Multilevel Mixture with Known Mixing Proportions: Applications to School and Individual Level Overweight and Obesity Data from Birmingham, England

By

Shakir Hussain¹ and Ghazi Shukur²

¹School of health and population since, University of Birmingham, Birmingham, UK.
²Department of Economic and Statistics, Linnaeus University, Sweden,
and Department of Economics, Finance and Statistics, Jönköping International Business School, Sweden

Abstract

Multilevel (ML) models allow for total variation in the outcome to be decomposed as level one and level two or ‘individual and group’ variance components. Multilevel Mixture (MLM) models can be used to explore unobserved heterogeneity that represents different qualitative relationships in the outcome. In this paper, we extend the standard MLM by introducing constraints to guide the MLM algorithm towards a more appropriate data partitioning. Our constraint-based methods combine the mixing proportions estimated by parametric Expectation Maximization (EM) of the outcome and the random component from the ML model. This forms a new Multilevel Mixture known (MLMk) mix method. This framework allows for smaller residual variances and permits meaningful parameter estimates for distinct classes in the coefficient space. We also provide an illustrative example demonstrating the advantage of the MLMk compared with the MLM approach. We show the benefit of our method using overweight and obesity from Body Mass Index (BMI) measurements for students in year 6. We apply these methods on multi-level BMI data to estimate student multiple deprivation and school sport effects.

Keywords: Parametric Expectation Maximization, Multilevel Mixture, Multilevel Mixture Known Mix, Overweight and Obesity Data.
1. Introduction

The methodology we present here is related to the general statistical method for the analysis of data with heterogeneous outcomes in nested and non-nested structures, for example, see [5]. Because traditional approaches to multilevel data have some deficiencies and disadvantages in presenting the heterogeneous outcome distribution, the mixture model, as an alternative approach, overcomes some of these deficiencies and provides proper outcome structure.

[7] developed an efficient Multilevel Mixture Model (MLM) that is more general and [10] illustrated MLM with three applications. The term multilevel mixture can have two meanings. The first refers to latent class models in which the probability of class membership is predicted by one or more covariates; this is a ‘known class’, that is, the class membership is a function of the predictors, see [12]. However, in other contexts, MLM is also used to refer to some specified number of latent classes as part of estimating the regression model; this is an ‘unknown class’, that is, the class membership is a function of both the outcome and the covariates [2]. [11] categorize the outcome and run separate regression on each class.

We focus on exploring outcome heterogeneity using parametric Expectation Maximization (EM) methods. These methods have been proposed for the exponential family by [7]. The EM algorithm is used to maximize the latent class model log-likelihood function. This optimal partitioning strategy can accommodate both repeated and individual outcomes.

Our constraint-based methods rely on user-provided constraints to guide the MLM algorithm towards a more appropriate data partitioning. We assume that some pre-existing knowledge about the desired partitioning is available and we provide this knowledge in the form of constraints, i.e., estimating the class membership of each individual.

The advantage of constraint-based modelling is that smaller residual variances provide meaningful individual and group parameter estimates in the latent coefficient space.
In a simple example we compare the performance of two models, Multilevel Mixture ‘unknown class’ MLM and Multilevel Mixture constraint-base ‘known class’ MLMk. The residual variance is smaller for all classes of MLMk compared to MLM. Significant estimated coefficients of the distinct latent at individual and group level are obtained. The disadvantage is that the Bayesian Information Criterion “BIC” shows smaller values for MLM.

Note that we need to run MLMk in two stages; in the first stage we provide constraints (labels) to guide the MLM algorithm by using parametric EM to generate a class indicator and, in the second stage, we run the multilevel mixture using the indicator generated in the first stage.

It is well known that overweight and obesity are key components of the Body Mass Index (BMI) distributions. Excessive BMI is defined as overweight when it exceeds the censoring point. A major problem in overweight research is to determine the degree to which subjects react under specific individual and group characteristics. Investigators have long struggled with the problem of differentiating subject ‘s’ homogeneity. A drawback to these approaches is that they do not accommodate the possibility that subjects may belong to different class memberships. The problem of identifying class membership can be formulated in terms of medically meaningful interpretations. One approach is to cluster the data. A closely related methodology is model-based clustering using a finite mixture model, see [9].

