjected motivation.
2. Good quality: highest scores on the identified and intrinsic motivations and moderate scores on external and introjected motivations.
3. High quantity: high scores on the introjected, identified and intrinsic motivations and moderate score on external motivation.
4. Low quantity: low scores on the external and introjected motivations and moderately low scores on the identified and intrinsic motivations.

These labels of profiles are in line with the motivational profiles that were found in previous research given in section 3.1.1.

To check that the profiles are clearly differentiated, we conducted a Games-Howell test. We conclude from the results of this test that:

- Profiles 1 and 2 are significantly different based on all the types of motivation.
- Profiles 1 and 3 are significantly different based on all the types of motivation.
- Profiles 1 and 4 are significantly different based on the external, introjected and intrinsic motivation.
- Profiles 2 and 3 are significantly different based on introjected, identified and intrinsic motivation.
- Profiles 2 and 4 are significantly different based on all the types of motivation.
- Profiles 3 and 4 are significantly different based on external, introjected and identified motivation.

In our hypothesis in section 3.1.2 we used the research of Wijnia and Baars, 2021 that the number of student profiles that can be distinguished based on students' motivation is 2 to 6 with a mode of 4 profiles. Our best fitted model is a model with 4 profiles, so therefore our analysis is in line with previous studies and our hypothesis holds.
To compare our results with the results of the previous studies we also have a small research question. This sub research question is *'Are the identified profiles comparable to those found in previous studies?'*. Using the research of previous studies in Wijnia and Baars, 2021 our answer to this question is that our identified profiles are indeed comparable to those found in previous studies.

During this research, we have gained a better understanding on the latent profiles are created in LPA and the mathematical formulations underlying LPA. We have found that previous research on LPA in social sciences lack of the mathematical formulations of LPA. In this thesis, we have attempted to make clear how the the models in LPA is estimated and specified. The explanations in this thesis will be of value to social science researchers who would like to make use of LPA in their research. A good understanding of the statistics in LPA would allow researchers to be able to better interpret the results from the analysis.

5 Discussion

In this chapter some of the problems we encountered will be discussed and we will give some recommendations for further research.

5.1. Models that can not be estimated

By performing latent profile analysis in R using the tidyLPA package, an error was stated that some models could not be estimated, due to non-convergence (see section 3.4). These models are the models with varying variances and fixed covariances for 3 to 6 profiles and the model with varying variances and varying covariances for 6 profiles. The problem is caused by singularity in the covariance matrix estimate (Scrucca et al., 2016). In future research it is a good idea to investigate this further. By doing so, researchers will gain a better understanding on how to better specify models when using LPA.

Thus a recommendation is to try to figure out why it follows from singularity in a covariance matrix that certain models with specific numbers of profiles can not be estimated. Fraley and Raftery, 2007 suggest using Bayesian methods to avoid that models with singularity in the covariance matrix can not be estimated, is to use Bayesian methods. From their paper it appears that using a Bayesian regularisation makes it possible to get estimates for every model that we wanted to estimate, so future research could investigate this and analyse these results.

5.2. Comparison with other universities

In this thesis we have discussed the small research question if our identified profiles are comparable to those found in previous studies. To research the impact of the innovations in PRIME, it might also be a good possibility to compare our identified profiles with those found in previous studies at other universities. By doing so you can also compare different types of teaching methods between universities and then find out if another method can increase the learning motivations of the students.

5.3. Academic performance

Another way to research the impact of the innovations in PRIME is to investigate how the identified profiles are related to learning outcomes and student success. PRIME has also collected the grades for the mathematics course of the students who have filled in the survey. These grades can be used to investigate if there are relations between grades and some specific types of motivations or combinations of motivations. Due to lack of time we are not able to investigate this, so therefore a suggestion for future research is to look into this question.

5.4. ANOVA and Games-Howell test

We have performed the Games-Howell test in our research to check if the student profiles are clearly differentiated. Due to lack of time we did not investigated this test in detail, therefore a recommendation for future research could be to do more research in the mathematics of this statistical test and look closely into the assumptions.

