The Evaluation of a Concomitant Variable Behaviour in a Mixture of Regression Models

Kristýna Vaňkátová[1] | *Palacký University Olomouc, Czech Republic*
Eva Fišerová | *Palacký University Olomouc, Czech Republic*

Abstract

Finite mixture of regression models are a popular technique for modelling the unobserved heterogeneity that occurs in the population. This method acquires parameters estimates by modelling a mixture conditional distribution of the response given explanatory variables. Since this optimization problem appears to be too computationally demanding, the expectation-maximization (EM) algorithm, an iterative algorithm for computing maximum likelihood estimates from incomplete data, is used in practice. In order to specify different components with higher accuracy and to improve regression parameter estimates and predictions the use of concomitant variables has been proposed. Based on a simulation study, performance and obvious advantages of concomitant variables are presented. A practical choice of appropriate concomitant variable and the effect of predictors' domains on the estimation are discussed as well.[2]

Keywords	***JEL code***
Mixture of regression models, linear regression, EM algorithm, concomitant variable	*C11, C38, C51, C52*

INTRODUCTION

The basic requirement for the proper use of a standard linear regression model is a homogeneity in the studied population. If this assumption is violated and a standard regression model is inapplicable due to several heterogeneous groups in data, an alternative approach to modelling by means of mixture of regression models can be utilized (DeSarbo and Cron, 1988; McLachlan and Peel, 2000). While a standard regression mainly aims to estimate regression parameters, a mixture of regression models is also used as a tool for data clustering and therefore works as a clusterwise regression.

Mixtures of linear regression models, originally called switching regressions, are a special case of mixture density models (also known as a mixture of distributions) that were initially studied by means of a moment-generating function (Pearson, 1894). Recently, however, a likelihood point of view has been preferred for mixture models with a fixed number of components. A standard technique to obtain the maximum likelihood estimates is the expectation-maximization (EM) algorithm (Dempster et al., 1977).

[1] 17. listopadu 12, 771 46 Olomouc, Czech Republic. E-mail: kristyna.vankatova@upol.cz, phone: (+420)733348062.

[2] This article is based on contribution at the conference *Robust 2016*.

In addition to the method of moments and the maximum-likelihood approach, a variety of other methods have been proposed for estimating parameters in mixture densities. These methods include graphic procedures; an estimate determined by a least squares criterion in the spirit of the minimum-distance method; a procedure based on a linear operator reducing the variances of the component densities; the confusion matrix method and related methods; a stochastic approximation algorithm; and a minimum chi-square estimation. A short description of these and related methods can be found in Redner and Walker (1984) along with necessary references.

Modelling of unobserved heterogeneity using a maximum likelihood methodology is presented for instance in Bengalia et al. (2009), De Veaux (1989), DeSarbo and Cron (1988), and Faria and Soromenho (2010). An extensive review of finite mixture models can be found in McLachlan and Peel (2000). The methodology of mixtures of regression models can be applied in various research fields, such as climatology, biology, economics, medicine and genetics; see e.g. Grün et al. (2012), Vaňkátová and Fišerová (2016), and Hamel et al. (2016).

Grün and Leisch (2008) proposed the concomitant variable models for the component weights that allow to allocate the data into the mixture components through other variables called concomitant. This extension can provide both more precise parameter estimates and better components identification. Since the concomitant variable is still a new concept in mixture modelling, the aim of this paper is to evaluate its role. Accordingly, a simulation study was conducted and results concerning the impact of the concomitant variable on the model quality are presented. Both precision of regression parameters and clusterwise properties of the model are addressed in cases of categorical and continuous concomitant variables. A practical choice of appropriate concomitant variable and the effect of predictors' domains on the estimation are discussed as well.

This paper is structured as follows. In Section 1, some fundamentals of mixtures of linear regression models with and without concomitant variables are presented. The theory behind parameters estimation is summarized in Section 2. Section 3 is dedicated to a simulation study investigating the performance of mixture models with and without concomitant variables. At the end, the conclusions of the study are drawn and additional comments are given.

1 REGRESSION MODELS

1.1 Mixtures of regression models

A mixture distribution (Pearson, 1894) is the probability distribution of a random variable obtained from a set of other random variables in such a way that, firstly, a random variable from the set is drawn according to given probabilities that sum to one; and that, secondly, the value of the selected variable is realized. Formally, the probability density function f can be represented by a convex combination of probability density functions f_i:

$$f(y) = \sum_{i=1}^{c} \pi_i f_i(y)\,, \tag{1}$$

where f_i are called component densities and $\pi_1, \dots, \pi_c$ are positive mixing proportions that sum to one. A Gaussian mixture distribution assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters (McLachlan and Peel, 2000).

Introduced by Goldfeld and Quandt (1976) as switching regressions, the mixture of regression models is formed analogously to the mixture distribution (1). Let Y_j denote the response variable, and let $\boldsymbol{x}_j$ denote the vector of predictors for the jth subject. Assuming the random errors are normally distributed and subpopulations are present within an overall population, the response variable Y_j is given as a finite sum (mixture) of conditional univariate normal densities $\boldsymbol{\varphi}$ with the expectation $\boldsymbol{x}_j^T \boldsymbol{\beta}_i$, and the variance

$\sigma_i^2, i = 1, \dots, c$. Following the mixture models structure, the conditional density of $Y_j \mid \boldsymbol{x}_j$ is defined by (Bengalia et al., 2009):

$$f(y_j|\boldsymbol{x}_j, \boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i \varphi(y_j|\boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2) = \sum_{i=1}^{c} \pi_i (2\pi\sigma_i^2)^{-\frac{1}{2}} \mathrm{e}^{\frac{-(y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_i)^2}{2\sigma_i^2}}. \tag{2}$$

The symbol $\boldsymbol{\Psi}$ denotes the vector of all unknown parameters for a mixture of regression models with c components:

$$\boldsymbol{\Psi} = \left(\pi_1, \dots, \pi_c, (\boldsymbol{\beta}_1^T, \sigma_1^2), \dots, (\boldsymbol{\beta}_c^T, \sigma_c^2)\right)^T,$$

where $\boldsymbol{\beta}_i$ denotes the q-dimensional vector of unknown regression parameters for the ith component and σ_i^2 is the unknown error variance for the ith component. The mixing proportions $\pi_1, \dots, \pi_c$ satisfy the conditions $\pi_i > 0$ and $\sum_{i=1}^{c} \pi_i = 1$. A more transparent way of expressing a mixture of regression models is:

$$Y_j = \begin{cases} \boldsymbol{x}_j^T \boldsymbol{\beta}_1 + \varepsilon_{1j} \text{ with probability } \pi_1, \\ \boldsymbol{x}_j^T \boldsymbol{\beta}_2 + \varepsilon_{2j} \text{ with probability } \pi_2, \\ \vdots \\ \boldsymbol{x}_j^T \boldsymbol{\beta}_c + \varepsilon_{cj} \text{ with probability } \pi_c, \end{cases} \tag{3}$$

where ε_{ij} are independent random errors with a normal distribution $\mathrm{N}(0, \sigma_i^2)$, $i = 1, \dots, c$, $j = 1, \dots, n$.

