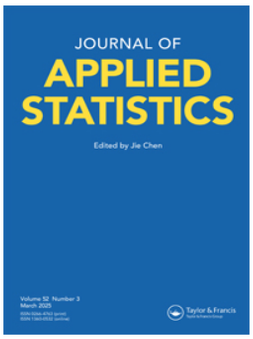

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cjas20

On the improved estimation of the normal mixture components for longitudinal data

Tapio Nummi, Jyrki Möttönen, Pasi Väkeväinen, Janne Salonen & Timothy E. O'Brien

To cite this article: Tapio Nummi, Jyrki Möttönen, Pasi Väkeväinen, Janne Salonen & Timothy E. O'Brien (07 Feb 2025): On the improved estimation of the normal mixture components for longitudinal data, Journal of Applied Statistics, DOI: 10.1080/02664763.2025.2459293

To link to this article: https://doi.org/10.1080/02664763.2025.2459293

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 07 Feb 2025.

Submit your article to this journal

Article views: 104

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=cjas20

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2025.2459293

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

On the improved estimation of the normal mixture components for longitudinal data

Tapio Nummi [a], Jyrki Möttönen [b], Pasi Väkeväinen [a], Janne Salonen [c] and Timothy E. O'Brien [d]

[a]Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland; [b]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland; [c]Finnish Public Sector Pension Provider Keva, Helsinki, Finland; [d]Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL, USA

ABSTRACT

When analyzing real data sets, statisticians often face the question that the data are heterogeneous and it may not necessarily be possible to model this heterogeneity directly. One natural option in this case is to use the methods based on finite mixtures. The key question in these techniques often is what is the best number of mixtures or, depending on the focus of the analysis, the best number of sub-populations when the model is otherwise fixed. Moreover, when the distribution of the response variable deviates from meeting the assumptions, it's common to employ an appropriate transformation to align the distribution with the model's requirements. To solve the problem in the mixture regression context we propose a technique based on the scaled Box-Cox transformation for normal mixtures. The specific focus here is on mixture regression for longitudinal data, the so-called trajectory analysis. We present interesting practical results as well as simulation experiments to demonstrate that our method yields reasonable results. Associated R-programs are also provided.

ARTICLE HISTORY
Received 23 March 2023
Accepted 18 January 2025

KEYWORDS
Box-Cox transformation; finite mixtures; mixture regression; number of mixture components; trajectory analysis

1. Introduction

A finite normal mixture model (see e.g. [5]) is useful to provide a sensible model in many practical data analysis applications, and this includes its extensions such as the skew-normal distribution [16] or the t-distribution [15]. Nonetheless, finding the best possible composition of densities, and the number of mixture components among many competing alternatives can be a complicated and challenging task. The number of mixture components needed may depend on the sample size, assumed component distribution, as well as the possible outcome transformation (scale of measurements) applied. We emphasize here that approximation of the density of the outcome (or response) variable, and identification of the possible number of sub-populations are often slightly separate tasks. The latter refers

CONTACT Tapio Nummi tapio.nummi@tuni.fi Tampere University, Kalevantie 4, 33014 Tampere, Finland

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

to the estimation of the number of possible latent classes, whereas the former refers to the best fit to the actually observed outcome distribution. The focus here is on the estimation of the number of latent classes or sub-populations.

In this paper we are focusing on the so-called *trajectory analysis* (TA), which is an application of finite mixture modeling for longitudinal data [19,20]. The method is widely used in many fields of science, and practitioners oftentimes face non-normal data distributions. It is quite possible that without proper outcome transformation the method may reveal incorrect number of sub-populations. Methodologically we propose a technique that is based on the *scaled Box-Cox transformation* [10,29] that makes the likelihood based inference possible over transformed outcomes. The subsequent analyses are then based on the components of transformed normal mixtures.

The proposed modeling approach has several advantages. First, the theory of TA with normal mixtures is well established and a number of implementations as software packages, such as packages in R, already exist. Second, a suitable outcome transformation can reduce the risk of generating either too many or too few trajectory groups, which can happen, for example, when the observed distribution is highly skewed.

Our theoretical and practical experience has shown that the proposed technique will make it much easier to present a reasonable solution to the problem of the best number of possible sub-populations. Consequently, it can also be argued that we can better rely on the interpretation of the practical meaning of the obtained solution, which is a very important feature when considering various real world data situations.

The approach presented here differs fundamentally from more traditional longitudinal analyses such as mixed models or the GEE approach (see e.g. [6]). This is because the data does not need to be homogeneous over the observed time interval. In fact, interest here often lies in the latent groups that generate this heterogeneity. These groups frequently reveal intriguing developmental paths that might remain concealed in a more traditional longitudinal analysis. Additionally, they can be utilized in further analyses when attempting to identify possible explanatory variables behind the latent groups that have been identified.

This paper is laid out as follows. In Section 2 we outline the normal mixture model in the trajectory analysis context. In Section 3 we introduce a transformation, which is based on the scaled Box-Cox transformation. In Section 4 we give an overview of the R code used in this study, and in Sections 5 and 6 we demonstrate the transformation with two real world data sets, as well as with two simulation experiments. Section 7 concludes our study with our summary. Additional plots are provided in Appendix 1. Appendix 2 includes all the necessary R code and instructions on how to use the developed `scaledbc` R package, available on GitHub, to implement the proposed transformation techniques. These tools can also be easily adapted for use with the reader's own longitudinal datasets.

2. Normal mixtures

Denote random vectors of longitudinal measurements as $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip_i})'$, where $\boldsymbol{y}_i$ presents the sequence of measurements on individual i over p_i periods, and the marginal probability density of $\boldsymbol{y}_i$ with possible time-dependent covariates $\boldsymbol{X}_i$ as $f(\boldsymbol{y}_i \mid \boldsymbol{X}_i)$, where

$i = 1, \ldots, n$. It is assumed that $f(\boldsymbol{y}_i \mid \boldsymbol{X}_i)$ follows a mixture of K densities

$$f(\boldsymbol{y}_i \mid \boldsymbol{X}_i) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i \mid \boldsymbol{X}_i), \quad \sum_{k=1}^{K} \pi_k = 1 \quad \text{with } \pi_k > 0,$$

where π_k is the probability of belonging to the cluster k and $f_k(\boldsymbol{y}_i \mid \boldsymbol{X}_i)$ is the density for the kth cluster. If the multivariate normal distribution is assumed we have for the kth component

$$f_k(\boldsymbol{y}_i \mid \boldsymbol{X}_i) = (2\pi)^{-p_i/2} |\boldsymbol{\Sigma}_{ik}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_{ik})' \boldsymbol{\Sigma}_{ik}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{ik})\right),$$

where $\boldsymbol{\mu}_{ik} = \boldsymbol{\mu}_{ik}(\boldsymbol{\theta}_k, \boldsymbol{X}_i)$ is a function of covariates $\boldsymbol{X}_i$ with parameter $\boldsymbol{\theta}_k$, and $\boldsymbol{\Sigma}_{ik}(\boldsymbol{\sigma}_k)$ is the variance-covariance matrix within the kth cluster involving a vector of unique covariance parameters $\boldsymbol{\sigma}_k$. In the most general case $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_{ik}$ are the unstructured mean and covariance matrices. However, often some more parsimonious structures are imposed either on $\boldsymbol{\mu}_k$ and/or $\boldsymbol{\Sigma}_{ik}$.

Here we focus on the normal linear model (i.e. the so-called *Trajectory analysis*; see e.g. [19,20]). It is then assumed that

$$\boldsymbol{\mu}_{ik} = \boldsymbol{X}_i \boldsymbol{\theta}_k \quad \text{and} \quad \boldsymbol{\Sigma}_{ik} = \sigma_k^2 \boldsymbol{I}_{p_i},$$

where $k = 1, \ldots, K$. Note that this conditional independence assumption within the kth cluster does not mean independence over the whole sample when the number of clusters K is greater than 1.