5.5. Comparison tidyLPA and mclust

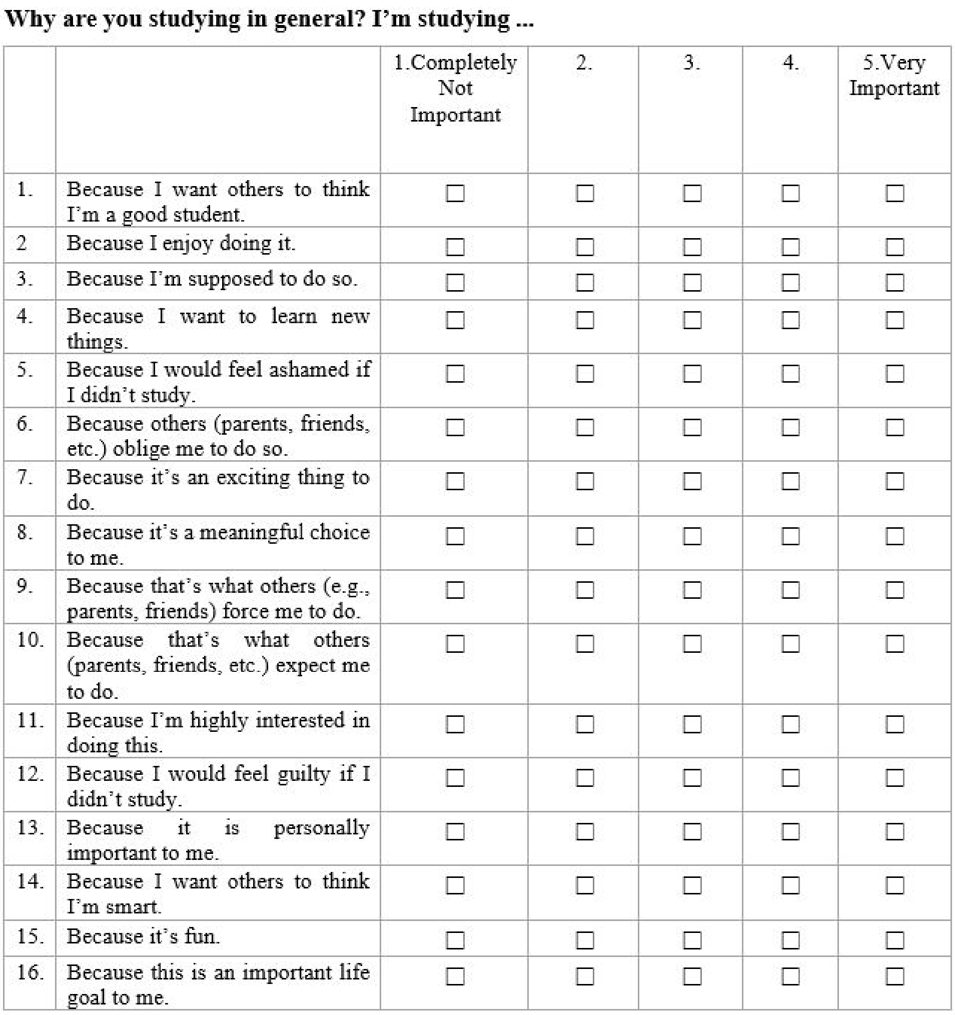
We have used both tidyLPA and mclust in R to perform latent profile analysis. We described both packages in section 3.3, but we did not really compare the two packages. Therefore we would recommend to do more research in the comparison of tidyLPA and mclust, see for example Wardenaar, 2021.

5.6. Previous research on LPA in the social sciences

During this research, we have found that it looks like that the research using LPA in social sciences just uses the statistical methods without checking the mathematics. For example the ANOVA test is used in many research of LPA to check if the profiles are clearly differentiated. But we showed in our research that the homogeneity of variance assumption of ANOVA is violated in our case. Maybe in these researches the mathematics is checked, but they just did not mention it. But this lack of mathematical formulation gives some uncertainties. Therefore this mathematical research is necessary to determine the validity of the statistical methods used.

