A semiparametric mixture regression model for longitudinal data

DOI:
10.1080/15598608.2017.1298062

Document Version
Accepted author manuscript

A semiparametric mixture regression model for longitudinal data

Authors: Tapio Nummi
– School of Information Sciences, University of Tampere, Finland (tan@uta.fi)

Janne Salonen
– Research Department, The Finnish Centre for Pensions, Finland (Janne.Salonen@etk.fi)

Lasse Koskinen
– School of Management, University of Tampere, Finland (Lasse.Koskinen@uta.fi)

Jianxin Pan
– School of Mathematics, The University of Manchester, UK (Jianxin.Pan@manchester.ac.uk)

Abstract:

- A normal semiparametric mixture regression model is proposed for longitudinal data. The proposed model contains one smooth term and a set of possible linear predictors. Model terms are estimated using the penalized likelihood method with the EM-algorithm. A computationally feasible alternative method that provides an approximate solution is also introduced. Simulation experiments and real data example are used to illustrate the methods.

Key-Words:

- *Curve Clustering; EM-algorithm; Finite Mixtures; Growth Curves.*

AMS Subject Classification:

- 62G05, 62B99, 62J07.

1. INTRODUCTION

Modeling of longitudinal data have been of special interest in statistics during recent decades. Depending on the context several approaches have been used: multivariate analysis, linear and generalized linear mixed and mixture models, structural equation models, Bayesian methods, quantile-regression etc. For comprehensive summaries of different approaches to longitudinal data analysis we can refer to Fitzmaurice et al. (2011) and Diggle et al. (2013), for example.

In our approach the focus is on the situation, where the studied population is not completely homogenous over time, but is instead comprised of groups of individuals with the same kind of mean developmental profiles. One approach to understanding such heterogeneity is to apply the theory of Finite Mixtures (FM). Nagin (1999 and 2005) and Jones et al. (2001) applies the generalized linear models theory to FM with the assumption that observations within a given mixture are independent. A further extension is to take some model parameters (e.g., polynomial coefficients) as random variables or (latent factors), see, e.g., Muthen and Khoo (1998). These random terms can then be used for modeling the correlation of the observations within a component mixture. The other kind of mixture regression application arises if part of the random model parameters arise from a mixture distribution (see e.g. Verbeke and Lesaffre, 1996).

The focus in the present study is especially on modeling the mean within the mixture using semiparametric regression techniques (Nummi et al. 2011 and Nummi et al. 2013). The mean consists of one time-dependent smooth term and a set of linear predictors that may or may not depend on time. Model terms are estimated using the penalized likelihood method with the EM algorithm. The present study also introduces a computationally feasible alternative that provides an approximate solution using an ordinary linear models methodology developed

for mixture regression. The data analysis part of the study consists of a simulation experiment and an analysis of real longitudinal data set of growth characteristics of Finnish children.

Section 2 introduces the basic multivariate normal mixture model and its parameter estimation with the maximum likelihood method. Then, the basic model is extended to the semiparametric mean model. Parameter estimation using penalized likelihood with the EM algorithm is introduced in detail. Section 3 introduces a method for obtaining a computationally feasible approximate solution for a semiparametric mean trajectory model and a simulation study was used to demonstrate the performance of the technique. The section closes by the real data analysis of growth curves of Finnish children. Finally, Section 4 summarizes the main results.

2. DESCRIPTION OF THE PROBLEM

2.1. Theoretical background

The aim is to identify clusters of individuals with the same kind of developmental curves. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip_i})'$ represent the sequence of measurements on individual i over p_i periods and let $f_i(\boldsymbol{y}_i|\boldsymbol{X}_i)$ denote the marginal probability distribution of $\boldsymbol{y}_i$ with possible time dependent covariates $\boldsymbol{X}_i$. It is assumed that $f_i(\boldsymbol{y}_i|\boldsymbol{X}_i)$ follows a mixture of K densities

$$f_i(\boldsymbol{y}_i|\boldsymbol{X}_i) = \sum_{k=1}^{K} \pi_k f_{ik}(\boldsymbol{y}_i|\boldsymbol{X}_i), \quad \sum_{k=1}^{K} \pi_k = 1 \text{ with } \pi_k > 0, \tag{2.1}$$