Our new approach is to build a student level screening device to uncover distributions within mixture models. Hence in this problem we shall be looking for a hidden component first, which could manifest itself as a heterogeneous component of the data. As the initial inspection of the overall data distribution does not indicate heterogeneity, we shall not make any prior assumptions about the nature of the possible heterogeneity. We shall attempt to fit a mixture of normal distributions to the BMI as a mixture of several normal variables using the parametric EM method. Covariate association with BMI at individual student level and group school level will be carried out using ML, MLM and MLMk.
This paper is organized as follows. Section 2 discusses the normal mixture model, its likelihood and its prior structure. In Section 3, we consider the multilevel mixture model and its likelihood as reflected in the paper by [7]. We illustrate our method using five artificial datasets and a real dataset in Section 4. Finally, Section 5 presents concluding remarks.

2. Estimating the Mixture Proportions by EM Algorithm

Mixture probability densities are usually used in clustering; the standard algorithm for learning clusters from the data is the EM method. Searching for and identifying clusters can be used in the classification of a new data point or for predicting missing data. A useful and popular class of models is mixture models; see [6]. Typically the EM model components are Gaussian density function and assumed to be generated by hypothetical Gaussian mixture. Because of their probabilistic nature, Gaussian mixtures are in principle preferred over models that partition a data set into discrete parts. In most applications where a new data item needs to be classified, it is more desirable to calculate the probability that this item belongs to a certain cluster than to assign it to strictly one specific cluster. We will use the parametric EM method to capture the unobserved heterogeneity of possible classes in Gaussian mixture measurement by estimating the mixture proportions in the form of categorical latent variables and estimate the first two moments in each class assuming bimodal distribution.

Let the random variable $S$ be a mixture of several $m \geq 2$ normal distributions. Hence $S_i \sim N(\mu_i, \sigma_i)$ for $i = 1, 2, \ldots, m$ then we may write

$$S = \sum_{i=1}^{m} R_i \cdot S_i,$$

where $R_i \in \{0, 1\}$ with $p(R_i = 1) = \pi_i$ such that $\sum \pi_i = 1$ and the joint distribution of the binary vector $(R_1, \ldots, R_m)$ is multinomial. If we generate multinomial variable $(R_1, \ldots, R_m)$ with $p(R_i = 1) = \pi_i$ then the density of $S$ is
\[ P(S \mid \pi, \theta) = \sum_{i=1}^{m} \pi_i P(S \mid \theta_i), \quad (2) \]

where the \( \pi_i \) is the mixing proportion and \( P(\cdot \mid \theta_i) \) is a normal density with the parameter \( \theta_i = (\mu_i, \sigma_i) \) for \( i = 1, 2, \ldots, m \). Hence \( S \) is a random variable whose distribution is given in terms of a mixture of normal distributions. Fitting a set of parameters from (2), \( \Theta = (\theta_1, \ldots, \theta_m, \pi_1, \ldots, \pi_{m-1}) \), by using the log-likelihood is known to be difficult, [4].

[3] use the mixture in Gaussian clustering with \( m \) components, where the following likelihood,

\[ L(\theta_1, \ldots, \theta_m; \pi_1, \ldots, \pi_m \mid s) = \prod_{i=1}^{m} \sum_{k=1}^{m} \pi_k f_k(s_i \mid \theta_k), \quad (3) \]

the density \( f_k \) and parameters \( \theta_k \) of the \( k \)th component in the mixture \( \pi_k \), is the probability that an observation belongs to the \( k \)th component \((\pi_k \geq 0; \sum_{k=1}^{m} \pi_k = 1)\).

The data can be viewed as consisting of \( n \) multivariate observations with \( r_i \) recoverable part of the \((s_i, z_i)\) in which \( s_i \) is observed and \( z_i \) is unobserved. The likelihood

\[ L(r_i \mid \theta) = \prod_{i=1}^{n} f(r_i \mid \theta), \quad (4) \]

is then maximized to obtain the estimate of \( \theta \). The observed data likelihood can be obtained by integrating \( z \) out of the complete data likelihood.

\[ L(s \mid \theta) = \int L(r \mid \theta) dz \quad (5) \]
The EM algorithm alternates between two steps, an ‘E’ step and an ‘M’ step, see [3]. The EM mixture model considers the complete data set \( r_i = (s_i, z_i) \), where \( z_i = (z_{i1}, \ldots, z_{im}) \) and is the unobserved portion of the data with

\[
z_{ik} = \begin{cases} 1 & \text{if } r_i \text{ belongs to group } k \\ 0 & \text{otherwise.} \end{cases}
\]

assuming that each \( z_i \) is independently and identically distributed from one draw with \( m \) categories with probabilities \((\pi_1, \ldots, \pi_m)\). In fact the coding of the allocation estimate in the E-step follows

\[
z_{ik} = \frac{\hat{\pi}_k f_k(s_i | \theta_k)}{\sum_{i=1}^{m} \hat{\pi}_i f_i(s_i | \theta_i)}
\]

and the M-step involves maximizing with respect to \( \theta \),

\[
\sum_{i=1}^{n} \sum_{k=1}^{m} z_{ik} \log(\hat{\pi}_k f_k(s_i | \hat{\theta}_k)).
\]

For an extensive discussion of the available implementation of the EM method for a variety of different parametric mixture models, see [6].