A

Survey table



Why are you studying in general? I'm studying ...

		1.Completely Not Important	2.	3.	4.	5.Very Important
1.	Because I want others to think I'm a good student.	□	□	□	□	□
2	Because I enjoy doing it.	□	□	□	□	□
3.	Because I'm supposed to do so.	□	□	□	□	□
4.	Because I want to learn new things.	□	□	□	□	□
5.	Because I would feel ashamed if I didn't study.	□	□	□	□	□
6.	Because others (parents, friends, etc.) oblige me to do so.	□	□	□	□	□
7.	Because it's an exciting thing to do.	□	□	□	□	□
8.	Because it's a meaningful choice to me.	□	□	□	□	□
9.	Because that's what others (e.g., parents, friends) force me to do.	□	□	□	□	□
10.	Because that's what others (parents, friends, etc.) expect me to do.	□	□	□	□	□
11.	Because I'm highly interested in doing this.	□	□	□	□	□
12.	Because I would feel guilty if I didn't study.	□	□	□	□	□
13.	Because it is personally important to me.	□	□	□	□	□
14.	Because I want others to think I'm smart.	□	□	□	□	□
15.	Because it's fun.	□	□	□	□	□
16.	Because this is an important life goal to me.	□	□	□	□	□

Figure A.1: Statements of the survey

B

R-code LPA using median composite scores

```
library(dplyr)
library(lavaan)
library(psych)
library(semPlot)
library(tidyLPA)
library(mclust)

#read file
df_all <- read.csv(”Data_final_survey.csv”, header = TRUE, sep = ”;”)

# select only 16 items for motivation
df_motv <- df_all %>%
  select (1, 64:79) %>%
  na.omit()

### LPA with median scores ###
df_motv_median <- df_motv %>%
  rowwise() %>%
 mutate (ex_med = median(c(Motivation3,Motivation6,Motivation9,Motivation10)),
       ij_med = median(c(Motivation1,Motivation5,Motivation12,Motivation14)),
       id_med = median(c(Motivation4,Motivation8,Motivation13,Motivation16)),
       in_med = median(c(Motivation2,Motivation7,Motivation11,Motivation15)))

median_model <- df_motv_median%>%
  select(ex_med, ij_med, id_med, in_med) %>%
  single_imputation() %>%
  estimate_profiles(2:6,
                    variances = c(”equal”, ”varying”, ”equal”, ”varying”),
                    covariances = c(”zero”, ”zero”, ”equal”, ”varying”))

median_model

df_motv_median %>%
  select(ex_med, ij_med, id_med, in_med)%>%
  single_imputation() %>%
  scale()%>%
  estimate_profiles(2:6,
                    variances = c(”equal”, ”varying”, ”equal”, ”varying”),
                    covariances = c(”zero”, ”zero”, ”equal”, ”varying”))%>%
  compare_solutions(statistics = ”BIC”)

df_motv_median %>%
  select(ex_med, ij_med, id_med, in_med)%>%
  single_imputation() %>%
  scale()%>%
  estimate_profiles(2,
                    variances = c(”varying”),
                    covariances = c(”varying”))%>%
  plot_profiles()
```