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

get

$$f_{ik}(\boldsymbol{y}_i|\boldsymbol{X}_i) = (2\pi)^{-\frac{p_i}{2}} \mid \boldsymbol{\Sigma}_{ik} \mid^{-\frac{p_i}{2}} \exp\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_{ik})'\boldsymbol{\Sigma}_{ik}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{ik})\}, \tag{2.2}$$

where $\boldsymbol{\mu}_{ik}$ is a function of covariates $\boldsymbol{X}_i$ with parameters $\boldsymbol{\theta}_k$ and $\boldsymbol{\Sigma}_{ik}$ is a variance-covariance matrix within the kth component, involving $\boldsymbol{\sigma}_k$, which is a vector of unique covariance parameters. The parameter estimates can then be obtained by maximizing the log-likelihood function for the entire set of N (independent) individuals $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$

$$l(\boldsymbol{\phi} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \sum_{i=1}^{N} \log f_i(\boldsymbol{y}_i|\boldsymbol{X}_i) \tag{2.3}$$

over all unknown parameters $\boldsymbol{\phi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_K)'$. A popular method for the Maximum Likelihood (ML) estimation is the EM (Expectation and Maximization) algorithm Dempster et al. (1977) that is often used, for example, for incomplete data problems. The EM algorithm is an iterative method consisting of two main steps. The E-step finds the expected log-likelihood under current parameter estimates, and the subsequent M-step maximizes the expected log-likelihood function. These two steps are then iterated until convergence. The mixture model EM algorithm implementation details can be found, for instance, in McLachlan and Peel (2000).

The basic mean model in applications is often a simple linear model, e.g. an appropriate low degree polynomial, in time. For many appropriately smooth curves, this provides a reasonable model. However, in certain cases, a low degree polynomial may not prove to be sufficient due to irregular or insufficient measuring points or otherwise complicated mean curve forms, for example. The aim here is to introduce a new, more flexible semiparametric model with one possible smooth term (time in our application) that can be used for mean curve modeling with normal mixture components. The important advantage is that smoothing is done separately for each mixture component and thus a very rich set of curves are available for modeling.

2.2. Modeling the conditional mean

The set of covariates $\boldsymbol{X}_i$ is divided into the parametric part $\boldsymbol{U}_i$ and to the non-parametric part $\boldsymbol{t}_i$, where $\boldsymbol{t}_i$ is the vector of measuring times $t_{i1}, \ldots, t_{ip_i}$. For the ith individual within the kth mixture we assume the semiparametric model

$$\boldsymbol{y}_{ik} = \boldsymbol{g}_{ik} + \boldsymbol{U}_i\boldsymbol{b}_k + \boldsymbol{\epsilon}_{ik}, \tag{2.4}$$

where $\boldsymbol{g}_{ik} = [g_k(t_{i1}), \ldots, g_k(t_{ip_i})]'$ is a smooth vector of twice differentiable functions evaluated at $\boldsymbol{t}_i$, $\boldsymbol{U}_i$ is a matrix of h covariates (constant term not included) and $\boldsymbol{b}_k$ is a parameter vector to be estimated. Note that the same measuring points are used for each individual, but the measurement sequence (number of measurements actually taken) may vary from individual to individual. The covariance matrix of random errors $\boldsymbol{\epsilon}_i$ for the kth group takes the simple form $\boldsymbol{\Sigma}_k = \sigma_k^2\boldsymbol{I}$ (Nagin 1999 and 2005). For more elaborated covariance modeling, we may refer to, for example, Ye and Pan (2006) and Leng et al. (2010).

We can define the so-called roughness matrix as $\boldsymbol{G} = \boldsymbol{\nabla}\boldsymbol{\Delta}^{-1}\boldsymbol{\nabla}'$ (from the penalty $\int g''^2$), where the non-zero elements of banded $p \times (p-2)$ and $(p-2) \times (p-2)$ matrices $\boldsymbol{\nabla}$ and $\boldsymbol{\Delta}$ are defined as

$$\nabla_{l,l} = \frac{1}{h_l},\ \nabla_{l+1,l} = -(\frac{1}{h_l} + \frac{1}{h_{l+1}}), \nabla_{l+2,l} = \frac{1}{h_{l+1}}$$

and

$$\Delta_{l,l+1} = \Delta_{l+1,l} = \frac{l_{k+1}}{6},\ \Delta_{l,l} = \frac{h_l + h_{l+1}}{3},$$