1.2 Mixtures of regression models with concomitant variables

The mixture of regression models consists of c components where each component follows a specific parametric distribution. Each individual component has been assigned a weight indicating the prior probability for an observation to come from this component. Hence, the mixture distribution is given by the weighted sum over c components with weights corresponding to the prior probabilities. If the weights depend on further variables, the latter are referred to as concomitant variables. The mixture of regression models with concomitant variables was introduced and is described in detail by Grün and Leisch (2008).

The mixture of regression models with s concomitant variables is in the form of:

$$f(y_j|\boldsymbol{x}_j, \boldsymbol{\omega}_j) = \sum_{i=1}^{c} \pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) \varphi(y_j|\boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2), \tag{4}$$

where $\boldsymbol{\omega}_j$ denotes the s-dimensional vector of concomitant variables for the jth observation. The symbol $\boldsymbol{\alpha}_i$ denotes the vector of parameters of the concomitant variable model for the ith component. The dimension of $\boldsymbol{\alpha}_i$ relates to the chosen concomitant model and the dimension of concomitant variables. Dimensions of these vectors remain the same over all observations.

The set of unknown parameters for a mixture of regression models with concomitant variables with c components is:

$$\boldsymbol{\Psi} = \left((\boldsymbol{\alpha}_1^T, \boldsymbol{\beta}_1^T, \sigma_1^2), \dots, (\boldsymbol{\alpha}_c^T, \boldsymbol{\beta}_c^T, \sigma_c^2)\right)^T.$$

The component weights π_i need to satisfy conditions:

$$\sum_{i=1}^{c} \pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) = 1, \qquad \text{for } j = 1, \dots, n, \tag{5}$$

$$\pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) > 0, \qquad \text{for } i = 1, \dots, c, \text{and } j = 1, \dots, n.$$

Although the function of a concomitant variable model may have an arbitrary form, it has to fulfill the conditions (5). In this paper, a multinomial logit model for the π_i is considered, as seen below:

$$\pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) = \frac{\mathrm{e}^{\omega_j^T \alpha_i}}{\sum_{h=1}^{c} \mathrm{e}^{\omega_j^T \alpha_h}} \quad \text{for} \quad i = 1, \dots, c,\ j = 1, \dots, n, \tag{6}$$

with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_c^T)^T$ and $\boldsymbol{\alpha}_1 \equiv \mathbf{0}$. This settings means that the first component is a baseline. The vector $\boldsymbol{\alpha}_i$ is s-dimensional, $i = 1, \dots, c$, provided the model contains s concomitant variables (Grün and Leisch, 2008).

A classical linear regression model can be applied to heterogeneous population problem only in a case when the component membership of every observation is deterministically known or described by the observable random variable. As the result of the first option (deterministically determined membership), c independent regression models are analyzed separately. Concerning the latter scenario, the cluster identification information in a form of a categorical random variable is included in the model as dummy variables (indicators of categories) together with interactions between predictors. However, both suggested approaches are inapplicable in the situation discussed in this paper since the cluster membership of observations is considered to be latent.

The categorical concomitant variable can be potentially used in a classical linear regression model as a random variable carrying the information about a cluster membership but the effect of such a variable on the estimated model is highly exaggerated. Also, a number of other problems arise in this case. For example, there is a problem with a number of categories versus a number of components. In addition, it is not ideal that the assignment of an observation to the cluster is no longer weighted but fixed as 1 or 0.

Mixtures of regression models are frequently used specifically for theirs clustering properties. Unlike classical clustering methods, mixture regression models are able to deal with clustering of the data following a certain function, therefore we refer to the clusterwise regression method.

2 PARAMETERS ESTIMATION

In order to obtain parameters estimates for a standard mixture of regression models with a fixed number of components c, the log-likelihood function is maximized:

$$\log L(\boldsymbol{\Psi}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_n, y_1, \dots, y_n) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{c} \pi_i \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right)\right). \tag{7}$$

Within the framework of mixture models the observations are viewed as an incomplete data. The data consists of triples $\left(\boldsymbol{x}_j^T, y_j, \boldsymbol{z}_j^T\right)^T$, where $\boldsymbol{z}_j$ is an unobserved vector indicating from which mixture component the observation $\left(\boldsymbol{x}_j^T, y_j\right)^T$ is drawn. More precisely, z_{ij} is equal to one if the observation $\left(\boldsymbol{x}_j^T, y_j\right)^T$ comes from the ith component; otherwise z_{ij} is zero. These values z_{ij} are unobservable and therefore treated as missing, and the data are augmented by estimates of the component memberships, i.e. the estimated posterior probabilities τ_{ij} (McLachlan and Peel, 2000). Using the Bayes rule, any jth observation can be assigned to the ith cluster with a probability given by:

$$\tau_{ij} = \frac{\pi_i \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right)}{\sum_{h=1}^{c} \pi_h \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_h, \sigma_h^2\right)}\,. \tag{8}$$

Since mixing proportions sum to unity, the log-likelihood function can be optimized using the Lagrange multipliers method with $\sum_{i=1}^{c} \pi_i = 1$ constraint. In order to obtain stationary equations, we compute the first order partial derivatives of the augmented log-likelihood function and equate them to zero. In the next step, it is a matter of few simple modifications to acquire a new system of equations obviously corresponding to stationary equations of another optimization problem formulated as (DeSarbo and Cron, 1988):

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{c} \sum_{j=1}^{n} \tau_{ij} \log \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right). \tag{9}$$

The function (9) is called the expected complete log-likelihood due to the fact it works with the estimated posterior probabilities τ_{ij} instead of unobservable values z_{ij}. This particular structure gainfully lends itself to the development of the EM algorithm (Dempster et al., 1977), an iterative procedure which alternates between an Expectation step and a Maximization step. The EM algorithm takes advantage of the expected likelihood that is in general easier to maximize than the original one.