Because the number of covariance parameters to be estimated is, due to simplifying assumptions, significantly reduced, it also greatly simplifies the log-likelihood function,

$$l(\boldsymbol{\phi} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i \mid \boldsymbol{X}_i)$$

to be maximized, and where $\boldsymbol{\phi} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \ldots, \boldsymbol{\theta}_K', \sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2, \pi_1, \pi_2, \ldots, \pi_{K-1})'$ is the full vector of free parameters to be estimated. These model assumptions have trade-offs. On the one hand, in many cases this computational simplicity can be of great advantage. On the other, it may have an effect on the types of groups the estimation algorithm aims to identify from the data. In line with the requirements given in [19], we assume here independent and homogeneous errors within the trajectory group. Note in passing, however, that allowing correlated and/or heterogeneous errors could possibly yield another group composition. This can also be the case if some mixture distribution other than normal is assumed.

In empirical applications, often a practical or heuristic interpretation of the identified groups is jointly emphasized with statistical evaluation criteria. As noted, our aim here is to present a novel approach for the identification and estimation of the possible subpopulations (latent groups) from the observed data.

The literature on mixture regression modeling is extensive. This includes books [17,20,21] as well as review articles such as [22,34]. In the context of longitudinal data analysis, mixture regression models based on the normal distribution are frequently utilized. Among these, trajectory analysis (also called *group-based trajectory modeling*) [19], *latent*

class growth analysis (LCGA), and *growth mixture models* (GMM) (see e.g. [13]) are typically the most frequently used. In GMMs, the parameters of the mean growth function can be perceived as a combination of fixed and random parameters within trajectory groups. On the other hand, LCGA constitutes a specific case of GMM, where variance and covariance parameters of growth parameters are assumed to be fixed at zero. This in turn leads to individuals within the group being rather homogeneous, resulting in a model similar to the trajectory analysis model.

The practitioner can select the number of clusters using, for example, the Akaike information criterion (AIC) [1]

$$\mathrm{AIC}(K) = -2l(\boldsymbol{\phi} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) + 2h,$$

the Bayesian information criterion (BIC) [27]

$$\mathrm{BIC}(K) = -2l(\boldsymbol{\phi} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) + \log(N)h,$$

or the Integrated completed likelihood (ICL) [3]

$$\mathrm{ICL}(K) = -2l_{\mathrm{ICL}}(\boldsymbol{\phi} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) + \log(N)h,$$

where

$$l(\boldsymbol{\phi} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i \mid \boldsymbol{X}_i)\right),$$

$$l_{\mathrm{ICL}}(\boldsymbol{\phi} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} [\pi_k f_k(\boldsymbol{y}_i \mid \boldsymbol{X}_i)]^{z_{ik}}\right),$$

h is the total number of free parameters, $N = \sum_{i=1}^{n} p_i$ is the total number of measurements and $z_{ik} = 1$ if $\boldsymbol{y}_i$ arises from cluster k and zero otherwise.

In the simulation study in Section 6 the BIC criterion has very precisely identified the correct number of clusters when the data was transformed using the scaled Box-Cox transformation. The ICL criterion performed slightly worse – and AIC clearly worse – than the BIC criterion. As such, in the empirical applications in Section 5, we focus on the BIC information criterion.

3. Power transformations

A commonly-used transformation method is the Box-Cox transformation, see [4]. For all the observations $y_{ij} > 0$ the Box-Cox transformation is

$$y_{ij}^{(\lambda)} = \begin{cases} (y_{ij}^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \log(y_{ij}), & \lambda = 0, \end{cases} \tag{1}$$

where λ is the transformation parameter. For fixed λ the log-likelihood function of the transformed observations $y_{ij}^{(\lambda)}$ becomes as

$$l^*(\boldsymbol{\phi}, \lambda) = l(\boldsymbol{\phi} \mid \boldsymbol{y}_1^{(\lambda)}, \ldots, \boldsymbol{y}_n^{(\lambda)}) + \log(J(\lambda)),$$

where

$$J(\lambda) = \tilde{y}^{N(1-\lambda)}$$

is the Jacobian and

$$\tilde{y} = \left(\prod_{i=1}^{n} \prod_{j=1}^{p_i} y_{ij} \right)^{1/N}$$

is the geometric mean. Maximization of the function $l^*(\boldsymbol{\phi}, \lambda)$ with respect to $\boldsymbol{\phi}$ gives the ML estimates $\hat{\boldsymbol{\phi}}_* = (\hat{\boldsymbol{\theta}}'_{*1}, \hat{\boldsymbol{\theta}}'_{*2}, \ldots, \hat{\boldsymbol{\theta}}'_{*K}, \hat{\sigma}^2_{*1}, \hat{\sigma}^2_{*2}, \ldots, \hat{\sigma}^2_{*K}, \hat{\pi}_{*1}, \hat{\pi}_{*2}, \ldots, \hat{\pi}_{*,K-1})'$.

Since the likelihood functions are now different for different values of λ the fits can not be compared with direct likelihood-based inference. Instead, here we apply the *scaled* transformation [4,10,29]

$$w_{ij}^{(\lambda)} = \begin{cases} (y_{ij}^{\lambda} - 1)/\lambda \tilde{y}^{\lambda-1}, & \lambda \neq 0 \\ \tilde{y} \times \log(y_{ij}), & \lambda = 0. \end{cases} \tag{2}$$

Since in this case $\log(J(\lambda)) = 0$ the log-likelihood function of the scaled observations becomes

$$l^{**}(\boldsymbol{\phi}, \lambda) = l(\boldsymbol{\phi} \mid \boldsymbol{w}_1^{(\lambda)}, \ldots, \boldsymbol{w}_n^{(\lambda)}), \tag{3}$$

where $\boldsymbol{w}_i^{(\lambda)} = (w_{i1}^{(\lambda)}, \ldots, w_{ip_i}^{(\lambda)})'$, $i = 1, \ldots, n$. It turns out that maximization of l^{**} for transformed observations with fixed λ is now equivalent to maximization of the original $l(\boldsymbol{\phi} \mid \cdot)$ with observations y_{ij} substituted by scaled observations $w_{ij}^{(\lambda)}$. As a result, models can now be compared under the log-likelihood $l(\boldsymbol{\phi} \mid \boldsymbol{w}_1^{(\lambda)}, \ldots, \boldsymbol{w}_n^{(\lambda)})$ with the likelihood based inference. Next, we denote

$$\hat{\boldsymbol{\phi}}_w = (\hat{\boldsymbol{\theta}}'_{w1}, \hat{\boldsymbol{\theta}}'_{w2}, \ldots, \hat{\boldsymbol{\theta}}'_{wK}, \hat{\sigma}^2_{w1}, \hat{\sigma}^2_{w2}, \ldots, \hat{\sigma}^2_{wK}, \hat{\pi}_{w1}, \hat{\pi}_{w2}, \ldots, \hat{\pi}_{w,K-1})'$$

as the ML estimate obtained from the log-likelihood

$$l(\boldsymbol{\phi} \mid \boldsymbol{w}_1^{(\lambda)}, \ldots, \boldsymbol{w}_n^{(\lambda)}).$$

For the case where $p_1 = p_2 = \ldots = p_n$, we can convert the parameter estimates to the original Box-Cox situation using

$$\begin{cases} \hat{\boldsymbol{\theta}}_{*k} = \tilde{y}^{\lambda-1} \hat{\boldsymbol{\theta}}_{wk}, \\ \hat{\sigma}_{*k} = \tilde{y}^{\lambda-1} \hat{\sigma}_{wk}, \\ \hat{\pi}_{*k} = \hat{\pi}_{wk}, \end{cases} \tag{4}$$

where $k = 1, \ldots, K$. A slight bias is introduced in this conversion; see discussion about this topic in [10]. Nevertheless, if the original measurement scale is required, e.g. for generating plots, it can be readily obtained, given that each individual's group membership can then be easily identified using the calculated posterior probabilities.