where $h_j = t_{j+1} - t_j, j = 1, 2, ..., (p-1)$ and $l = 1, 2, ..., (p-2)$ (see e.g. Green and Silverman, 1994). The penalized log-likelihood function is now

$$l(\boldsymbol{\phi} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \sum_{i=1}^{N} log\{\sum_{k=1}^{K} \pi_k f_{ik}\} - \sum_{k=1}^{K}\{\frac{\alpha_k}{2}\boldsymbol{g}_k'\boldsymbol{G}\boldsymbol{g}_k\}, \tag{2.5}$$

where α_k is a smoothing parameter and $\boldsymbol{\phi}$ is a vector of unknown parameters. Maximizing this log-likelihood is computationally intensive. The next section shows how the solution can be obtained using the iterative EM algorithm.

2.3. Estimation with the EM algorithm

In this section, we show how the semiparametric mixture model can be estimated using the EM algorithm. In this implementation, estimation is viewed as a missing data problem (see also McLachlan and Peel, 2000). We denote

$$\boldsymbol{y}_i^* = (\boldsymbol{y}_i', \boldsymbol{z}_i')',$$

where $z_{ik} = 1$ if $\boldsymbol{y}_i$ stemmed from the kth component; otherwise, $z_{ik} = 0$. The vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ can now be seen as realized values of random vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ from the multinomial distribution. The complete-data, joint log-likelihood function of $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ can be written as

$$l_c(\boldsymbol{\phi}) = \sum_{i=1}^{N}\{\sum_{k=1}^{K} z_{ik}[log(\pi_k) + log(f_{ik})]\} - \sum_{k=1}^{K}\frac{\alpha_k}{2}\boldsymbol{g}_k'\boldsymbol{G}\boldsymbol{g}_k. \tag{2.6}$$

The algorithm's E step is simply to calculate the conditional expectation of $l_c(\boldsymbol{\phi})$ under current parameter estimates $\hat{\boldsymbol{\phi}}$ and the observed data. This yields

$$E(Z_{ik} \mid \hat{\boldsymbol{\phi}}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \frac{\hat{\pi}_k f_{ik}(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \hat{\boldsymbol{\xi}}_k)}{\sum_{l=1}^{K}\hat{\pi}_l f_{il}(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \hat{\boldsymbol{\xi}}_l)} = \hat{z}_{ik}, \tag{2.7}$$

where $\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_K$ are vectors consisting of estimates of mixing distribution mean and variances. In the (M step) the expected log-likelihood for the completed data

$$E[l_c(\boldsymbol{\phi})] = \sum_{i=1}^{N}\{\sum_{k=1}^{K}\hat{z}_{ik}[log(\pi_k) + log(f_{ik})]\} - \sum_{k=1}^{K}\frac{\alpha_k}{2}\boldsymbol{g}_k'\boldsymbol{G}\boldsymbol{g}_k \tag{2.8}$$

is maximized. Note that for the kth component we may denote $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$, $\boldsymbol{U} = (\boldsymbol{U}_1', \ldots, \boldsymbol{U}_N')'$ and $\boldsymbol{W}^k = diag(\boldsymbol{W}_{k1}, \ldots, \boldsymbol{W}_{kN})$, where $\boldsymbol{W}_{ki} = \hat{z}_{ik}\boldsymbol{I}_i$. The expected log-likelihood for the kth component ($\times 2$) can be written as

$$-\frac{1}{\sigma_k^2}[\boldsymbol{y} - (\boldsymbol{U}\boldsymbol{b}_k + \boldsymbol{N}\boldsymbol{g}_k)]'\boldsymbol{W}^k[\boldsymbol{y} - (\boldsymbol{U}\boldsymbol{b}_k + \boldsymbol{N}\boldsymbol{g}_k)] - N_k log(\sigma_k^2) - \alpha_k\boldsymbol{g}_k'\boldsymbol{G}\boldsymbol{g}_k \tag{2.9}$$