The EM algorithm is widely exploited in practice. The estimators are viewed as some form of a local maximum likelihood estimator (Behboodian, 1970). However, it is not guaranteed that the EM algorithm provides a global maximum. A complication may occur in the case of normal mixtures with component specific variances, where the log-likelihood is unbounded and attains +∞ for certain values of the parameter space. For this specific case, the EM algorithm adds to its advantage and provides, according to many practitioners, rather reasonable solutions unlike algorithmic approaches of global character such as a gradient function based techniques. Although the EM algorithm is often used, there is surprisingly little theoretical knowledge available for this estimator. In fact, it might be unclear to which extent asymptotic properties of the EM algorithm estimators, such as consistency, asymptotic efficiency and asymptotic normality, hold (Nityasuddhi and Böhning, 2003).

In the E-step, posterior probabilities τ_{ij} are estimated. Consequently, the expected complete log-likelihood is maximized in the M-step and the vector of unknown parameters $\boldsymbol{\Psi}$ is updated. The (*k+1*)th iteration of the EM algorithm can be summarized as follows:

E-step: Given the observed data $\boldsymbol{y}$ and current parameter estimates $\widehat{\boldsymbol{\Psi}}^{(k)}$ in the kth iteration, replace the missing data z_{ij} by the estimated posterior probabilities:

$$\hat{\tau}_{ij}^{(k)} = \frac{\hat{\pi}_i^{(k)} \varphi\left(y_j \middle| \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_i^{(k)}, \hat{\sigma}_i^{2(k)}\right)}{\sum_{h=1}^{c} \hat{\pi}_h^{(k)} \varphi\left(y_j \middle| \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_h^{(k)}, \hat{\sigma}_h^{2(k)}\right)}. \tag{10}$$

M-step: Given the estimates $\hat{\tau}_{ij}^{(k)}$ for the posterior probabilities τ_{ij} (which are functions of $\widehat{\boldsymbol{\Psi}}^{(k)}$), obtain new estimates $\widehat{\boldsymbol{\Psi}}^{(k+1)}$ of the parameters by maximizing the expected complete log-likelihood:

$$Q\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}^{(k)}\right) = \sum_{i=1}^{c} \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \log \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right). \tag{11}$$

This maximization problem is equivalent to solving the weighted least squares problem, where the vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ of observations and the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ are each weighted by $\sqrt{\hat{\tau}_{ij}^{(k)}}$. That means that we get:

$$\widehat{\boldsymbol{\beta}}_i^{(k+1)} = \left(\boldsymbol{X}^T \boldsymbol{W}_i^{(k)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}_i^{(k)} \boldsymbol{y} \tag{12}$$

for estimates of regression parameters, assuming the $n \times n$ matrix $\boldsymbol{W}_i^{(k)} = Diag\left\{\sqrt{\hat{\tau}_{i1}^{(k)}}, \dots, \sqrt{\hat{\tau}_{in}^{(k)}}\right\}$ is a diagonal matrix of weights, and:

$$\hat{\sigma}_i^{2(k+1)} = \frac{\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_i^{(k+1)}\right)\boldsymbol{W}_i^{(k)}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_i^{(k+1)})}{n} \tag{13}$$

for the error variance estimate.

Thus, the entire set of $\widehat{\boldsymbol{\beta}}_i^{(k+1)}$ is derived by performing c separate weighted least squares analyses. In the same spirit, $\hat{\sigma}_i^{2(k+1)}$ is estimated and, lastly, the estimates of the mixing proportions π_i are updated using:

$$\hat{\pi}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)}}{n} \quad \text{for } i = 1, \dots, c. \tag{14}$$

The principle of parameters estimation is very similar for the mixture of regression models with concomitant variables. The expected complete log-likelihood function can be derived analogously to the previous case. Thus, the EM algorithm for the mixture models with concomitant variables follows the following two steps (Grün and Leisch, 2008):

E-step: Given the observed data $\boldsymbol{y}$ and current parameter estimates $\widehat{\boldsymbol{\Psi}}^{(k)}$ in the kth iteration, replace the missing data z_{ij} by the estimated posterior probabilities τ_{ij}:

$$\hat{\tau}_{ij}^{(k)} = \frac{\pi_i(\boldsymbol{\omega}_j, \widehat{\boldsymbol{\alpha}}_i^{(k)})\varphi\left(y_j \middle| \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_i^{(k)}, \hat{\sigma}_i^{2(k)}\right)}{\sum_{h=1}^{c} \pi_i(\boldsymbol{\omega}_j, \widehat{\boldsymbol{\alpha}}_h^{(k)})\varphi\left(y_j \middle| \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_h^{(k)}, \hat{\sigma}_h^{2(k)}\right)}. \tag{15}$$

M-step: Given the estimates $\hat{\tau}_{ij}^{(k)}$ for the posterior probabilities τ_{ij} (which are functions of $\widehat{\boldsymbol{\Psi}}^{(k)}$), obtain new estimates $\widehat{\boldsymbol{\Psi}}^{(k+1)}$ of the parameters $\boldsymbol{\Psi}$ by maximizing:

$$Q\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}^{(k)}\right) = Q_1\left(\boldsymbol{\beta}_i, \sigma_i^2, i = 1, \dots, c; \widehat{\boldsymbol{\Psi}}^{(k)}\right) + Q_2\left(\boldsymbol{\alpha}, \widehat{\boldsymbol{\Psi}}^{(k)}\right), \tag{16}$$

where:

$$Q_1\left(\boldsymbol{\beta}_i, \sigma_i^2, i = 1, \dots, c; \widehat{\boldsymbol{\Psi}}^{(k)}\right) = \sum_{i=1}^{c} \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \log\left(\varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right)\right) \tag{17}$$

and:

$$Q_2\left(\boldsymbol{\alpha}, \widehat{\boldsymbol{\Psi}}^{(k)}\right) = \sum_{i=1}^{c} \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \log\left(\pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i)\right). \tag{18}$$

Formulas Q_1 and Q_2 can be maximized separately. The formula Q_1 is maximized using the weighted ML estimation of linear models with weights $\sqrt{\hat{\tau}_{ij}^{(k)}}$. The maximization of Q_1 gives new estimates $\widehat{\boldsymbol{\beta}}_i^{(k+1)}, \hat{\sigma}_i^{2(k+1)}$, $i = 1, \dots, c$. The term Q_2 is maximized by means of the weighted ML estimation of multinomial logit models and provides new estimates $\widehat{\boldsymbol{\alpha}}_i^{(k+1)}$.