To summarize the estimation procedure, the selected number of trajectory groups is assessed across various values denoted by K. The value of K that gives the optimal fit to the data can be chosen utilizing model selection criteria such as BIC, AIC, or ICL. Note that the criteria BIC and AIC are here based on $l(\boldsymbol{\phi} \mid \boldsymbol{w}_1^{(\lambda)}, \ldots, \boldsymbol{w}_n^{(\lambda)})$ and the integrated

completed likelihood is based on $l_{\mathrm{ICL}}(\boldsymbol{\phi} \mid \boldsymbol{w}_1^{(\lambda)}, \ldots, \boldsymbol{w}_n^{(\lambda)})$. The transformation parameter λ is estimated at multiple calculation points for each K value, and the optimal value is determined using the log-likelihood analysis. Following this methodology, the parameter estimates ϕ are computed with both K and λ held constant, based on their respective chosen and estimated values.

As was discussed in the introduction, *over-extraction* and *under-extraction* of the number of trajectory groups is a fundamental issue in finite mixture modeling. The over-extraction of mixture components was brought into attention by Bauer and Curran [2] pointing out that as finite mixture models can have two separate purposes, the results can have two separate interpretations. The mixture components can be a meaningful approximation of a sub-population structure, but often with non-normal data, the mixture components will carry no useful information regarding the structure of the population. Bauer and Curran [2] demonstrate this in a simulation experiment, showing that a two-component solution can provide the best fit to non-normal data generated from a unitary distribution. A transformation used to normalize the data before the analysis could enable the researcher to look for sub-populations in situations where the analysis would otherwise use mixture components as tools to approximate a non-normal distribution. Relatedly, see also an overview of the discussion and literature given in [31].

4. Implementation of the scaled Box-Cox transformation in R

To facilitate our analysis, computations were carried out in the R Environment for Statistical Computing [25]. All the proposed methods are implemented in our R package `scaledbc` which is available in GitHub. Our implementation is based on the function `stepFlexmix` in R package `flexmix` [7–9,14]. For a tutorial of installing and using our R package `scaledbc`, see Appendix 2. Our package contains three functions for fitting the mixture model: `stepmixl_orig`, which fits the model onto the original, untransformed data, `stepmixl_bf`, which implements a brute-force method of the transformation, performing a grid-search of all λ values entered by the user as a parameter, and `stepmixl_opt`, which is an optimized version of the transformation method that searches the range $[-5, 5]$ with 0.04 interval using 33 calculation points. See [31] for an explanation of the optimization algorithm. We next demonstrate the use of our algorithm using both typical data as well as simulated data.

The provided package also contains some tools for model selection with criteria AIC, BIC and ICL, residual diagnostics in the form of a Shapiro-Wilks test and evaluating clustering performance using cluster purity.

5. Some real data applications

5.1. Birth weight distribution

For our first illustration, we start with the simple analysis of birth weight distribution. The original data set contains several growth measurements of five birth cohorts of 4,223 children collected in Finland [33]. For this example we study the birth weight of the cohort born in 2001 ($n = 766$). The data are plotted in Figure 1. For these data, our aim is to test

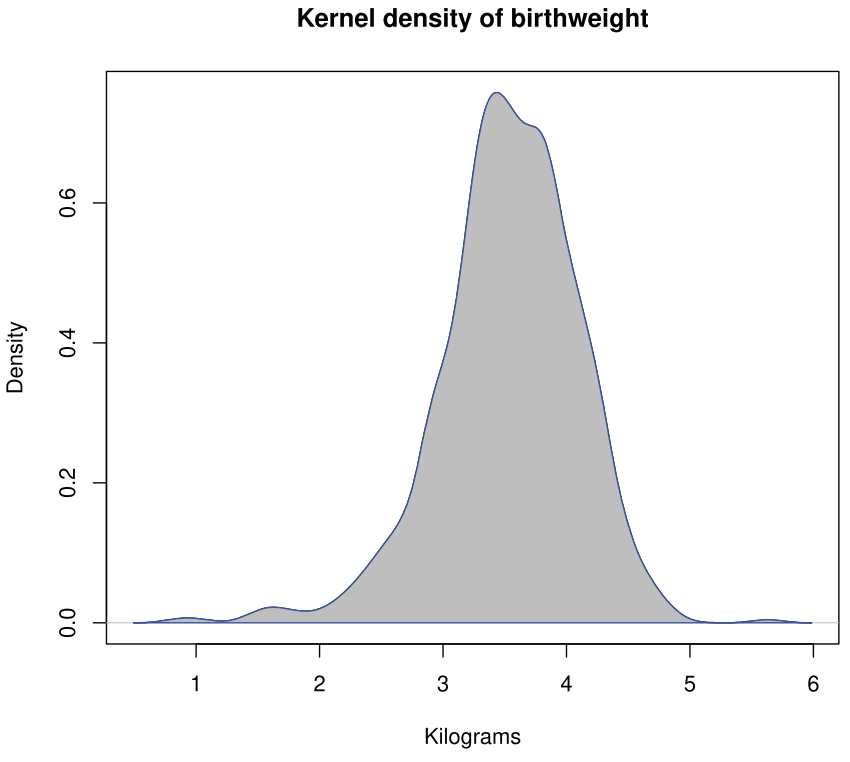

Figure 1. Smoothed birth weight distribution of the Finnish cohort born in 2001 ($n = 766$).

how many normal mixture components are needed, and if more than one component is needed, to determine whether it would be possible to transform the distribution into one component form.

For these data, direct application of mixture modeling clearly points out to two-component normal mixture model with Bayesian information criterion values, (BIC($K = 1$) $= 1299.784$ and BIC($K = 2$) $= 1264.076$). Interestingly, when we apply the scaled transformation given in Equation (2) to these data and compare the fits, with BIC($K = 2, \lambda = 1$) $= 1270.72$ and BIC($K = 1, \lambda = 1.94$) $= 1263.48$, the analysis suggests the one-population model, and where λ is now estimated using the maximum likelihood method (see also the Figure A5 in Appendix 1). These results indicate that this is in fact only a one-population model, whereas if the transformation had not been done, one would have ended up with a two-population model and this may not be the best solution if the focus is indeed on the estimation of the number of sub-populations. This illustration underscores that a clear omnibus criterion is needed to highlight that mixtures should be chosen in conjunction with a possible transformation.

5.2. Alcohol consumption trajectories

Our second illustration involves the Northern Sweden Cohort Study [12], which consists of pupils who in 1981 attended the last year of compulsory school (age 16) in Luleå (in northern Sweden). A total of 1,005 individuals participated in all follow-up surveys in 1983,

Table 1. Results of the fits of the mixture models with $K = 1, \ldots, 7$ for $\lambda \in [-2, 2]$ with transformation parameter estimates $\hat{\lambda}$, estimated values of the log-likelihood function *l* and the values of information criterion AIC and BIC.

K	$\hat{\lambda}$	*l*	AIC	BIC
1	0.16	−13631.51	27273.02	27298.75
2	0.00	−13358.42	26736.85	26791.46
3	0.00	−13261.83	26553.68	26637.18
4	0.00	−13201.68	26443.38	26555.77
5	0.00	−13143.63	26337.32	26478.60
6	0.00	−13113.49	26286.99	**26457.16**
7	0.00	−13097.14	**26264.31**	26463.37

1986, 1995 and 2007. For this illustration, we analyzed the alcohol consumption (converted into absolute alcohol in centiliters) of the subset of female participants at the age of 16, 18, 21, 30 and 42 years.