C R-code tidyLPA

```
library(dplyr)
library(lavaan)
library(psych)
library(semPlot)
library(tidyLPA)
library(mclust)

#read file
df_all <- read.csv(”Data_final_survey.csv”, header = TRUE, sep = ”;”)

# select only 16 items for motivation
df_motv <- df_all %>%
  select (1, 64:79) %>%
  na.omit()

## Converting ordinal data into composite value

#function for weighted sd
weighted.sd <- function (x, w) {
  sum.w <- sum(w)
  sum.w2 <- sum (w^2)
  mean.w <- sum (x*w)/ sum(w)
  x.sd.w <-sqrt ((sum.w/ (sum.w^2-sum.w2)) * sum(w*(x-mean.w)^2))
  return(x.sd.w)
}

#function for rescale to create composite score
re.scale <- function(f.scores, raw.data, loadings) {
  fz.scores <- (f.scores + mean(f.scores))/(sd(f.scores))
  means <- apply(raw.data, 1, weighted.mean, w=loadings)
  sds <- apply(raw.data, 1, weighted.sd, w=loadings)
  grand.mean <- mean(means)
  grand.sd <-mean(sds)
  final.scores <- ((fz.scores * grand.sd)+ grand.mean)
  return (final.scores)
}

#function to apply to each motivation type
get.scores.fun <- function(data) {
  fact <- factanal(data, factors =1, scores =”regression”)
  f.scores <- fact $scores[,1]
  f.loads <- fact$loadings [,1]
  rescaled.scores <-re.scale(f.scores,data,f.loads)
  output.list <- list (rescaled.scores, f.loads)
  names(output.list) <- c(”rescaled.scores”, ”factor.loadings”)
  return(output.list)
}

# Subset data
ext <- df_motv %>% select (Motivation3,Motivation6, Motivation9, Motivation10)
ijt <- df_motv %>% select (Motivation1,Motivation5, Motivation12, Motivation14)
idt <- df_motv %>% select (Motivation4,Motivation8, Motivation13, Motivation16)
int <- df_motv %>% select (Motivation2,Motivation7, Motivation11, Motivation15)


###Apply function to get factor loadings
##extrinsic: extract rescaled factor scores
ext.score.loadings <- get.scores.fun(ext)
ext.composite.score<- ext.score.loadings$rescaled.scores

#introjected: extract rescaled factor scores
ijt.score.loadings <- get.scores.fun(ijt)
ijt.composite.score<- ijt.score.loadings$rescaled.scores


#identified: extract rescaled factor scores
idt.score.loadings <- get.scores.fun(idt)
idt.composite.score<- idt.score.loadings$rescaled.scores

#intrinsic: extract rescaled factor scores
int.score.loadings <- get.scores.fun(int)
int.composite.score<- int.score.loadings$rescaled.scores

### create data frame with composite scores
df_motv_composite <- data.frame(ext.composite.score,ijt.composite.score,
                                idt.composite.score, int.composite.score)


### Perform LPA
composite_model <- df_motv_composite%>%
  select(ext.composite.score, ijt.composite.score,
        idt.composite.score, int.composite.score) %>%
  single_imputation() %>%
  estimate_profiles(2:6,
                    variances = c(”equal”, ”varying”, ”equal”, ”varying”),
                    covariances = c(”zero”, ”zero”, ”equal”, ”varying”))
composite_model

composite_model_compare <- df_motv_composite%>%
  select(ext.composite.score, ijt.composite.score,
        idt.composite.score, int.composite.score) %>%
  single_imputation() %>%
  estimate_profiles(2:6,
                    variances = c(”equal”, ”varying”, ”equal”, ”varying”),
                    covariances = c(”zero”, ”zero”, ”equal”, ”varying”))%>%
  compare_solutions()
```

```
composite_model_compare

composite_model_best <- df_motv_composite%>%
  select(ext.composite.score, ijt.composite.score,
        idt.composite.score, int.composite.score) %>%
  single_imputation() %>%
  estimate_profiles(4,
                    variances = c(”varying”),
                    covariances = c(”varying”))
plot_profiles(composite_model_best)

composite_model_best_scaled <- df_motv_composite%>%
  select(ext.composite.score, ijt.composite.score,
        idt.composite.score, int.composite.score) %>%
  single_imputation() %>%
  scale()%>%
  estimate_profiles(4,
                    variances = c(”varying”),
                    covariances = c(”varying”))
plot_profiles(composite_model_best_scaled)

## Table means and standard deviations
output <- get_data(composite_model_best)

by(cbind(output = output$ext.composite.score,
         output= output$ijt.composite.score,
         output = output$idt.composite.score,
         output = output$int.composite.score),
   output$Class, describe)


## ANOVA

res.aov.ext <- aov(output$ext.composite.score~output$Class, data = output)

res.aov.ijt <- aov(output$ijt.composite.score~output$Class, data = output)

res.aov.idt <- aov(output$idt.composite.score~output$Class, data = output)

res.aov.int <- aov(output$int.composite.score~output$Class, data = output)


### Games-Howell-test
library(PMCMRplus)
resext <- gamesHowellTest(res.aov.ext)
resijt <- gamesHowellTest(res.aov.ijt)
residt <- gamesHowellTest(res.aov.idt)
resint <- gamesHowellTest(res.aov.int)

summaryGroup(resext)
summaryGroup(resijt)
summaryGroup(residt)
summaryGroup(resint)
```