Initial values of regression parameters may be based on a random division of observations into c components, i.e. on initial $\hat{\tau}_{ij}^{(0)}$ probabilities, where for each observation y_j only one of these c probabilities equals to 1 and the other ones are set to zero. The EM algorithm is stopped when the (relative) change of the log-likelihood is smaller than a chosen tolerance.

The number of components can be chosen by comparing information criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC) of various models, each with a different number of components.

3 SIMULATION STUDY

A simulation study is conducted to assess the performance of both a standard mixture of regression models and a mixture of regression models containing concomitant variables. The standard regression model could only be applied in the case of statistically significant concomitant variables that could be used as additional explanatory variables. However, such a model lacks the clustering properties that are essential in the following analysis; therefore, only mixtures of regression models are considered.

The study is mainly focused on the impact of the concomitant variable on the model quality (the accuracy of estimation and clustering), practical choice of appropriate concomitant variable and the effect of predictors' domains on the estimation. Accordingly, data are simulated under a two and three component mixture of linear regressions and concomitant variables are considered either categorical or continuous. The statistical software R (R Core Team, 2016) containing several extension packages for the estimation of a mixture of regression models is used. The results are built on flexmix package, introduced in Leisch (2004).

3.1 Design of the study

Each observation $(\boldsymbol{x}_j^T, y_j)^T$, $j = 1, \dots, n$, is generated by the following scheme. Firstly, the component membership is determined. Assuming the observation comes from the ith component with the probability π_i, it is possible to randomly select the component membership by means of the outcome of the multinomial distribution with mixing proportions as multinomial probabilities. With established membership, the value of the predictor x for the assigned ith component is randomly generated from a given distribution (a uniform distribution on the interval $[x_L, x_U]$ or a normal distribution with parameters μ_x and σ_x^2). Next, a normal random error ε_{ij} with the mean 0 and variance σ_i^2 is generated. Finally, the observed value y_j is computed using the regression model form $\boldsymbol{x}_j^T \boldsymbol{\beta}_i + \varepsilon_{ij}$, where the true values of regression parameters $\boldsymbol{\beta}_i$ are considered. Two typical positions of the true regression lines are considered, in which the lines are either parallel or concurrent. The effect of these alternative positions is also studied.

In order to examine the performance of both mixture models (with and without a concomitant variable), the following statistical characteristics of estimators of $\boldsymbol{\Psi}$ are calculated:

- The mean square error of the regression parameter estimates over all replications:

$$MSEPAR(\hat{\psi}_p) = \frac{1}{M} \sum_{m=1}^{M} \left(\psi_p - \hat{\psi}_p^{(m)}\right)^2, \tag{19}$$

where ψ_p is the pth parameter of the vector $\boldsymbol{\Psi}$. While ψ_p is a true parameter, $\hat{\psi}_p^{(m)}$ is the final estimate of a given parameter in the mth replication, $m = 1, \dots, M$. We desire to examine MSEPAR for all mixture model parameters, i.e. for regression coefficients $\boldsymbol{\beta}_i$, error variances σ_i^2, and mixing proportions π_i, $i = 1, \dots, c$. For mixing proportions, however, the true values of component weights are not constant and vary over replications, denoted as $\pi_i^{(m)}$ (with increasing sample size, these values converge to the true mixing proportions). Hereby, the mean square error of mixing proportions is computed according to:

$$MSEPAR(\hat{\pi}_i) = \frac{1}{M} \sum_{m=1}^{M} \left(\pi_i^{(m)} - \hat{\pi}_i^{(m)}\right)^2. \tag{20}$$

- The mean variance of estimated regression parameters:

$$VAR(\hat{\psi}_p) = \frac{1}{M}\sum_{m=1}^{M} \operatorname{var}\left(\hat{\psi}_p^{(m)}\right). \tag{21}$$

  Here, $\operatorname{var}\left(\hat{\psi}_p^{(m)}\right)$ represents the estimate of a variance of the pth parameter estimator in the mth replication. The variance-covariance matrix of the regression parameters estimators is estimated by the inverted negative Hesse matrix of the full likelihood of the model (Grün and Leisch, 2008).
- The misclassification error:

$$\mathrm{Err}_{\mathrm{M}} = 1 - \frac{1}{nM}\sum_{m=1}^{M}\sum_{j=1}^{n} I(\hat{\boldsymbol{z}}_j = \boldsymbol{z}_j), \tag{22}$$

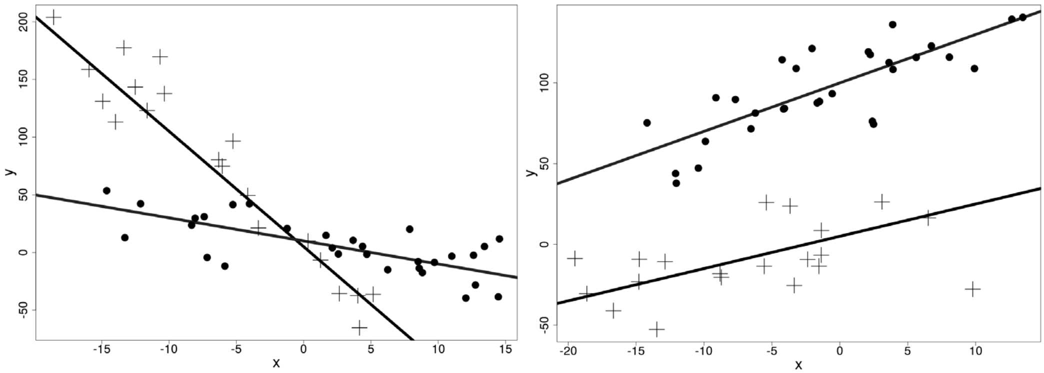
  where $\boldsymbol{z}_j$ is the true component membership of each observation and $\hat{\boldsymbol{z}}_j$ is its estimate. The misclassification error states a mean ratio of incorrectly assigned observations over all replications.