The distribution of alcohol consumption is *highly skewed* and there is also a high probability of zero observations (i.e. these are *semi-continuous* data). One possible solution to this kind of situation is to apply mixture modeling which provides tools to both of the problems discussed above. Here the so-called *broken stick* model was applied for transformed observations:

$$y^{(\lambda)} = \beta_0 + \beta_1 t + \beta_2 (t - \kappa_1)_+ + \varepsilon,$$

where t is age, $\kappa_1 = 21$, $(\cdot)_+$ equals$(\cdot)$ if $(\cdot) \geq 0$ and 0 otherwise. The chosen model is thus a continuous curve of combined straight lines at the age of 21.

As noted in Table 1, using the AIC measure, the minimum is obtained for $K = 7$, whereas with BIC measure, the minimum is obtained for $K = 6$. Our choice here is $K = 6$, and this is in line with the earlier studies of these data [23,32].

The estimate of the transformation parameter $\hat{\lambda}$ suggest the log transformation be used here for each of the tested $k = 1, \ldots, 7$. The estimated respective mixture proportions are $\pi_1 = 0.2669$, $\pi_2 = 0.1510$, $\pi_3 = 0.0777$, $\pi_4 = 0.0894$, $\pi_5 = 0.2789$ and $\pi_6 = 0.1361$, and this is noted in the sub-headings of Figure 2,.

Next, identified trajectories are illustrated by fitting smoothed curves for each of the trajectory groups, as shown in Figure 2. Note that group 4 is a low consumption group whereas groups 6 is a high consumption group. The largest group, however, is group 5 where alcohol consumption starts at the age of 18 years (when it is legal to buy alcohol) and the consumption stabilizes at a constant level for older ages.

5.3. Wage earnings trajectories preceding disability retirement

In the Finnish statutory pension system, there is occupational rehabilitation and earnings-related disability pension available for the working age employees in case of declining work-ability. Declining work-ability leading to permanent disability may cause difficulties in employment and career opportunities over the life-course. For our next practical illustration, we study wage earnings trajectories of Finnish municipal sector employees retiring in 2017, using data from administrative registers of Finnish Public Sector Pension Provider

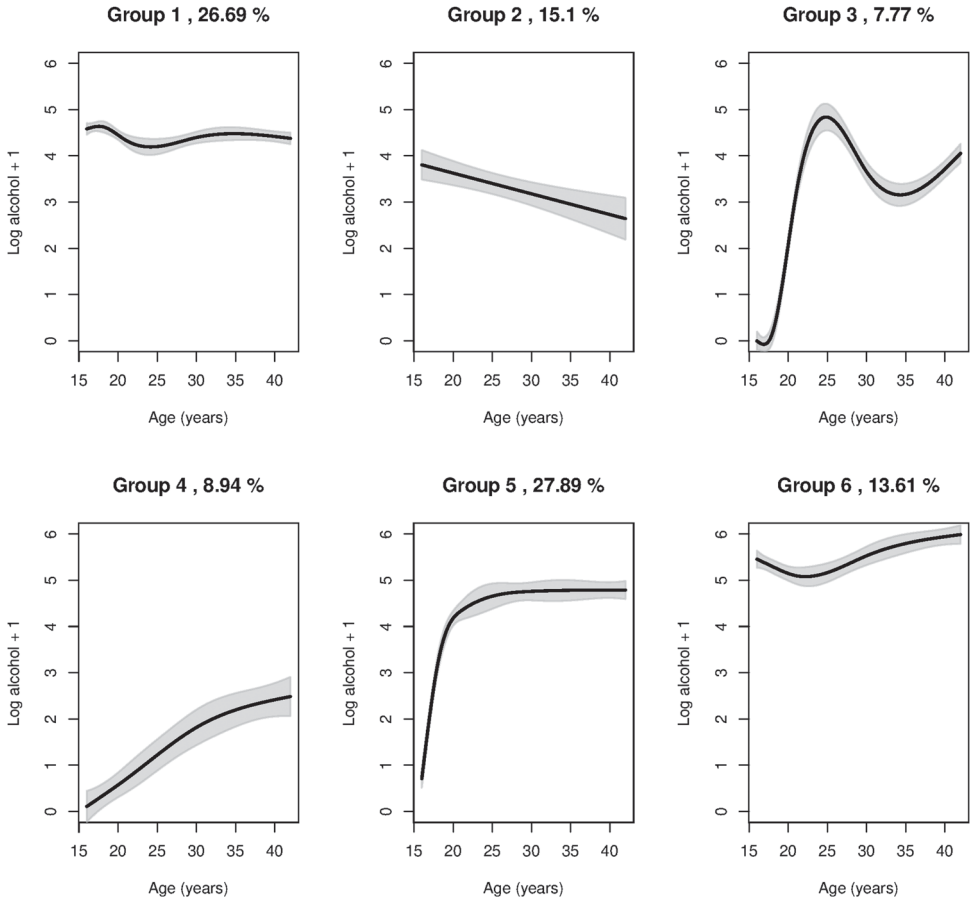

Figure 2. Illustration of the identified alcohol consumption trajectories. Smoothed log-transformed alcohol consumption is plotted as a function of age for each trajectory group. The gray area around the smoothed curve is the approximate 95% confidence interval, and the group size estimates are given in individual plot titles.

Keva. The data consists of a sample of 448 individuals born in 1956, whose wage earnings are followed for 13 years, specifically from 2005 to 2017.

The distribution of wage earnings overall is *right-skewed* and there is also a high probability of zero observations (14%), which again indicates semi-continuous data. Oftentimes wage trajectories are modeled with age polynomials (see e.g. [24]). Here the following quadratic model was applied across trajectory groups:

$$y^{(\lambda)} = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon,$$

where y is yearly wage, t is age and λ is a transformation parameter. As regards the transformed analysis, we can see from Table 2 that BIC minimum value is obtained for $K = 6$. The estimate of the transformation parameter $\hat{\lambda} = 0.16$ suggest the log transformation is appropriate. The estimated mixture proportions are then $\pi_1 = 0.0870$, $\pi_2 = 0.1495$, $\pi_3 = 0.1986$, $\pi_4 = 0.0691$, $\pi_5 = 0.1808$ and $\pi_6 = 0.3147$. Note that the corresponding

Table 2. Results of the fits of the mixture models with $k = 1, \ldots, 7$ for $\lambda \in [-2, 2]$ with transformation parameter estimates $\hat{\lambda}$, estimated values of the log-likelihood function l and the values of information criterion AIC and BIC.

k	$\hat{\lambda}$	l	AIC	BIC
1	0.48	−61993.7	123995.4	124022.1
2	0.12	−57753.1	115524.2	115584.2
3	0.08	−56947.6	113923.3	114016.7
4	0.12	−56654.9	113348.0	113474.7
5	0.12	−56160.5	112369.0	112529.1
6	0.16	−56090.5	**112239.1**	**112432.6**
7	0.20	−56264.7	112577.5	112737.6

analysis with information criteria indicate eight groups for the original non-transformed data (results not given here).

Identified trajectories are illustrated by the group averages displayed in Figure 3. We see that for the large groups (3, 5 and 6) wages are not affected by the upcoming disability retirement (i.e. approximately 70 percent). These groups constitute the main part of the sample. For groups 2 and 4, wages (and employment) are strongly affected by the forthcoming permanent disability. For group 1, where wages have initially risen strongly, the earnings seems to fall slightly before permanent disability.

6. Simulation studies

6.1. Study 1: Testing the performance of the method

In addition to the previous real-world illustrations, we also investigated the performance of our proposed technique with simulated data sets. The simulation-design used here is based on the extensive work presented in [31]. Our aim here is to investigate how well our technique is able to estimate the actual value of the transformation parameter λ and the actual number of sub-populations. As such, our results also provide interesting empirical information of how well the information criteria used works in practice.