### 3. Multilevel Mixture with Unknown Mixing

We start by writing a simple empty multilevel model

\[
y_{ij} = \bar{\beta} + \beta_{0j} + e_{ij},
\]

where \( y_{ij} \) denotes the outcome for the \( i^{th} \) individual of level one in the \( j^{th} \) group of level two, \( \bar{\beta} \) represents the grand mean, \( \beta_{0j} \) is a random variable representing ‘between-units’ variability and \( e_{ij} \) is a random variable representing ‘within-units’ variability.
The distributions of the random variables are assumed to be

\[ \beta_{0j} \sim N(0, \sigma_{\beta}^2), \quad e_{ij} \sim N(0, \sigma_e^2), \]  

(8)

where \( \sigma_{\beta}^2 \) and \( \sigma_e^2 \) are the variances of the between items (level two) and within items (level one) effects respectively.

The model in (7) may include some covariates \( x_{ij} \) at level one

\[ y_{ij} = \bar{\beta} + \beta_{0j} + \beta_1 x_{ij} + e_{ij}, \]  

(9)

and other covariates \( \omega_j \) at level two

\[ \beta_{0j} = \lambda_{00} + \lambda_{01} \omega_j + \epsilon_j, \]  

(10)

where both \( e_{ij} \) and \( \epsilon_j \) are normal with mean zero and variances \( \sigma_e^2 \) and \( \sigma_{\epsilon}^2 \).

With a heterogeneous population and when regression of \( y_{ij} \) on \( x_{ij} \) vary across some latent class variable C with m categories, the residual \( \eta_{ij} \) in

\[ y_{ij|C_j=c} = \bar{\beta}_c + \beta_{0cj} + \beta_{1cj} x_{ij} + \eta_{ij}, \]  

(11)

and may have a covariance matrix \( \theta_c \) reflecting the heterogeneity. The coefficients may also reflect a large source of variation at level one and level two.

The probability of being in a given latent class with respect to a base class may vary as a function of a two-level multinomial logistic regression. In level one we have
\[ p(C_{ij} = c \mid x_{ij}) = \frac{\exp(a_{ij} + b_{c} x_{ij})}{\sum_{c=1}^{m} \exp(a_{v_j} + b_{c} x_{ij})} , \]  

(12)

while in level two we have the following three equations, the random intercept, random slope of equation (11) and the random intercept of equation (12):

\[ \beta_{0cj} = \lambda_{00c} + \lambda_{01c} \omega_{0j} + u_{0j} \]  

(13)

\[ \beta_{1cj} = \lambda_{10c} + \lambda_{11c} \omega_{1j} + u_{1j} \]  

(14)

\[ \alpha_{cj} = \lambda_{20c} + \lambda_{21c} \omega_{2j} + u_{2cj} \]  

(15)

Since MLM considers (12) as part of the optimization progression, sample allocations will be one element of the likelihood maximization procedure. The sample membership will not be constant and we may have different class membership estimate in MLM and in MLMk.

If the outcome is homogeneous and does not comprise distinct sub-populations, then one can use ML rather than MLM or MLMk.

4. Introducing the MLMk Method

The individual latent class generated by parametric EM algorithm presented in Section 2

\[ f(s) = \sum_{k=1}^{m} \pi_k f_k(s), \text{ the E-step } \frac{\hat{\pi}_k f_k(s_i \mid \hat{\theta}_i)}{\sum_{l=1}^{m} \hat{\pi}_l f_l(s_i \mid \hat{\theta}_i)} , \]

gives a fixed number of classes, m, probability density function, f_k, and \( \pi_k \) the probability that an observation comes from the kth mixture components (\( \pi_k \in (0,1) \) and \( \sum_{k=1}^{m} \pi_k = 1 \)). Equation (12) becomes constant and no estimation for \( C_{ij} \) is carried out; this is highlighted in
\[
p(C_{ij} = c \mid x_{ij}) = \frac{\exp(a_{cj} + b_{c}x_{ij})}{\sum_{s=1}^{K} \exp(a_{sj} + b_{s}x_{ij})} = \frac{\hat{\pi}_k f_k(s_i \mid \hat{\theta}_k)}{\sum_{s=1}^{m} \hat{\pi}_s f_s(s_i \mid \hat{\theta}_s)}.
\]

Where the right hand side was determined in the early stage in Section 2.