For the simplification, mixtures of regression lines are only considered in the following simulations. This simplification is not restrictive. The similar results are also valid in more complex regression models such as models with a polynomial trend.

3.2 Two component mixtures of linear regression models

For two component models, samples of three different sizes ($\mathrm{n} = 50, 100, 300$) are considered. Values of the predictor x are drawn from a uniform distribution on the interval $[-20,15]$ for both components. True parameter values (regression lines' coefficients and variances) are shown in Table 1 along with true mixing proportions. Scatter plots for samples of size 50 together with true regression lines are demonstrated in Figure 1. The number of replications is set to $M = 200$ considering how slow the algorithm is in practice.

Figure 1 Scatter plots for two configurations of mixtures of two regression lines of a sample of size 50 together with true regression lines



Source: Own construction

The concomitant variable is chosen as a categorical variable with four levels. Each of these levels labels the corresponding component with approximately 90% accuracy; values 1 and 2 label the first component, while values 3 and 4 label the second one. Since the concomitant variable is a univariate categorical variable, we can create three dummy variables that reflect the original variable in terms of a linear regression model.

Table 1 True parameter values for a two component mixture of regression lines

Position	β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	π_1	π_2
Parallel	100	3	15	5	2	20	6/10	4/10
Concurrent	10	–2	15	5	–10	20	6/10	4/10

Source: Own construction

Let us consider for example the level one as the reference category. Then, every level of the concomitant variable can be replaced with the 4-dimensional vector $\boldsymbol{\omega_j} = \left(1, \delta_{j2}, \delta_{j3}, \delta_{j4}\right)^T$, where δ_{jl} is an indicator of the level l for the jth observation, i.e. $\delta_{jl} = 1$ if the jth observation is labelled to the lth level, otherwise $\delta_{jl} = 0$. The resulting logit model is of the form:

$$\text{logit}\left[\pi_i\left(\boldsymbol{\omega_j}, \boldsymbol{\alpha_i}\right)\right] = \alpha_{i0} + \alpha_{i2}\delta_{j2} + \alpha_{i3}\delta_{j3} + \alpha_{i4}\delta_{j4}, \;\; i = 1,2, \;\; j = 1, \dots, n, \tag{23}$$

meaning that the 90% accuracy of classification by a concomitant variable corresponds to a vector of parameters $\boldsymbol{\alpha}_1 = (\alpha_{10}, \alpha_{12}, \alpha_{13}, \alpha_{14})^T = (0,0,0,0)^T$ and $\boldsymbol{\alpha}_2 = (\alpha_{20}, \alpha_{22}, \alpha_{23}, \alpha_{24})^T = (-2.2,0,4.4,4.4)^T$. The vector $\boldsymbol{\alpha}_1$ is set to zero as the theory in Section 2 determines. To demonstrate the basic scheme of a concomitant model, we aim to show the selected probabilities π_i given by the multinomial logit model (23); for the clarity, the level l is also indicated:

$$\pi_1\left(l = 1, \boldsymbol{\omega}_j = (1{,}0{,}0{,}0)^T, \boldsymbol{\alpha}_1\right) = \frac{\mathrm{e}^{\alpha_{10}}}{\mathrm{e}^{\alpha_{10}} + \mathrm{e}^{\alpha_{20}}} = 0.9\,, \tag{24}$$

$$\pi_1\left(l = 3, \boldsymbol{\omega}_j = (1{,}0{,}1{,}0)^T, \boldsymbol{\alpha}_1\right) = \frac{\mathrm{e}^{\alpha_{10}+\alpha_{13}}}{\mathrm{e}^{\alpha_{10}+\alpha_{13}} + \mathrm{e}^{\alpha_{20}+\alpha_{23}}} = 0.1\,, \tag{25}$$

$$\pi_2\left(l = 1, \boldsymbol{\omega}_j = (1{,}0{,}0{,}0)^T, \boldsymbol{\alpha}_2\right) = \frac{\mathrm{e}^{\alpha_{20}}}{\mathrm{e}^{\alpha_{10}} + \mathrm{e}^{\alpha_{20}}} = 0.9\,, \tag{26}$$

$$\pi_2\left(l = 3, \boldsymbol{\omega}_j = (1{,}0{,}1{,}0)^T, \boldsymbol{\alpha}_2\right) = \frac{\mathrm{e}^{\alpha_{20}+\alpha_{23}}}{\mathrm{e}^{\alpha_{10}+\alpha_{13}} + \mathrm{e}^{\alpha_{20}+\alpha_{23}}} = 0.1\,. \tag{27}$$

Software R provides a detailed summary for the concomitant model, so that both parameters estimates and their significance test statistics are displayed.

The effect of a concomitant variable on the estimation in mixture models is visible on the resulting statistical characteristics of estimators, such as the mean square errors (MSEPAR), the mean variances (VAR) and the misclassification errors (Err_M), see Tables 2 and 3. It is rather obvious the concomitant variable helps to optimize parameters estimates in both regression lines configurations (parallel and concurrent). Its benefit is apparent mainly for a small sample size. In case of a parallel model of a sample of size 50, the MSEPAR is about 1.7-fold to 3.2-fold smaller for a concomitant model than for a standard mixture. With an increasing sample size, the MSEPAR from both models are comparable. The accuracy of estimators is slightly higher in a model with a concomitant variable. The same tendency is also valid for the accuracy of mixing proportions (Table 3). For two component mixtures, the MSEPAR of both mixing proportions is the same. The MSEPAR of $\hat{\pi}_1$ is minor in both mixture models, and the difference is most significant for the parallel position of regression lines. It is less than 0.1%, with the exception for a sample of size 50, when the mixture model with a concomitant variable is used. For the standard mixture, the MSEPAR ($\hat{\pi}_1$) is 2-fold greater for both parallel and concurrent position, except sample size

Table 2 The mean square error (*MSEPAR*) and the mean variance (*VAR*) of the regression parameters, and standard error estimates for a two component mixture of regression models