We used three different longitudinal models in our simulation experiments. One hundred data sets (each with sample size 200) from each model were generated.

Our models were as follows:

$$\text{One-cluster model (M1):} \quad y_{ij} = e^{1+0.2x_j+\epsilon_{ij}},$$

$$\text{Two-cluster model (M2):} \quad y_{ij} = \begin{cases} e^{1+\epsilon_{ij}}, & \text{with } \pi_1 = 0.6 \\ e^{1+x_j+\epsilon_{ij}}, & \text{with } \pi_2 = 0.4 \end{cases}$$

and

$$\text{Three-cluster model (M3):} \quad y_{ij} = \begin{cases} e^{1+\epsilon_{ij}}, & \text{with } \pi_1 = 0.5 \\ e^{1+x_j+\epsilon_{ij}}, & \text{with } \pi_2 = 0.3 \\ e^{1+x_j+x_j^2+\epsilon_{ij}}, & \text{with } \pi_3 = 0.2 \end{cases}$$

Here ϵ_{ij} are independently distributed as $\epsilon_{ij} \sim N(0, 0.5^2)$, $x_j = \frac{1}{9} \times (j-1)$, $j = 1, \ldots, 10$ and $i = 1, \ldots, 200$. Note that Model M1 is a one-population model, Model M2 is a two-population model and Model M3 is a three-population model.

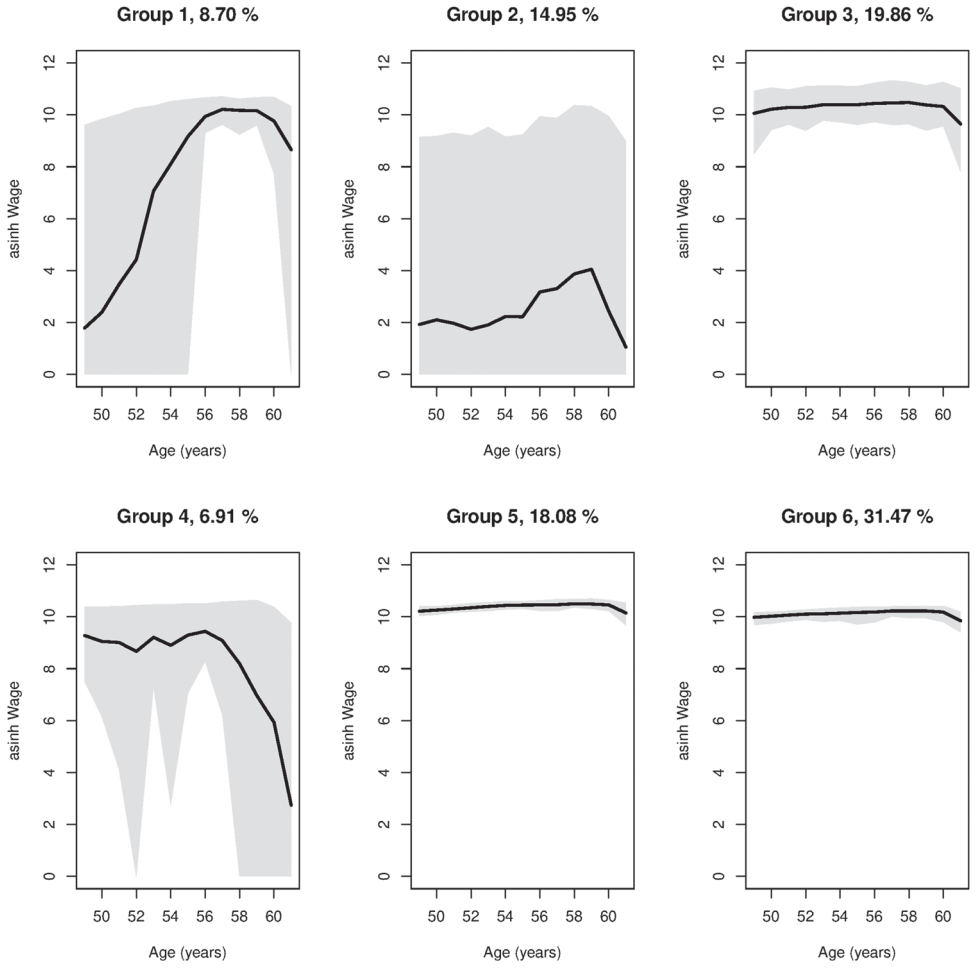

Figure 3. Illustration of the identified wage trajectories. Group averages are plotted by age. The gray area around the average curve is the approximate 95% confidence interval, and the group size estimates are given in individual plot titles.

Next, λ, the Box-Cox parameter, and K, the number of sub-populations, were estimated from each simulated data set. The scaled Box-Cox transformation was tested with $\lambda \in [-2, 2]$ using grid search with 0.05 intervals. Estimation was repeated 10 times for each tested λ value to ensure the convergence. The estimated values of λ were found to vary between -0.3 and 0.15 which is indeed quite close to the nominal value of $\lambda = 0$. The results are summarized in Table 3. We note that using our proposed methodology with the scaled Box-Cox transformation, identification of the number of clusters work superbly for these data sets. It is also interesting to see how different model selection criteria work here. Clearly, the BIC measure performed best and ICL slightly worse. These results notwithstanding, however, it should be underscored that the results obtained are limited to the chosen experimental setting; for example, the models and dimensions of the data sets may influence the performance.

Table 3. Identified number of clusters k using the values of information criterion functions AIC, BIC and ICL for the following models applied to the transformed data under three different settings: the one-cluster model M1, the two-cluster model M2 and the three-cluster model M3.

	AIC			BIC			ICL		
k	$M1$	$M2$	$M3$	$M1$	$M2$	$M3$	$M1$	$M2$	$M3$
1	85	0	0	100	0	0	100	0	0
2	12	83	0	0	100	0	0	100	8
3	3	15	85	0	0	100	0	0	92
4	0	1	13	0	0	0	0	0	0
5	0	1	2	0	0	0	0	0	0

Table 4. Identified number of clusters k using the values of information criterion functions AIC, BIC and ICL for the following models: the one-cluster model M1, the two-cluster model M2 and the three-cluster model M3, applied to untransformed data.

	AIC			BIC			ICL		
k	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	0	0	0	0	0	0	0	0	0
2	11	0	0	75	0	0	88	17	0
3	51	0	5	17	27	6	6	73	13
4	19	28	7	5	67	35	5	18	61
5	19	72	88	3	6	59	1	2	26

Next, as shown in Table 4 the corresponding analysis was made for the same data sets without the scaled Box-Cox transformation. Clearly these results lead one to conclude that when no transformation is made (or even considered), an incorrect determinations of the number of groups can easily be made. Specifically, recall that for models 1, 2 and 3, the corresponding number of groups are also 1, 2, and 3, respectively, but Table 4 underscores that when no transformation is made, each of the information measures consistently indicate a larger number of latent groups. Readers concerned solely with model fit to the original distribution, for which numerous methods have been developed, may choose other methods, but this is not the focus of our study.