This provides labels or constraints to guide the algorithm towards a more appropriate data partitioning. The MLMk format for m known classes with probability \((\pi_1, \pi_2, \ldots, \pi_m)\) is:

\[
y_{ij\mid c_{ij}=1} = \beta_{01j} + \beta_{11}x_{ij} + \eta_{ij} \quad \text{with probability } \pi_1
\]
\[
y_{ij\mid c_{ij}=2} = \beta_{02j} + \beta_{12}x_{ij} + \eta_{2j} \quad \text{with probability } \pi_2
\]
\[
\vdots \quad \vdots \quad \vdots \quad \vdots 
\]
\[
y_{ij\mid c_{ij}=m} = \beta_{0mj} + \beta_{1m}x_{ij} + \eta_{mij} \quad \text{with probability } \pi_m
\]

and the format of the random intercepts for m classes is:

\[
\begin{align*}
\beta_{01j} &= \gamma_{01} + \gamma_{11}u_{ij} + \zeta_j \\
\beta_{02j} &= \gamma_{02} + \gamma_{12}u_{2j} + \zeta_j \\
\vdots \quad \vdots \quad \vdots \\
\beta_{0mj} &= \gamma_{0m} + \gamma_{1m}u_{mj} + \zeta_j.
\end{align*}
\] (16)

Five artificial data sets are generated and a test is presented in Subsection 4.1. Application of the MLMk method on overweight and obesity data is presented in Subsections 4.2 and 4.3.

### 4.1 Illustrative Example

A small example was run to illustrate how different methods of multilevel mixture models, MLM and MLMk perform. We conducted all computation in R and Mplus. In
this illustration, five independent realizations $x_1, x_2, \ldots, x_{4095}$ of the predictor are generated from a normal distribution of size $n = 4095$. Note that the size $n = 4095$ used to form unbalance data structure.

With the three random components of size (1365) each, we allow the parametric EM algorithm to estimate the following:

- the number of components $m$;
- the mixture proportion $\pi_m$; and
- the components-specific parameters $(\mu_m, \sigma_m)$.

The mean of the three components were set to (-16, 1, 2), the standard deviations were set to (5, 1.5, 4.5). The component one, two or three is chosen at random with equal probability of $(1/3)$.

<table>
<thead>
<tr>
<th>Parameters setting</th>
<th>$\mu_m$</th>
<th>$\sigma_m$</th>
<th>$\pi_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-16.00, 1.00, 2.00</td>
<td>5.00, 1.50, 4.50</td>
<td>0.33, 0.33, 0.33</td>
<td></td>
</tr>
<tr>
<td>-16.78, 1.23, 1.61</td>
<td>5.20, 1.55, 5.07</td>
<td>0.32, 0.42, 0.25</td>
<td></td>
</tr>
<tr>
<td>-15.94, 0.98, 2.37</td>
<td>5.38, 1.60, 4.59</td>
<td>0.34, 0.46, 0.20</td>
<td></td>
</tr>
<tr>
<td>-16.99, 1.30, 2.79</td>
<td>5.65, 1.72, 4.60</td>
<td>0.33, 0.43, 0.24</td>
<td></td>
</tr>
<tr>
<td>-17.63, 1.35, 2.13</td>
<td>5.76, 1.74, 4.89</td>
<td>0.33, 0.45, 0.22</td>
<td></td>
</tr>
<tr>
<td>-16.01, 0.86, 2.76</td>
<td>5.44, 1.55, 4.45</td>
<td>0.34, 0.45, 0.21</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Simulation setting and parameters estimates.

Applying the standard parametric EM algorithm given in Appendix A Part 1 (R code), we automatically obtain class membership indicator $CL_m$ and the probability and uncertainty of sample membership generated by $z_{ik}$ in equation (6). The second and third columns of Table 1 show the estimated parameters of five data sets.
These estimates agree with the setting parameters in the first row used for the simulation. The mixture proportion in column four for each set shows different size of class membership.

The above simulated five data sets were formulated in terms of identifying three profiles with meaningful interpretations. The MLMk models assume the existence of distinct latent sub-groups in the population.

Class membership indicator “CL_m” is introduced to MLMk in Part 2. Maximum likelihood estimation will be used for estimating fixed and random effect at individual and group level. We will preserve the fixed and random effects coefficients to check interpretability of MLM and MLMk.

In Appendix A Parts 2 and 3, we present the (Mplus) code for estimating MLM and MLMk models, respectively. The BIC and the individual and group level residuals variances, using MLM and MLMk, are presented in Table 2 below.