where $N_k = \sum_{i=1}^{N} p_i\hat{z}_{ik}$. The solutions are obtained at

$$\hat{\boldsymbol{b}}_k = [\tilde{\boldsymbol{U}}'\boldsymbol{U}]^{-1}\tilde{\boldsymbol{U}}'\boldsymbol{y} \quad \text{and} \quad \boldsymbol{N}\hat{\boldsymbol{g}}_k = \boldsymbol{S}(\boldsymbol{y} - \boldsymbol{U}\hat{\boldsymbol{b}}_k),$$

where $\tilde{\boldsymbol{U}} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{W}^k\boldsymbol{U}$ and $\boldsymbol{S} = \boldsymbol{N}(\boldsymbol{N}'\boldsymbol{W}^k\boldsymbol{N} + \alpha_k\boldsymbol{G})^{-1}\boldsymbol{N}'\boldsymbol{W}^k$ is the smoother matrix, where $\boldsymbol{N}$ is an incidence matrix. Note that the maximizing curve $\hat{\boldsymbol{g}}_k$ is a natural cubic smoothing spline with knots at the design points $t_1, \ldots, t_p$. The conditions for uniqueness of the solutions turns out to be identical to the fully parametric regression with explanatory variables $\boldsymbol{t}_i$ and $\boldsymbol{U}_i$ (Green and Silverman, 1994). Estimates for σ_k^2 and π_k can be obtained from

$$\hat{\sigma}_k^2 = \frac{1}{N_k}[\boldsymbol{y} - (\boldsymbol{U}\hat{\boldsymbol{b}}_k + \boldsymbol{N}\hat{\boldsymbol{g}}_k)]'\boldsymbol{W}^k[\boldsymbol{y} - (\boldsymbol{U}\hat{\boldsymbol{b}}_k + \boldsymbol{N}\hat{\boldsymbol{g}}_k)] \text{ and } \hat{\pi}_k = \sum_{i=1}^{N} \hat{z}_{ik}/N$$

with $\sum_{k=1}^{K} \hat{\pi}_k = 1$. A further simplification of the M-step is easily obtained for complete and balanced data (parametric part dropped) using

$$\hat{\boldsymbol{g}}_k = (\hat{\pi}_k N\boldsymbol{I} + \alpha_k\boldsymbol{G})^{-1} \sum_{i=1}^{N} \hat{z}_{ik}\boldsymbol{y}_i$$

and

$$\hat{\sigma}_k^2 = \frac{1}{N_k} \sum_{i=1}^{N} \hat{z}_{ik}(\boldsymbol{y}_i - \hat{\boldsymbol{g}}_k)'(\boldsymbol{y}_i - \hat{\boldsymbol{g}}_k).$$

To update the value of the smoothing parameter α_k the following idea is introduced. The profile log-likelihood for the kth component given $\boldsymbol{y}$, $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_N$, $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N$ and $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N$ is written as a function of the smoothing parameter only. This yields to $l(\alpha) = -N_k - N_k log[\hat{\sigma}_k^2(\alpha)]$ and the maximum is obtained when $\hat{\sigma}_k^2(\alpha)$ is minimized with respect to α. When $\alpha_1, \ldots, \alpha_K$ are updated also the estimates for $\sigma_1^2, \ldots, \sigma_K^2$, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K$ and $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_K$ are readily available. Since each component is smoothed individually, the method allows a very flexible modeling tool within each of the K components of the mixture model. The EM steps are iterated until convergence. However, in some cases, the algorithm may converge to a local maximum. Therefore, in practice many initial values are usually tested. For more detailed considerations of the EM algorithm in a similar kind of context we can refer to Fariaa and Soromenhobre (2010) and to Basford and McLahlan (1985).

Identifiability is a crucial issue in mixture modeling. This topic for normal mixture is studied quite extensively in Titterington et al. (1985) and McLachlan

and Peel (2000). For the studies of normal mixture regression we can refer to Huang and Yao (2012) and of normal nonparametric mixture regression to Huang et al. (2013). Especially, the results in the later paper are applicable here since the semiparametric regression model of this paper can be considered as a special case of their more general class of models.

Selection of the number of components K is a subject of lively scientific debate. Many statistical criteria have been presented for the purpose, of which the most important are the information criterion functions, especially AIC and BIC. In practice also the overall fit and the interpretability of the components must be taken into account. See McLachlan and Rathnayake (2014) for a review article of the topic.