6.2. Study 2: A comparison with skew-t growth mixture model

We next selected the skew-t GMMs method [13,18] for comparison since it is quite commonly used and offers a sufficiently flexible and versatile alternative to the method we have presented. It is well justified to choose a setup for the simulation that basically does not favor either of the methods under consideration. For this purpose, we chose for the independent random errors $\epsilon_{ij} \sim Gamma(2, 1)$, that produces rather skewed errors. The following two-cluster setups were considered here

$$\text{Model 4:} \quad y_{ij} = \begin{cases} 1 + \epsilon_{ij}, & \text{with } \pi_1 = 0.5 \\ 1 + 2x_j + \epsilon_{ij}, & \text{with } \pi_2 = 0.5 \end{cases}$$

and

$$\text{Model 5:} \quad y_{ij} = \begin{cases} 1 + \epsilon_{ij}, & \text{with } \pi_1 = 0.8 \\ 1 + 2x_j + \epsilon_{ij}, & \text{with } \pi_2 = 0.2 \end{cases}$$

Here $i = 1, \ldots, n_k, j = 1, \ldots, 6$, with $x_j = \frac{j-1}{5}$, $n_{k_1} = 200$, $n_{k_2} = 500$ and $\epsilon_{ij} \sim Gamma$ (2, 1). Thus, note that we investigated two different sample sizes and two mixture proportions, with each model replicated one hundred times to ensure thorough testing. Table 5 provides the results of the four different setups (Table 6) used in the simulation; see [31] for additional details.

We conducted computations utilizing an optimized implementation of the transformation method, which again utilizes the R package *flexmix* (Appendix 2). For skew-t GMMs, we utilized the Mplus software ([18] and [13]) through the R interface provided by the MplusAutomation library [11]. We utilized 40 different starting values for skew-*t* GMMs, with the default value of 500 iterations. For the transformation method, we employed the default value of 200 from the *flexmix* package. The computations were done using Linux Mint 21.3, Intel® Core™ i7-8700 @ 3.20 GHz × 6. We were able to speed up the skew-*t* GMM estimation program significantly by using parallel computing. Unfortunately, the *flexmix* program does not have a parallel computation option. It is also worth noting that since R is an interpreted language primarily utilized for statistical computing and graphics, it may not exhibit the same efficiency as implementations based on languages such as C++ for computationally intensive tasks. In comparing methods, our assessment focused on the following three key aspects:

A: **Selection of correct number of clusters.** We employed models selected via the Bayesian Information Criterion (BIC) [27] for this purpose.
B: **Agreement between identified and true clusters.** For this purpose, we utilized the Rand index [26].
C: **Computational efficiency.** We measured the running times. It's important to note that the runs were conducted in entirely different computational environments, thus making direct exact comparisons unfeasible.

In what follows, we provide a comprehensive summary of the findings of our simulation study 2.

Aspect A: Selection of the Correct Number of Clusters. With a sample size of $n = 200$, the optimized Box-Cox transformation method (BC) significantly excelled in identifying the

Table 5. Identified number of clusters *k* using Bayesian information criterion (BIC) for the optimized transformation method (BC) and for the skew-t GMM method (ST).

	S1		*S2*		*S3*		*S4*	
k	*BC*	*ST*	*BC*	*ST*	*BC*	*ST*	*BC*	*ST*
1	0	87	13	99	0	14	0	87
2	99	13	86	1	76	86	100	12
3	1	0	1	0	24	0	0	0
4	0	0	0	0	0	0	0	1

Note: S1-S4 are the four different setups listed in Table 6.

Table 6. Summary of four different simulation setups used in simulation study 2.

Setup	Model	n	π_1	π_2
*S*1	4	200	0.5	0.5
*S*2	5	200	0.8	0.2
*S*3	4	500	0.5	0.5
*S*4	5	500	0.8	0.2

correct number of groups for setups S1 and S2. Conversely, with a sample size of $n = 500$, the skew-t growth mixture model (ST) exhibited slight superiority under conditions of identical relative proportions (S3). However, when the relative proportions were not identical (S4), the transformation method once again demonstrated clear superiority. Only in setup S3 did the ST method exhibit slightly better performance. One plausible explanation for this slightly unexpected outcome lies in the functionality of the criterion utilized. While

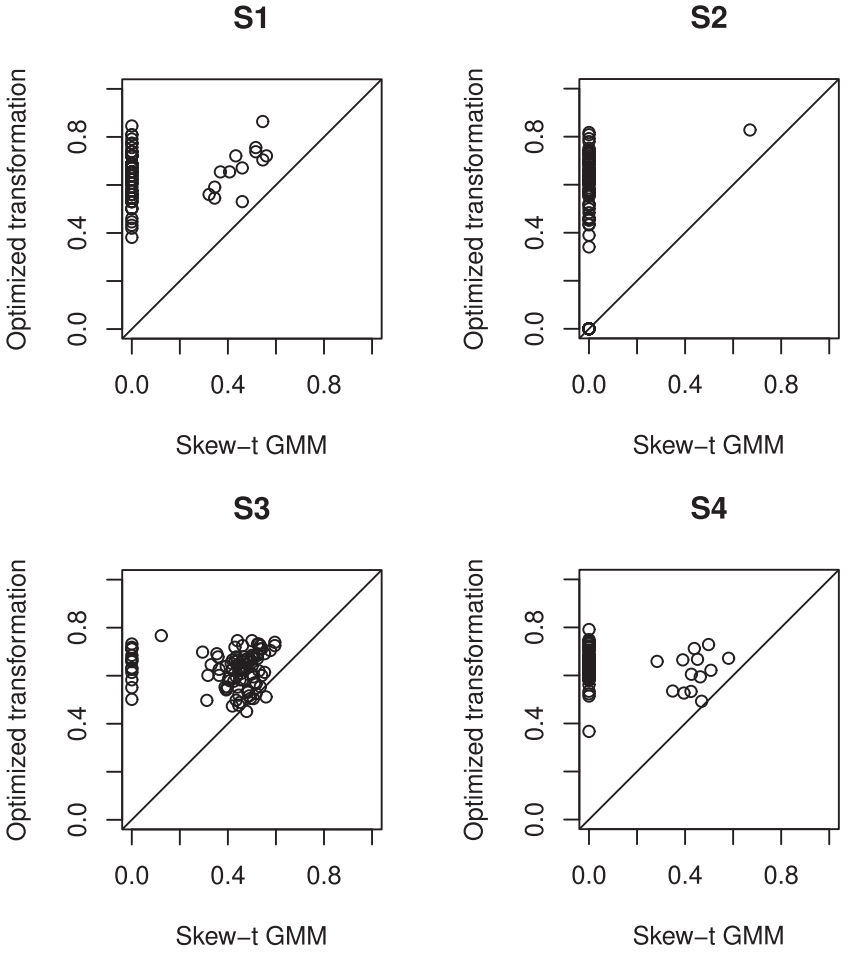

Figure 4. Adjusted Rand indices comparing true classifications to the estimated classifications. S1-S4 are the four different setups.

Table 7. User times (in seconds) for computing estimates of the number of clusters with the optimized transformation method and the skew-t GMM method.

Setup	Method	Min.	Q1	Median	Mean	Q3	Max.
*S*1	Scaled Box-Cox	85.06	112.70	124.20	126.20	137.70	187.6
	Skew-t GMM	26.15	30.92	33.79	38.04	37.66	129.4
*S*2	Scaled Box-Cox	89.84	116.90	141.60	145.30	165.70	223.6
	Skew-t GMM	20.92	27.99	30.35	38.77	34.74	278.6
*S*3	Scaled Box-Cox	152.90	186.40	205.30	209.70	229.40	280.8
	Skew-t GMM	59.50	70.98	74.87	80.95	85.10	155.1
*S*4	Scaled Box-Cox	164.60	209.30	239.10	247.10	281.40	347.3
	Skew-t GMM	53.19	66.91	73.07	73.85	78.62	139.8

Note: The skew-t GMM estimates were calculated using parallel computing. S1-S4 are the four different setups.

the Bayesian Information Criterion (BIC) is generally employed in these models, there are instances, particularly with larger sample sizes (≥ 500), where the adjusted BIC version [28] may be deemed more favorable in determining the optimal number of clusters (see [30], page 183).