Parallel

MSEPAR		β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2
n = 50	standard	24.1535	0.4206	10.3827	74.1701	1.1376	32.1636
	concomitant	8.7147	0.1405	4.9659	44.6211	0.3585	14.7015
n = 100	standard	7.1550	0.1165	3.5546	13.8645	0.3098	8.7427
	concomitant	4.0437	0.0606	1.8087	14.9526	0.1335	6.7200
n = 300	standard	1.7086	0.0244	0.7518	5.0777	0.0452	1.9845
	concomitant	1.3909	0.0166	0.6388	4.3677	0.0403	1.6901
VAR							
n = 50	standard	8.6111	0.1200	0.0207	33.4614	0.3381	0.0310
	concomitant	7.4635	0.1003	0.0181	25.9662	0.2717	0.0273
n = 100	standard	4.1556	0.0566	0.0098	13.4476	0.1390	0.0157
	concomitant	3.8123	0.0517	0.0091	13.6837	0.1399	0.0144
n = 300	standard	1.3314	0.0180	0.0033	4.6263	0.0464	0.0052
	concomitant	1.2862	0.0173	0.0030	4.5854	0.0454	0.0048

Concurrent

MSEPAR		β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2
n = 50	standard	12.6919	0.1459	5.9812	56.5914	0.5026	18.6374
	concomitant	9.1072	0.1245	5.2192	45.6830	0.3635	14.0798
n = 100	standard	4.6781	0.0528	2.8718	28.0681	0.2231	9.3213
	concomitant	4.5898	0.0609	2.4462	16.5644	0.1752	5.8688
n = 300	standard	1.4862	0.0210	0.8740	7.1206	0.0641	2.5656
	concomitant	1.2614	0.0155	0.7271	5.9025	0.0569	2.4797
VAR							
n = 50	standard	8.9620	0.1091	0.0258	39.0396	0.3477	0.0371
	concomitant	8.4352	0.1076	0.0206	31.1324	0.3036	0.0303
n = 100	standard	4.5738	0.0542	0.0125	20.3584	0.1774	0.0189
	concomitant	3.9854	0.0506	0.0103	16.3407	0.1526	0.0159
n = 300	standard	1.5218	0.0181	0.0041	6.6889	0.0582	0.0062
	concomitant	1.3803	0.0177	0.0034	5.2361	0.0489	0.0054

Source: Own construction

of 50. For the smallest sample size in this study, the difference in MSEPAR of $\hat{\pi}_1$ is more significant considering parallel configuration of regression lines. In this case the model with a concomitant variable achieves more than 9-fold better results in π_i estimation.

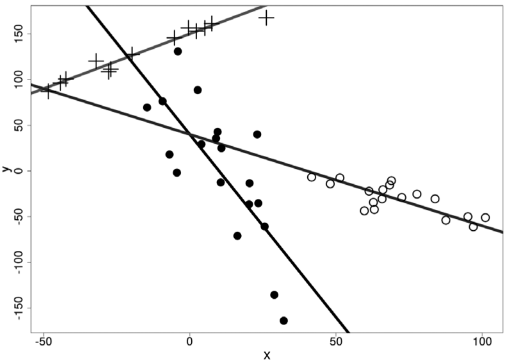
In addition, the misclassification error is considerably smaller when additional information on membership of observations is taken into account (Table 3). The misclassification error also depends on the configuration of regression lines in the mixture. For the mixture of concurrent lines, the misclassification error is more than 10% when the standard mixture model is used, while it decreases by half using the concomitant model. The classification is better for parallel regression lines. Although the misclassification error is still worse in the standard mixture (2% in contrast to 0.6% for a sample of size 50), with an increasing sample size the differences become negligible (0.6% and 0.3% for a sample of size 300).

Table 3 The misclassification errors ($\mathrm{Err_M}$) and the mean square errors for the estimate of the first mixing proportion ($\mathrm{MSEPAR}(\hat{\pi}_1)$) for a two component mixture of regression lines

	n = 50		n = 100		n = 300	
Position	standard	concomitant	standard	concomitant	standard	concomitant
$\mathrm{Err_M}$						
Parallel	0.0209	0.0061	0.0080	0.0038	0.0057	0.0034
Concurrent	0.1112	0.0502	0.1083	0.0437	0.1040	0.0416
$\mathrm{MSEPAR}(\hat{\pi}_1)$						
Parallel	0.0019	0.0002	0.0002	< 0.0001	< 0.0001	< 0.0001
Concurrent	0.0028	0.0017	0.0014	0.0007	0.0004	0.0002

Source: Own construction

Figure 2 The scatter plot of a three component mixture of a sample of size 50. True regression lines visualized

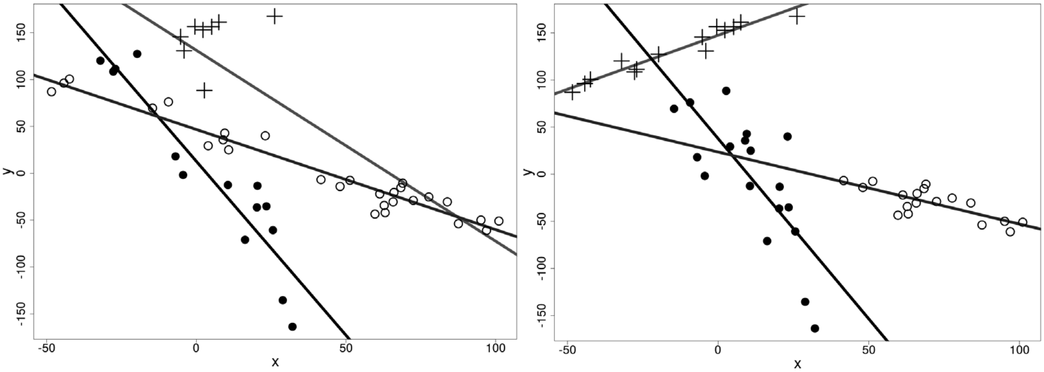


Source: Own construction

It should be noted that similarly to other clustering problems it is not guaranteed in general that the misclassification error converges to zero as *n* tends to infinity. It rather converges to some fixed value depending on the variance of parameter estimators and the distance or the angle between regression lines.

It is worth mentioning that even a random choice of a concomitant variable (a concomitant variable is generated as a completely random variable with zero correlation to the observation's component membership) does not affect this kind of mixture models in a negative way. It merely causes that the results given by a model including the concomitant variable are comparable to the results of a standard mixture model. This is due to the fact that a multinomial logit model describing

Figure 3 Fitted regression lines via a standard mixture of regression lines (left) and a mixture model with a continuous concomitant variable (right)



Source: Own construction

the effect of a concomitant variable on mixing proportions is only secondary in a process of clustering. A possible contribution of concomitant variables can be assessed by statistical significance of parameters in a multinomial logit model (6).