<table>
<thead>
<tr>
<th>Residual variance</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
<th>Set 4</th>
<th>Set 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLM</td>
<td>MLMk</td>
<td>MLM</td>
<td>MLMk</td>
<td>MLM</td>
</tr>
<tr>
<td>Class 1</td>
<td>3.49</td>
<td>4.04</td>
<td>4.54</td>
<td>4.18</td>
<td>3.80</td>
</tr>
<tr>
<td>Class 2</td>
<td>1.82</td>
<td>1.47</td>
<td>2.10</td>
<td>1.81</td>
<td>2.01</td>
</tr>
<tr>
<td>Class 3</td>
<td>22.70</td>
<td>21.12</td>
<td>23.26</td>
<td>24.37</td>
<td>23.11</td>
</tr>
<tr>
<td>Individual</td>
<td>28.01</td>
<td>26.63</td>
<td>29.90</td>
<td>30.36</td>
<td>28.92</td>
</tr>
<tr>
<td>Group</td>
<td>76.16</td>
<td>22.80</td>
<td>77.35</td>
<td>19.2</td>
<td>96.07</td>
</tr>
<tr>
<td>BIC</td>
<td>21574</td>
<td>27133</td>
<td>21582</td>
<td>26600</td>
<td>21825</td>
</tr>
</tbody>
</table>

*Table 2: BIC and Level 1 and level 2 residual variances of five MLM and MLMk fit.*

The MLM method shows huge group residual variance compared to MLMk. The two methods use different class memberships; MLMk uses the constraints provided by parametric EM presented in Table 1, unlike MLM. Note that in Table 2 “Individual” residual variances represent the sum of Class 1, Class 2, and Class 3. The BIC indicates better fit of MLM.

Figure 1 left, below, presents MLM individual (Res1, Res2, Res3) and group (ResG) residual variances. The same terms are used for the MLMk model on the right of this figure.
The five simulated data sets comprise individual and group covariates. When running ML models, the individual and the group covariates show strong association with the outcome. The variance components and the BIC for both MLM and MLMk will be compared. The MLM results indicate that when accounting for group covariate, there are no distinct group mixture components. This provides the motivation for MLMk models.

<table>
<thead>
<tr>
<th>Parameter estimate</th>
<th>ML</th>
<th>MLM</th>
<th>MLMk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual level</td>
<td>-0.932 (0.053)</td>
<td>-2.276 (0.499)</td>
<td>-2.821 (0.057)</td>
</tr>
<tr>
<td>Latent 1</td>
<td></td>
<td>0.133 (0.155)</td>
<td>0.456 (0.192)</td>
</tr>
<tr>
<td>Latent 2</td>
<td></td>
<td>0.999 (1.175)</td>
<td>-0.032 (0.035)</td>
</tr>
<tr>
<td>Latent 3</td>
<td></td>
<td>-0.859 (0.386)</td>
<td>-0.618 (0.258)</td>
</tr>
</tbody>
</table>

| Group level        |          | -0.313 (0.473) | -0.618 (0.258) |
| Latent 1           |          | 0.084 (0.128)  | -0.055 (0.366) |
| Latent 2           |          | 0.006 (0.087)  | 0.025 (0.422)  |

Table 3: ML, MLM and MLMk parameter estimate for one simulated data set.

Estimation of individual and group level parameters for one simulated data set are presented in Table 3. The estimate of individual parts shows two significant effects for MLMk and only one significant effect for MLM. At group level there are no significant effects estimated by MLM; this result contradicts the simulation setting for the parameter estimate and the estimate from the ML model. However, MLMk shows significant effects of the first latent at the group level.
4.2 Overweight and Obesity Data

This study uses routine individual and school level data collected in 2006/7 across primary schools in Birmingham, a large multicultural industrial city in central England. All schools are invited to participate in the annual National Child Measurement Programme. All children in reception year (age 4-5) and year 6 (age 10-11) are eligible to be measured and parental consent is obtained on an ‘opt out’ basis (i.e. parental consent is implied unless parents actively refuse).

BMI was calculated using weight/height (kg/cm\(^2\)), and children were defined as obese, overweight, not overweight or obese using the UK 1990 BMI reference curves for children, [1]. The definitions used were the 85\(^{th}\) centile and above and the 95\(^{th}\) centile and above for overweight and obese, respectively. This is in line with most UK based epidemiological studies on obesity in children. All schools nationally are invited to participate in an annual survey, undertaken as part of the national Physical Education, School Sport and Club Links strategy (http://www.teachernet.gov.uk/teachingandlearning/subjects/pe/nationalstrategy/).