Aspect B: Agreement between identified and true clusters. The adjusted Rand indices are plotted in Figure 4 for the simulated data sets (setups S1, S2, S3 and S4), computed using the BC and ST methods, again with BIC used for cluster identification. With values ranging between 0 and 1 and 1 representing perfect agreement with true groupings, these indices offer insights into method performance. Overall, it appears that the BC method consistently outperformed the ST method across all the chosen setups considered here.

Aspect C: Computational efficiency. In the analysis of run-time performance, an initial observation highlights that the BC method exhibited longer average run-times compared to the ST method (see Table 7). This discrepancy was largely anticipated, given that the ST method utilized on parallel computing with 4 cores, while BC utilized only one core. Adjusting for this discrepancy by dividing the average run-times by four, for example, would give a more favorable rating for the BC method. Further examination reveals that as sample sizes increased from two hundred to five hundred, run-times for the ST method nearly doubled, whereas the increase in run-times was slightly less pronounced for the BC method. This suggests a degree of robustness in the BC method, especially considering the disparity in computing resources between the two methods. Summarizing these observations, it may be argued that under entirely comparable environments, the BC method might provide at least as effective an alternative as ST method.

It was found that the transformation method BC performed better than skew-t GMM method ST on the identification of the correct number of clusters as well as producing clusters that represent the sub-populations more accurately. The run-times are, of course, heavily dependent on the computer capacity and on the practical implementation of the methods. In sum, it appears the BC method offers a computationally efficient alternative when compared to the ST method.

7. Concluding remarks

Our work underscores that one of the advantages of using the log-likelihood of the scaled observations is that it provides a general framework that can be used in the analysis of mixtures. With the proposed approach, models can easily be compared with likelihood based inference and with ordinary model selection tools, such as AIC, BIC or ICL information

criteria. In particular, we believe that the method we propose helps to better estimate the number of possible sub-populations. The proposed technique can be implemented with standard statistical software developed for finite mixture models, and this is facilitated by the provided R functions.

In terms of follow-up research and recommendations, conducting slightly more comprehensive simulation experiments and comparing our method with several alternative approaches using varying sample sizes would be of interest; we hope our work will spark further results in this regard. One interesting future topic would be to investigate how to apply the technique to the situation of a multivariate trajectory analysis, where several response variables are jointly used to identify trajectory groups. Additionally, exploring different distributions beyond the mixture normal distribution within our method could provide valuable insights. Moreover, investigating the performance of the method under less stringent assumptions regarding the correlation structure would be equally intriguing.

Acknowledgments

We would like to thank the editor and two anonymous referees for their constructive criticism, which improved our manuscript in many ways. Additionally, T. E. O'Brien also gratefully acknowledges the financial support of the Fulbright Finland Foundation for funding his research visit.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethics approval and consent to participate

The Alcohol consumption study was approved by the Regional Ethical Review Board in Umeå, Sweden. Response to a survey was considered as consent to participate in the study. Moreover, participation was voluntary and participants could drop out at any time they chose during the follow-up period.

ORCID

Tapio Nummi https://orcid.org/0000-0001-8682-6951
Jyrki Möttönen https://orcid.org/0000-0002-6270-2556
Pasi Väkeväinen https://orcid.org/0000-0002-2680-3501
Janne Salonen https://orcid.org/0000-0002-0595-6226
Timothy E. O'Brien https://orcid.org/0000-0002-2408-7281

References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Autom. Control 19 (1974), pp. 716–723. Available at https://doi.org/10.1109/TAC.1974.1100705.
[2] D.J. Bauer and P.J. Curran, *Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes*, Psychol. Methods 8 (2003), pp. 338–363. Available at https://psycnet.apa.org/doi/10.1037/1082-989X.8.3.338.
[3] C. Biernacki, G. Celeux, and G. Govaert, *Assessing a mixture model for clustering with the integrated completed likelihood*, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000), pp. 719–725. Available at https://doi.org/10.1109/34.865189.

[4] G.E.P. Box and D.R. Cox, *An analysis of transformations*, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 26 (1964), pp. 211–252. Available at https://www.jstor.org/stable/2984418.
[5] B. Everitt and D. Hand, *Finite Mixture Distributions*, Chapman & Hall, London, 1981. Available at https://doi.org/10.1007/978-94-009-5897-5.
[6] G.M. Fitzmaurice, N.M. Laird, and J.H. Ware, *Applied Longitudinal Analysis*, John Wiley & Sons, Hoboken, NJ, 2012.
[7] B. Grün and F. Leisch, *Fitting finite mixtures of generalized linear regressions in R*, Comput. Stat. Data Anal. 51 (2007), pp. 5247–5252. Available at https://doi.org/10.1016/j.csda.2006.08.014.
[8] B. Grün and F. Leisch, *FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters*, J. Stat. Softw. 28 (2008), pp. 1–35. Available at https://www.jstatsoft.org/v28/i04/.
[9] B. Grün and F. Leisch, *flexmix: Flexible Mixture Modeling* (2023). Available at https://CRAN.R-project.org/package=flexmix, R package version 2.3-19.
[10] M.J. Gurka, L.J. Edwards, K.E. Muller, and L.L. Kupper, *Extending the Box–Cox transformation to the linear mixed model*, J. R. Stat. Soc. Ser. A (Stat. Soc.) 169 (2006), pp. 273–288. Available at https://www.jstor.org/stable/3559673.
[11] M.N. Hallquist and J.F. Wiley, *MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus*, Struct. Equ. Modeling 25 (2018), pp. 621–638. Available at https://www.tandfonline.com/doi/full/10.1080/10705511.2017.1402334.
[12] A. Hammarström, *The Northern Swedish Cohort*, Available at https://ki.se/en/imm/the-northern-swedish-cohort (2021). Accessed: 11 Feb 2022.
[13] T. Jung and K.A.S. Wickrama, *An introduction to latent class growth analysis and growth mixture modeling*, Soc. Personal. Psychol. Compass 2 (2008), pp. 302–317. Available at https://doi.org/10.1111/j.1751-9004.2007.00054.x.
[14] F. Leisch, *FlexMix: A general framework for finite mixture models and latent class regression in R*, J. Stat. Softw. 11 (2004), pp. 1–18. Available at https://www.jstatsoft.org/v11/i08/.
[15] T.I. Lin, J.C. Lee, and W.J. Hsieh, *Robust mixture modeling using the skew t distribution*, Stat. Comput. 17 (2007), pp. 81–92. Available at https://doi.org/10.1007/s11222-006-9005-8.
[16] T.I. Lin, J.C. Lee, and S.Y. Yen, *Finite mixture modelling using the skew normal distribution*, Stat. Sin. 17 (2007), pp. 909–927. Available at https://www3.stat.sinica.edu.tw/statistica/oldpdf/a17n35.pdf.
[17] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, Hoboken, NJ, 2000. Available at http://dx.doi.org/10.1002/0471721182.
[18] L.K. Muthén and B.O. Muthén, *Mplus (Version 8)[computer software].(1998–2017)*, Los Angeles, CA: Muthén & Muthén (2017).
[19] D.S. Nagin, *Analyzing developmental trajectories: A semiparametric, group-based approach*, Psychol. Methods 4 (1999), pp. 139–177. Available at https://doi.org/10.1037/1082-989X.4.2.139.
[20] D. Nagin, *Group-Based Modeling of Development*, Harvard University Press, Cambridge, MA, 2005.
[21] S.K. Ng, L. Xiang, and K.K.W. Yau, *Mixture Modelling for Medical and Health Sciences*, 1st ed., Chapman & Hall/CRC, New York, 2019. Available at https://doi.org/10.1201/9780429195181.
[22] H.L. Nguena Nguefack, M.G. Pagé, J. Katz, M. Choinière, A. Vanasse, M. Dorais, O.M. Samb, and A. Lacasse, *Trajectory modelling techniques useful to epidemiological research: A comparative narrative review of approaches*, Clin. Epidemiol. 2020 (2020), pp. 1205–1222. Available at https://doi.org/10.2147/CLEP.S265287.
[23] T. Nummi, *Some models and their extensions for longitudinal analyses*, in *Data Science and SDGs: Challenges, Opportunities and Realities*, B.K. Sinha and M.N.H. Mollah, eds., Springer, Singapore, 2021, pp. 15–30. Available at https://doi.org/10.1007/978-981-16-1919-9_2.
[24] T. Nummi, J. Möttönen, J. Salonen, and T.E. O'Brien, *Some applications of mixture regression modelling techniques*, Gujarat J. Stat. Data Sci. 38 (2022), pp. 151–165. Available at https://www.thegsa.in/wp-content/uploads/2023/06/gjsds-vol-38-issue-01-oct-2022.pdf.
[25] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. Available at https://www.R-project.org/.