D

R-code mclust

```
library(tidyverse)
library(mclust)
library(hrbrthemes)

explore_model_fit <- function(df_motv_composite, n_profiles_range = 1:6,
                             model_names = c(”EII”, ”VVI”, ”EEE”, ”VVV”)) {
 x <- mclustBIC(df_motv_composite, G = n_profiles_range, modelNames = model_names)
  y <- x %>%
    as.data.frame.matrix() %>%
    rownames_to_column(”n_profiles”) %>%
    rename(‘Constrained variance, fixed covariance‘ = EII,
           ‘Freed variance, fixed covariance‘ = VVI,
           ‘Constrained variance, constrained covariance‘ = EEE,
           ‘Freed variance, freed covariance‘ = VVV)
  y
}


fit_output <- explore_model_fit(df_motv_composite, n_profiles_range = 1:6)

library(forcats)

to_plot <- fit_output %>%
  gather(‘Covariance matrix structure‘, val, -n_profiles) %>%
  mutate(
    ‘Covariance matrix structure‘ = as.factor(‘Covariance matrix structure‘
    ),
    val = abs(val))
# this is to make the BIC values positive
(to align with more common formula / interpretation of BIC)


ggplot(to_plot, aes(x = n_profiles, y = val, color = ‘Covariance matrix structure‘,
        group = ‘Covariance matrix structure‘)) +
  geom_line() +
  geom_point() +
  ylab(”BIC (smaller value is better)”) +
  theme_ipsum_rc()

create_profiles_mclust <- function(df_motv_composite,
                                   n_profiles,
                                   variance_structure = ”freed”,
                                   covariance_structure = ”freed”){

  if (variance_structure == ”constrained” & covariance_structure == ”fixed”) {

    model_name <- ”EEI”

  } else if (variance_structure == ”freed” & covariance_structure == ”fixed”) {

    model_name <- ”VVI”

  } else if (variance_structure == ”constrained” & covariance_structure ==
    ”constrained”) {

    model_name <- ”EEE”

  } else if (variance_structure == ”freed” & covariance_structure == ”freed”) {

    model_name <- ”VVV”

  } else if (variance_structure == ”fixed”) {

   stop(”variance_structure cannot equal ’fixed’ using this function; change this to
    ’constrained’ or ’freed’ or try one of the models from mclust::Mclust()”)

  }

  x <- Mclust(df_motv_composite, G = n_profiles, modelNames = model_name)

  print(summary(x))

  dff <- bind_cols(df_motv_composite, classification = x$classification)

  proc_df <- dff %>%
    mutate_at(vars(-classification), scale) %>%
    group_by(classification) %>%
    summarize_all(funs(mean)) %>%
    mutate(classification = paste0(”Profile ”, 1:n_profiles)) %>%
    mutate_at(vars(-classification), function(x) round(x, 3)) %>%
    rename(profile = classification)

  return(proc_df)

}


m4 <- create_profiles_mclust(df_motv_composite, 4, variance_structure = ”freed”,
      covariance_structure = ”freed”) #best model from tidylpa

m4 %>%
  gather(key, val, -profile) %>%
  ggplot(aes(x = profile, y = val, fill = key, group = key)) +
  geom_col(position = ”dodge”) +
  ylab(”Z-score”) +
  xlab(””) +
  scale_fill_discrete(””) +
  theme_ipsum_rc()

#model 2 with 3 profiles, gives error
m2_3 <- create_profiles_mclust(df_motv_composite, 3, variance_structure = ”freed”,
        covariance_structure = ”fixed”)
```

Bibliography

Akogul, S., & Erisoglu, M. (2017). An approach for determining the number of clusters in a model-based cluster analysis. *Entropy*, *19*(9). https://doi.org/10.3390/e19090452

Bijma, F., Jonker, M., & van der Vaart, A. (2017). *An introduction to mathematical statistics* (Second). Amsterdam University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Deci, E. L., & Ryan, R. M. (2000). The ”what”and ”why”of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Fox, J. (2016). *Applied regression analysis and generalized linear models* (Third). SAGE Publications.