Data extracted from schools used in this study included: number of minutes per week spent in physical education, averaged across year groups; proportion of pupils spending at least two hours in high quality physical education; and proportion of pupils who participated in a club (Club). Home postcodes were linked to UK Indices of Multiple Deprivation (IMD) scores: http://www.communities.gov.uk/communities/neighbourhoodrenewal/deprivation/deprivation07/ to provide a measure of deprivation. IMD scores are derived from social, housing and economic indicators, and are assigned to lower super output areas (LSOAs). Birmingham has 641 LSOAs.

Two questions are of interest: which individual level factors significantly predict overweight and obesity, and to what extent do they explain the observed level 1 variance; and do any of the measured school level physical activity variables explain the observed level 2 variance, and to what extent?
Year 6 data will be used to illustrate our method. Student level variable, IMD and school level variable and proportion of pupils who participated in a club (Club) will be considered. Here, a single covariate is used for simplicity of illustration, but further covariates can clearly be added. A total of 5566 students in 147 schools make individual and group level sample sizes.

### 4.3 ML, MLM and MLMk Models for Overweight Data

We fit a sequence of normal mixture models with increasing numbers of components using the parametric EM algorithm to BMI measurements for students in Year 6. A best model can be estimated by fitting with different parameterization and/or a different number of classes and then applying a statistical criterion for model selection. Table 4 gives the estimates and the Bayesian Information Criterion ‘BIC’, [8]. Note that the estimated means differ by four BMI units and the BIC is (-29072) for the two-component model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>m=2</th>
<th>m=3</th>
<th>m=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$</td>
<td>17.27</td>
<td>16.55</td>
<td>15.81</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>21.81</td>
<td>19.56</td>
<td>17.73</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>23.59</td>
<td>19.56</td>
<td>20.30</td>
</tr>
<tr>
<td>$\mu_4$</td>
<td></td>
<td>23.59</td>
<td>40.52</td>
</tr>
<tr>
<td>$\sigma_1^2$</td>
<td>3.12</td>
<td>1.95</td>
<td>1.29</td>
</tr>
<tr>
<td>$\sigma_2^2$</td>
<td>15.07</td>
<td>4.45</td>
<td>1.34</td>
</tr>
<tr>
<td>$\sigma_3^2$</td>
<td>17.85</td>
<td>4.51</td>
<td></td>
</tr>
<tr>
<td>$\pi_1$</td>
<td>0.57</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td>$\pi_2$</td>
<td>0.43</td>
<td>0.41</td>
<td>0.24</td>
</tr>
<tr>
<td>$\pi_3$</td>
<td>0.20</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>$\pi_4$</td>
<td></td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>-29072</td>
<td>-28922</td>
<td>-28934</td>
</tr>
</tbody>
</table>

*Table 4: EM estimate of BMI for 5566 students in 147 schools.*

For m = 3, the mean estimate differs by 3 and 4 with BIC reduced by (150). When m = 4, the mean estimate differs by (1.9, 2.6, 3.2) and BIC increased by (12). The three latent classes are distinguished by the level of the first and second moments.
Monotone increase in the variance (1.95, 4.45, 17.85), and maximum BIC is attained for three latent classes. The plot of the BIC in Figure 2 indicates a best fit with $m = 3$ classes.

![Figure 2: BIC for BMI; E curve stands for equal variance and V for unequal variance.](image)

The parameters estimated by the EM method indicate heterogeneity of the data, which seems to consist of three different types of BMI. This is, obviously, related to a high volume of BMI in the interval (11.89 to 41.52) for class three ‘overweight’, ‘moderate’ for class two and ‘normal’ for class one.

The box plot in Figure 3 shows the parameter estimate of the three mixtures. The three latent classes are ordered from low to high BMI: 16.55 (class 1, 39%), 19.56 (class 2, 41%) and 23.59 (class 3, 20%). The mean differences correspond to (3 to 4) standard deviations of the BMI. Outliers show in class 3.
Figure 3 Box plot of the three class parameter estimate

The density estimate \( z_{ik} \) in equation (6) for membership in component \( k \) is presented in Figure 4; this density allows complete data representation.

Figure 4 BMI densities of three latent classes.

Heavy tail density shows for class 3 ‘overweight’. This may lead to high residual variance. The class indicator \( CL_m \) generated by parametric EM will be used for the MLMk model next.

The overweight and obesity study raises two main questions: which individual level predictors predict being obesity; and does any of the measured school level predictors explain being obesity?
This study focuses on nested sources of variability: students nested within schools. Group and individual level predictors will be considered.