[26] W.M. Rand, *Objective criteria for the evaluation of clustering methods*, J. Am. Stat. Assoc. 66 (1971), pp. 846–850.
[27] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. 6 (1978), pp. 461–464. Available at http://www.jstor.org/stable/2958889.
[28] S.L. Sclove, *Application of model-selection criteria to some problems in multivariate analysis*, Psychometrika 52 (1987), pp. 333–343.
[29] J. Spitzer, *Variance estimates in models with the Box-Cox transformation: Implications for estimation and hypothesis testing*, Rev. Econ. Stat. 66 (1984), pp. 645–652. Available at https://www.jstor.org/stable/1935988.
[30] A. Tolvanen, *Latent Growth Mixture Modeling: A Simulation Study, 111*, University of Jyväskylä, Jyväskylä, 2007.
[31] P. Väkeväinen, *Longitudinal mixture modeling with Box-Cox transformation: The effects of Box-Cox transformation on latent class identification in trajectory analysis*, Master's thesis, Tampere University, 2022. Available at https://urn.fi/URN:NBN:fi:tuni-202204213372.
[32] P. Virtanen, T. Nummi, T. Lintonen, H. Westerlund, B. Hägglöf, and A. Hammarström, *Mental health in adolescence as determinant of alcohol consumption trajectories in the Northern Swedish Cohort*, Int. J. Public Health 60 (2015), pp. 335–342. Available at https://doi.org/10.1007/s00038-015-0651-5.
[33] N. Vuorela, *Body Mass Index, Overweight and Obesity Among Children in Finland - A Retrospective Epidemiological Study in Pirkanmaa District Spanning Over Four Decades*, Tampere University Press, Tampere, 2011. Available at https://urn.fi/urn:isbn:978-951-44-8446-9.
[34] H. Zhang and Y. Huang, *Finite mixture models and their applications: A review*, Austin. Biom. Biostat. 2 (2015), pp. 1013. Available at https://austinpublishinggroup.com/biometrics/fulltext/biometrics-v2-id1013.pdf.

Appendices

Appendix 1. Visualization of the model fit in birth weight application

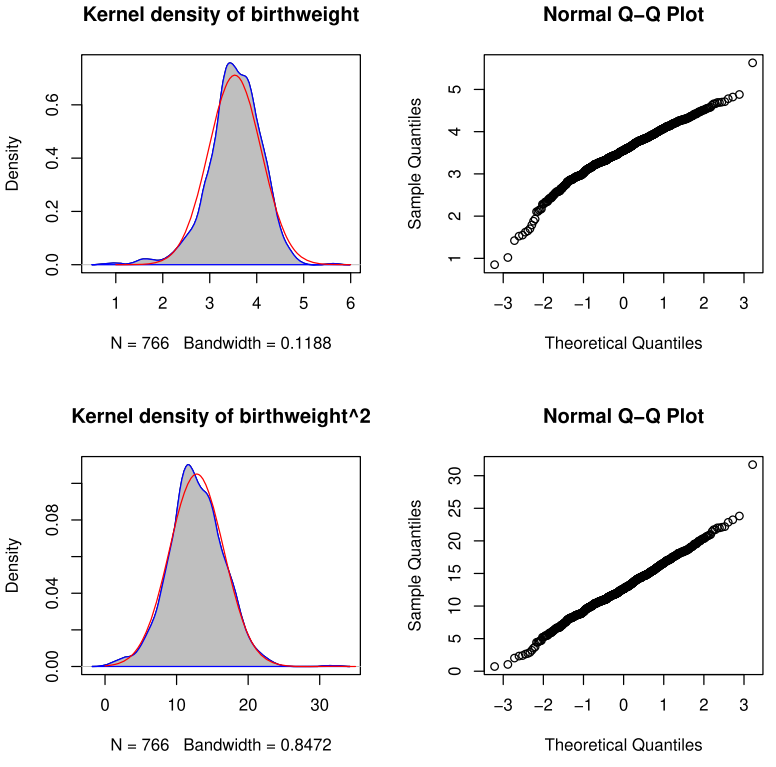

Figure A5. Density plots and fitted distributions with corresponding qqplot.

Appendix 2. Usage of the functions in the R package scaledbc

```
# Installing the R package scaledbc from GitHub using devtools.
# To install devtools:
install.packages("devtools")
# To install and load scaledbc
devtools::install_github("jymottonen/scaledbc")
library(scaledbc)
# Note: packages need to be installed only once, but loaded
#       after every R reboot.

# Data generation using the getSample function.
# Parameters:
# n: number of individuals
# p: number of measurements for each individual
# a: the intercept (applies to each sub-population)
# b: linear growth (sub-populations 2 and 3)
# c: quadratic growth (sub-population 3)
# pi: the sub-population prior probabilities
#     in this example a one-population sample with
#     linear growth is generated.
# v: variance parameter
set.seed(877)
data <- getSample(n=200, p=10, a=1, b=0.2, c=0, pi=c(0,1,0), v=0.5)
# Note: function getSample2 can be used to generate data from a
#       Gamma distribution.
#
# The function stepmixl_orig analyses the data without a transformation
# and outputs results in a similar form as the other functions.
#
# Parameters:
# y: the response variable
# x: the predictor variables
# id: the identifiers of individuals
# K: number of components (vector of multiple values)
# classes: known classes of individuals,
#          used for calculating cluster purity (optional)
ex_orig <- stepmixl_orig(y=data[[1]]$y, x=cbind(data[[1]]$time,
    data[[1]]$time2), id=data[[1]]$id, K=1:5, classes=data[[3]])
# Summary of the key metrics
summary(ex_orig)
# A plot of model selection criteria AIC, BIC and ICL
plot(ex_orig)
# Prior probability estimates of the 2-component model
ex_orig$models[[2]]@prior
# The proportion of clusters where residuals passed
# a Shapiro-Wilk test of normality at alpha = 0.05
ex_orig$pass
# Cluster purity will always be 1 with one-population data
ex_orig$purity

# The function stepmixl_bf performs a brute-force grid search
# with all the lambda calculation points passed in parameter lambdas
ex_bf <- stepmixl_bf(y=data[[1]]$y,
    x=cbind(data[[1]]$time, data[[1]]$time2), id=data[[1]]$id,
    K=1:5, classes=data[[3]], lambdas=seq(-2,2,0.05))
summary(ex_bf)
plot(ex_bf)
# The prior probability of the 1-component model will of course be 1
ex_bf$models[[1]]@prior
ex_bf$pass
ex_bf$purity

# The function stepmixl_opt uses an optimized implementation of the
# transformation. This implementation uses 33 calculation points
# of lambda to search the range [-5, 5] with 0.04 intervals.
ex_opt <- stepmixl_opt(y=data[[1]]$y,
    x=cbind(data[[1]]$time, data[[1]]$time2),
    id=data[[1]]$id, K=1:5, classes=data[[3]])
summary(ex_opt)
plot(ex_opt)
ex_opt$models[[1]]@prior
ex_opt$pass
ex_opt$purity
```