The mixture model tends to maintain its behaviour no matter how many categories the concomitant variable has. Favourable characteristics of mixture models containing concomitant variables are preserved even when the concomitant variable is continuous. The superiority of the mixture of regression models using concomitant variables does not deteriorate with a rising number of components.

3.3 Three component mixtures of linear regression models

In this section, a three component mixture and a continuous concomitant variable represented by the normally distributed predictor itself are investigated. The aim is to show how problematic the usage of mixture models is when components are defined on different parts of the predictor space, which is in our case the x-axis. In other words, if the values of the predictor x are generated from a uniform distribution, the interval $[x_L, x_U]$ is not the same for all components. Assuming normally distributed predictor, the mean μ_x is different for each component. Even relatively small nuances significantly affect estimates in a negative way, as it can be seen in the following example. In this type of a configuration of mixtures of regression functions, the estimation can be improved by using the predictor as a concomitant variable.

Table 4 True parameter values for a three component mixture of regression lines and probability distributions of a predictor

reg. parameters	β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	β_{30}	β_{31}	σ_3^2
	150	1.2	10	40	–4	40	40	–1	10
mixing proportions π_i	3/10			5/10			2/10		
$x \sim \mathrm{N}(\mu_x, \sigma_x^2)$	$\mathrm{N}(-20, 20^2)$			$\mathrm{N}(10, 20^2)$			$\mathrm{N}(70, 20^2)$		

Source: Own construction

The design of the mixture model containing three regression lines is presented in Table 4. The predictor x is considered as a concomitant variable ω and the logit of the mixing proportion π_i is assumed to be a linear function of a concomitant variable, as seen below:

$$\mathrm{logit}\big[\pi_i(\omega_j, \boldsymbol{\alpha}_i)\big] = \alpha_{i0} + \alpha_{i1}\omega_j, \quad i = 1,2,3, \qquad j = 1, \dots, n. \tag{28}$$

Let us recall that the vector $\boldsymbol{\alpha}_1$ is set to zero. Apparently, the mixing proportions can be expressed as:

$$\pi_i(\omega_j, \boldsymbol{\alpha}_i) = \frac{\mathrm{e}^{\alpha_{i0}+\alpha_{i1}\omega_j}}{\mathrm{e}^{\alpha_{10}+\alpha_{11}\omega_j} + \mathrm{e}^{\alpha_{20}+\alpha_{21}\omega_j} + \mathrm{e}^{\alpha_{30}+\alpha_{31}\omega_j}}\,, \quad i = 1,2,3, \ \ j = 1, \dots, n. \tag{29}$$

An example of such a mixture for a sample of size 50 is visualized in Figure 2. Apart from visualization of a data coming from a given three component mixture, true regression lines for individual clusters are demonstrated. As it was indicated above, this particular mixture of regression models causes severe inaccuracy in estimates. This problematic phenomenon is noticeable in Figure 3, where one fitted regression line from a standard mixture model is completely inaccurate due to incorrect classification, while a model with a concomitant variable fits all lines correctly.

In this type of configuration, a standard mixture model in many cases does not even estimate the right number of components, let alone remotely accurate regression parameters and component memberships of observations. 2 000 simulations were performed and the ratio of these highly imprecise estimates was 79% for a sample size of 50 and even 93% for a sample size of 300 (Table 5), which indicates that this is a systematic effect. Conversely, the proportion of inaccurate estimates obtained from a model with a concomitant variable is significantly smaller, accounting for only 27% for a sample size of 50 and decreasing to 6% for a sample size of 300. The estimates so dissimilar to the true parameter values that individual components cannot be recognized or efficiently identified were considered highly inaccurate. In practice, acceptance intervals for regression parameters $\boldsymbol{\beta}_i$ from a given component may be used. These intervals are as wide as possible to allow identification of a component and its distinction from the remaining components in the model. If no component or more components correspond to some acceptance interval, the whole mixture model is marked as inaccurate (see Figure 3, left).

Table 5 The ratio of entirely inaccurate estimates of parameters in a three component mixture of regression lines with a different space of the predictor from 2 000 simulations. The misclassification errors are evaluated from 200 correctly fitted models as well as the mean square errors for mixing proportion estimates

	n = 50		**n = 100**		**n = 300**	
	standard	concomitant	standard	concomitant	standard	concomitant
Inaccurate param. ratio	0.7850	0.2700	0.8200	0.1400	0.9250	0.0600
Err_M	0.2316	0.0703	0.1925	0.0456	0.1742	0.0392
$\mathrm{MSEPAR}(\hat{\pi}_1)$	0.0066	0.0018	0.0062	0.0005	0.0047	0.0002
$\mathrm{MSEPAR}(\hat{\pi}_2)$	0.0098	0.0046	0.0054	0.0010	0.0036	0.0002
$\mathrm{MSEPAR}(\hat{\pi}_3)$	0.0254	0.0023	0.0187	0.0003	0.0154	<0.0001

Source: Own construction

Table 6 The mean square errors (MSEPAR) and the mean variances (VAR) of the regression parameters, and standard error estimates for a three component mixture of regression lines with a different space of the predictor. Characteristics are calculated from 200 correctly fitted models

MSEPAR		β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	β_{30}	β_{31}	σ_3^2
n = 50	standard	35.7349	0.0388	42.2389	268.0571	0.3215	90.3459	277.8042	0.0527	33.0111
	concomitant	21.4600	0.1367	32.7982	154.9637	0.4086	74.5576	365.5693	0.0667	14.7752
n = 100	standard	7.6366	0.0096	6.4402	89.8041	0.1335	34.0492	194.2027	0.0367	12.4032
	concomitant	9.1976	0.0515	12.2703	55.8374	0.1272	24.2869	126.9932	0.0241	5.4217
n = 300	standard	2.3826	0.0029	2.2685	28.5231	0.0348	10.2868	91.1527	0.0178	5.2866
	concomitant	1.4447	0.0144	0.9264	21.0425	0.0384	6.3211	32.4438	0.0062	0.9661
VAR		β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	β_{30}	β_{31}	σ_3^2
n = 50	standard	16.3324	0.0314	0.0742	116.0794	0.1842	0.0329	40.9276	0.0097	0.0512
	concomitant	11.8080	0.0927	0.0425	103.0376	0.2121	0.0262	164.2889	0.0319	0.0497
n = 100	standard	6.0410	0.0086	0.0364	53.9657	0.0841	0.0151	26.5701	0.0056	0.0315
	concomitant	4.7048	0.0353	0.0235	49.8695	0.0975	0.0117	81.7340	0.0155	0.0260
n = 300	standard	2.0882	0.0030	0.0115	18.1297	0.0299	0.0047	13.1638	0.0027	0.0125
	concomitant	1.3823	0.0108	0.0080	16.6439	0.0314	0.0040	27.1625	0.0050	0.0087