In the initial ML model we wish to fit

\[
y_i \sim N(\alpha_{ij} + \beta x_i, \sigma_y^2), \quad \text{for } i = 1, \ldots, 5566
\]

\[
\alpha_j \sim N(\gamma_0 + \gamma_i u_j, \sigma_{\alpha}^2), \quad \text{for } j = 1, \ldots, 147
\]

(17)

where \((\sigma_y^2 \text{ and } \sigma_{\alpha}^2)\) are individual and group residual variances and the individual and group covariates are \((x_i \text{ and } u_j)\) respectively.

We start by fitting the ML model in (17); the estimate for the overweight data presented in Table 5 below:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimate (SD)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student Level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>18.9393 (0.160)</td>
<td>0.0000</td>
</tr>
<tr>
<td>IMD score</td>
<td>0.0097 (0.003)</td>
<td>0.0031</td>
</tr>
<tr>
<td>Residual Variance</td>
<td>13.1047 (0.436)</td>
<td>0.0000</td>
</tr>
<tr>
<td><strong>School Level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Club</td>
<td>-0.0075 (0.0118)</td>
<td>0.5420</td>
</tr>
<tr>
<td>Residual Variance</td>
<td>0.1757 (0.0590)</td>
<td>0.0042</td>
</tr>
<tr>
<td><strong>BIC</strong></td>
<td>30236</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: ML estimate of BMI using model 17.

The ML model recognizes that the student level covariate IMD has a positive significant effect on all students’ BMI. This suggests that a 100 unit difference in IMD causes almost a unit increase in student BMI. For group level covariates, Club is in the right direction but not significant. Interclass correlation of the two variances is \(0.1757 / (0.1757 + 13.1047) = 0.01\). This indicates that the variance accounted for by school level is 1%.

An examination of the fitted model by scatter plot of residual versus predicted values is presented in Figure 5.
Figure 5: scatter plot of the residual versus the predicted values.

The results indicate positive departure from the within-group error assumptions that the residuals are not symmetrically distributed around zero and the variance is not constant. We are not satisfied with the fit using the ML model.

Next we will fit two models: the MLM from [7]

\[
\begin{align*}
    y_{ij|c=c} &= \beta_{0c} + \beta_{1c} x_{ij} + \eta_i \\
    p(C_j = c | x_{ij}) &= \frac{\exp(a_{ij} + b_s x_{ij})}{\sum_{s=1}^k \exp(a_{ij} + b_s x_{ij})}, \\
    \beta_{0c} &= \gamma_{00c} + \gamma_{01c} u_{0j} + \zeta_{0j} \\
    a_{ij} &= \gamma_{10c} + \gamma_{11c} u_{1j} + \zeta_{1j}
\end{align*}
\]

(18)

and our proposed MLMk model
\begin{align*}
y_{ij|c_1=1} &= \beta_{01j} + \beta_{11}x_{ij} + \eta_{1ij} \quad \text{with probability } \pi_1 = 0.39 \\
y_{ij|c_2=2} &= \beta_{02j} + \beta_{12}x_{ij} + \eta_{2ij} \quad \text{with probability } \pi_2 = 0.41 \\
y_{ij|c_3=3} &= \beta_{03j} + \beta_{13}x_{ij} + \eta_{3ij} \quad \text{with probability } \pi_3 = 0.20 \\
\beta_{01j} &= \gamma_{01} + \gamma_{11}u_{1j} + \zeta_j \\
\beta_{02j} &= \gamma_{02} + \gamma_{12}u_{2j} + \zeta_j \\
\beta_{03j} &= \gamma_{03} + \gamma_{13}u_{3j} + \zeta_j.
\end{align*}

Table 6 shows the estimate of the two models. We have different membership probability. For MLMk we use the estimate \((\pi_1, \pi_2, \pi_3)\) from parametric EM in Table 4. When using the MLM model, the probability for most likely latent class membership is \((0.778, 0.725\) and \(0.812)\), this shows clear shift in estimating individual class membership.

The conventional ML model makes the researcher believe that the individual variable (IMD) has significant effect on all students’ BMI. This is not true for the first latent class (normal level of BMI). The two models MLM and MLMk dismiss this finding.

The key result is that school level covariate ‘Club’ comes out negative and significant in latent class 2 (moderate level of BMI) and class 3 (Overweight) with the MLMk model only. In other words, our research using MLMk would conclude that increasing school percentage of club participation causes a reduction in BMI for two distinct classes “Moderate” and “Overweight”, but not for “Normal”.

This means that there is a variation in the response between the latent classes. For the first latent class ‘Normal’, no effect of Club is seen on student BMI; the magnitude of significant effects for class 2 to 3 is 16:32. The effects double for overweight students.
Table 6: estimate of MLM and MLMk models in 18 and 19.