Fraley, C., & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, *24*, 155–181. https://doi.org/10.1007/s00357-007-0004-z

Hesamian, G. (2016). One-way anova based on interval information. *International Journal of Systems Science*, *47*(11), 2682–2690. https://doi.org/10.1080/00207721.2015.1014449

Johnson, S. K. (2021). Latent profile transition analyses and growth mixture models: A very non-technical guide for researchers in child and adolescent development. *New Directions for Child and Adolescent Development*, 111–139. https://doi.org/10.1002/cad.20398

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *12*(3), 209–229. http://www.jstor.com/stable/2986915

Lee, S., & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test? *Korean journal of anesthesiology*, *71*(5), 353–360. https://doi.org/10.4097/kja.d.18.00242

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons, Inc.

Morgan, G. B., Hodge, K. J., & Baggett, A. R. (2016). Latent profile analysis with nonnormal mixtures: A monte carlo examination of model selection using fit indices. *Computational Statistics & Data Analysis*, *93*, 146–161. https://doi.org/https://doi.org/10.1016/j.csda.2015.02.019

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson & M. Kaptein (Eds.), *Modern statistical methods for hci*. Springer International Publishing. https://doi.org/10.1007/978-3-319-26633-6_12

Pastor, D. A., Barron, K. E., Miller, B., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, *32*(1), 8–47. https://doi.org/10.1016/j.cedpsych.2006.10.003

Peugh, J., & Fan, X. (2013). Modeling unobserved heterogeneity using latent profile analysis: A monte carlo simulation. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 616–639. https://doi.org/10.1080/10705511.2013.824780

Prime. (n.d.). https://www.tudelft.nl/en/eemcs/the-faculty/departments/applied-mathematics/education/prime/

Rosenberg, J. M. (2020). Package 'tidylpa'. *CRAN*. https://cran.r-project.org/web/packages/tidyLPA/tidyLPA.pdf

Rosenberg, J. M. (2021). Introduction to tidylpa. https://data-edu.github.io/tidyLPA/articles/Introduction_to_tidyLPA.html

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., van Lissa, C. J., & Schmidt, J. A. (2018). Tidylpa: An r package to easily carry out latent profile analysis (lpa) using open-source or commercial software. *Journal of Open Source Software*, *3*(30), 978. https://doi.org/10.21105/joss.00978

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, *8*(1), 289–317. https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf

Scrucca, L., & Raftery, A. E. (2015). Improved initialisation of model-based clustering using gaussian hierarchical partitions. *Advances in data analysis and classification*, *9*(4), 447–460. https://doi.org/10.1007/s11634-015-0220-z

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*(4), 103445. https://doi.org/https://doi.org/10.1016/j.jvb.2020.103445

Starkweather, J. (2012). How to calculate empirically derived composite or indicator scores. *Benchmarks Online*. https://it.unt.edu/sites/default/files/benchmarks-02-2012.pdf

Tein, J.-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 640–657. https://doi.org/10.1080/10705511.2013.824781

Vansteenkiste, M. (2009). Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of educational psychology*, *101*(3), 671–688. https://doi.org/10.1037/a0015083

Wardenaar, K. J. (2021). Latent profile analysis in r: A tutorial and comparison to mplus. 10.31234/osf.io/wzftr

Wijnia, L., & Baars, M. (2021). The role of motivational profiles in learning problem-solving and self-assessment skills with video modeling examples. *Instructional Science*, *49*, 67–107. https://doi.org/10.1007/s11251-020-09531-4

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079