Source: Own construction

In order to evaluate the quality of estimates, all entirely inaccurate estimated models were identified and discarded. Quality characteristics for both types of models evaluated from 200 correctly estimated models are reported in Table 6. In contrast to the previous simulation study (Table 2), the superiority of a model with a concomitant variable is not so apparent and the accuracy of estimators from both models is comparable. The misclassification error is still much worse in a standard mixture and the same goes for the MSEPAR of mixing proportion estimates (Table 5). However, it should be kept in mind that the quality characteristics were calculated from the correctly estimated models and that a standard mixture tends to give entirely inaccurate results. Therefore, for this type of regression function configuration, a mixture model with a concomitant variable should be only used for the estimation.

CONCLUSION

The paper is focused on a concomitant variable introduced by Grün and Leisch (2008) and its role in the mixture of regression models. Two representative simulation studies were performed in order to assess the quality of regression estimates and clustering properties of both a standard mixture of regression models and a mixture of regression models with concomitant variables. Obviously, the possibilities of mixture models setting are various and this paper is focused only on two of them. However, the models presented here were chosen as a representative sample, assuming at the same time that each model works with different number of components, diverse distributions of predictor, various regression lines configuration and, most importantly, distinct characters of the concomitant variable.

The results of both studies indicate that the concomitant variables present a beneficial extension of mixture models. In case of a categorical concomitant variable, the results are straightforward and provide evidence in favour of a mixture model including a concomitant variable, since for this model, both the mean square error and the mean variance of estimates are, with very few exceptions, smaller. These characteristics are not so unambiguous for a three component mixture and a covariate as a concomitant variable. However, these indicators are only valid for a small portion of estimates that are close enough to the true values of parameters. In practice, the ratio of highly inaccurate estimates is more informative and is significantly reduced as a concomitant variable is added into the model.

Clustering properties are assessed through the mean misclassification error of each model. Again, a concomitant variable enhances estimated component membership in both cases, especially for a small sample size. In general, concomitant variables themselves prove to be useful in the mixture of regression models. Particularly, the concomitant variable in a form of the predictor itself seems to be a common choice for reasonable regression parameters estimates.

As models in the mixture get more complicated, estimates can become less precise and reliable. Nevertheless, the conclusions of the simulation study remain similar as the concomitant variables still enhance the performance of the mixture of regression models for both categorical and continuous concomitant variables.

ACKNOWLEDGMENT

The authors are grateful to the referees for their helpful comments that significantly improved this article. The authors also gratefully acknowledge the support received from the grant IGA_PrF_2016_025 and IGA_PrF_2017_019 of the Internal Grant Agency of Palacký University Olomouc.

References

BEHBOODIAN, J. On a mixture of normal distributions. *Biometrika*, 1970, 57, pp. 215–217.

BENGALIA, T., CHAUVEAU, D., HUNTER, D. R., YOUNG, D. S. Mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, 2009, 32(6), pp. 1–29.

DE VEAUX, R. D. Mixtures of Linear Regressions. *Computational Statistics & Data Analysis*, 1989, 8(3), pp. 227–245.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. Maximum Likelihood from Incomplete Data Via the EM-Algorithm. *Journal of the Royal Statistical Society*, Series B, 1977, 39, pp. 1–38.

DESARBO, W. S. AND CRON, W. L. A Maximum Likelihood Methodology for Clusterwise Linear Regression. *Journal of Classification*, 1988, 5(2), pp. 249–282.

FARIA, S. AND SOROMENHO, G. Fitting Mixtures of Linear Regressions. *Journal of Statistical Computation and Simulation*, 2010, 80(2), pp. 201–225.

GOLDFELD, S. M. AND QUANDT, R. E. Techniques for Estimating Switching Regressions. In: *Studies in Nonlinear Estimation*, Cambridge, Massachussets, Ballinger, 1976, pp. 3–35.

GRÜN, B. AND LEISCH. F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, 2008, 28(4), pp. 1–35.

GRÜN, B., SCHARL, T., LEISCH, F. Modelling Time Course Gene Expression Data with Finite Mixtures of Linear Additive Models. *Bioinformatics*, 2012, 28(2), pp. 222–228.

HAMEL, S., YOCCOZ, N. G., GAILLARD, J. M. Assessing Variation in Life-history Tactics within a Population Using Mixture Regression Models: A Practical Guide for Evolutionary Ecologists. *Biological Reviews*, Cambridge Philosophical Society, 2017, 92(2), pp. 754–775.

LEISCH, F. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 2004, 11(8), pp. 1–18.

MCLACHLAN, G. AND PEEL, D. *Finite Mixture Models.* New York: John Wiley & Sons, 2000.

NITYASUDDHI, D. AND BÖHNING, D. Asymptotic Properties of the EM Algorithm Estimate for Normal Mixture Models with Component Specific Variances. *Computational Statistics & Data Analysis*, 2003, 41, pp. 591–601.

PEARSON, K. Contributions to the Mathematical Theory of Evolution. *The Royal Society*, 1894, 185, pp. 71–110.

QUANDT, R. E. AND RAMSEY, J. B. Estimating Mixtures of Normal Distributions and Switching Regressions. *Journal of the American Statistical Association*, 1978, 73, pp. 730–752.

R CORE TEAM. R: A *Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing, 2016.

REDNER, R. A. AND WALKER, H. F. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, 1984, 26, pp. 195–239.

VAŇKÁTOVÁ, K. AND FIŠEROVÁ, E. Analysis of Income of EU Residents Using Finite Mixtures of Regression Models. In: *34th International Conference Mathematical Methods in Economics MME 2016 – Conference proceedings.* Liberec: Technical University of Liberec, 2016, 1, pp. 875–880.