It is interesting to note that the IMD and Club influence on BMI is different in the three latent classes. The latent classes are distinguished not only by the level of the BMI in the first part of our analysis, but also by the strength of the relation with IMD, (0.001, 0.006, 0.010) and Club, (0.002, -0.016, -0.032) in table 6.

In the first stage of our method, model based clustering can identify distinct classes of BMI and in the second stage MLMk can discover associations effects of covariates on each BMI latent class.

The individual residual variances will be used for comparison of fit of MLM and MLMk models.
Figure 6: Individual and school residual variances in 3 latent classes for MLM and MLMk.

The Figure 6 residual variances estimate says that the MLM is double MLMk in magnitude. The residual variance in latent class three is more than double that in latent class 2 for MLM, and is four times greater in MLMk; this is made clearer in the parameter estimate from parametric EM, see Figure 3. The level one variation is much larger than the level two variations, indicating greater unobserved heterogeneity on the individual student level. This analysis shows that the conventional ML model and the MLM are outperformed by the MLMk model.

5. Concluding Remarks

This paper presents a constraint-based method for allowing information regarding mixing proportions to be used in multilevel mixture models. Our approach provides more flexibility than the standard multilevel mixture; this flexibility can offer smaller variance at individual and group levels and provide medically meaningful BMI profile in group and individual coefficient space.

For the MLMk method to achieve its full potential objectives, the parametric EM method was used to guide the MLM algorithm towards the appropriate outcome partitioning. The MLM method requires the user to choose the number of mixing components, unlike MLMk, where the number of mixing is based on the BIC of the
outcome from EM in the first stage. The disadvantage of MLMk is that it produces a slightly higher BIC.

The MLMk method clearly shows that the association between IMD score and club with BMI varies between three medical latent classes. This feature is not accommodated in the ML model. The MLMk model, which relies on known mixtures, is able to detect significant differences with clubs which are not detected by MLM using unknown mixtures.

References
Appendix A

R code for generating multilevel mixture data

```r
J<-90
n<-J*(J+1)/2
group<-rep(1:J,1:J)
mu.a<-5
sigma.a<-2
a<-rnorm(J,mu.a,sigma.a)
b<-3
x<-rnorm(n,2,1)
sigma.y1<-3
sigma.y2<-4
sigma.y3<-0.5
y1<-rnorm(n/3,-2*a[group]+b*x,sigma.y1)
y2<-rnorm(n/3,a[group]+(b/2)*x,sigma.y2)
y3<-rnorm(n/3,0.5*a[group]+(b/4)*x,sigma.y3)
y<-matrix(c(y1,y2,y3), ncol=1)
library(mclust)
CL1<-Mclust(y)
table(CL1$class)
u<-runif(J,0,9)
u.full<-u[group]
y123.dat<-cbind(y,x,u.full,group)

library(arm)
E2<-lmer(y~x+u.full+(1|group),data=data.frame(y123.dat))
A<-cbind(y123.dat, LET=CL1$class)
write.csv(A, "F:\A.csv")
```

MPLUS code for MLM model estimation

```plaintext
DATA:  FILE IS D:\A.csv;
  VARIABLE: NAMES ARE y1 x1 w1 clus C;
  USEVARIABLES ARE y1 x1 w1;
  CLASS = C (3);
  WITHIN = x1 ;
  BETWEEN = w1;
  CLUSTER = clus;
  ANALYSIS: TYPE = TWOLEVEL MIXTURE;
  START = 0;
  MODEL:
    %WITHIN%
    %OVERALL%
    y1 ON x1 ;
    %c#1%
    y1 ON x1 ;
    y1;
```
MPLUS code for MLMk model estimation
DATA:  FILE IS D:\A.csv;
VARIABLE:NAMES ARE y1 x1 w1 clus g;
USEVARIABLES ARE y1 x1 w1;
CLASSES = cg (3);
KNOWNCLASS = cg (g = 1 g = 2 g = 3);
WITHIN = x1;
BETWEEN = w1;
CLUSTER = clus;
ANALYSIS:  TYPE = TWOLEVEL MIXTURE;
  START = 0;
MODEL:
  %WITHIN%
  %OVERALL%
  y1  ON x1;
  %cg#1%
  y1  ON x1;
  y1;
  %cg#2%
  y1  ON x1;
  y1;
  %cg#3%
  y1  ON x1;
  y1;
%BETWEEN%
%OVERALL%
 y1  ON w1;
%cg#1%
 y1  ON w1;
%cg#2%
 y1  ON w1;
%cg#3%
 y1  ON w1;