In practical implementations, individuals are often assigned to groups or clusters $c_1, \ldots, c_K$ according to posterior probabilities $\hat{z}_{ik}$. This is often done using maximum posterior probability $max\{\hat{z}_{ik}\}$ or by random integers generated using $\hat{z}_{ik}$ as probabilities. This assignment of individuals to specific clusters can be seen as an important contribution to longitudinal data analysis. This is because many important latent characteristics manifest themselves only when analyzing longitudinal data. However, further statistical analysis of the identified clusters must be accomplished very carefully since they are not fixed constructs, but are based on probabilities.

3. DATA ANALYSIS

3.1. Computing using an approximation

In the following, we present a simple method to estimate the semiparametric model using standard statistical software (e.g. Jones et al. 2001, Leisch 2004, Muthen and Muthen 2007) developed for mixture regression. The method is based on the spline approximation. For the ith individual in the kth trajectory group (indices dropped), we have the semiparametric model

$$\boldsymbol{\mu} = \boldsymbol{g} + \boldsymbol{U}\boldsymbol{b}, \tag{3.1}$$

where we have the estimate $\hat{\boldsymbol{b}} = [\tilde{\boldsymbol{U}}'\boldsymbol{U}]^{-1}\tilde{\boldsymbol{U}}'\boldsymbol{y}$ and $\hat{\boldsymbol{g}} = \boldsymbol{S}_\alpha(\boldsymbol{y} - \boldsymbol{U}\hat{\boldsymbol{b}})$, $\boldsymbol{S}_\alpha = (\boldsymbol{I} + \alpha\boldsymbol{G})^{-1}$ and $\tilde{\boldsymbol{U}} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{U}$. The whole semiparametric curve is then fitted by

$$\hat{\boldsymbol{\mu}} = \boldsymbol{S}\boldsymbol{y} + \tilde{\boldsymbol{U}}\hat{\boldsymbol{b}}. \tag{3.2}$$

For the smoother matrix $\boldsymbol{S}$ we can show that

$$\boldsymbol{S} = \boldsymbol{M}(\boldsymbol{I} + \alpha\boldsymbol{\Lambda})^{-1}\boldsymbol{M}', \tag{3.3}$$

where $\boldsymbol{M}$ is the matrix of p orthogonal eigenvectors of the roughness matrix $\boldsymbol{G}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix of corresponding p eigenvalues $\lambda_1, \ldots, \lambda_p$. Note that $\boldsymbol{G}$ and $\boldsymbol{S}$ share the same set of eigenvectors, but in the reverse order. Subsequently, we assume that eigenvectors $\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_p$ of $\boldsymbol{M}$ are ordered according to the eigenvalues $\gamma = 1/(1 + \alpha\lambda)$ of $\boldsymbol{S}$. The sequence of these eigenvectors appears to increase in complexity like a sequence of orthogonal polynomials and the first two eigenvalues are always 1 (corresponding eigenvectors span a straight line model, see e.g. Ruppert et al., p. 79, 2005). We can then approximate $\boldsymbol{S}$ by $\boldsymbol{P} = \boldsymbol{M}_c\boldsymbol{M}_c'$, where $\boldsymbol{M}_c$ contains the first c eigenvectors of $\boldsymbol{M}$. The number c of needed eigenvectors can be estimated using ordinary model selection criteria

like AIC, BIC, etc. (for more details see Nummi et al. 2011 and Nummi et al. 2013). The fit of the model (3.2) is approximated by fitting the approximating mean model

$$\boldsymbol{\mu}_* = \boldsymbol{M}_c\boldsymbol{\gamma} + \boldsymbol{U}\boldsymbol{b}. \tag{3.4}$$

Thus estimating the semiparametric mean model is now returned to the linear model framework. Therefore we can quite easily apply the common mixture regression statistical software for our analysis.

A simulation study was conducted to test how well the approximation method perform when the data are generated using different, but closely behaving, curve forms. Following models were used to simulate the data

a) $y_j = 0.1 + 1.5x_j - 0.1x_j^2 + d_a z_j + \epsilon_j,$

b) $y_j = 0.1 + 1.5x_j - 0.1x_j^2 + d_b z_j + \epsilon_j,$

where $\epsilon_j \sim N(0, 0.25)$, $z_j = cos(0.5\pi x_j)$, $x_j = j, j = 1, \ldots, 10$, $d_a = 0.8$ and $d_b = 0$. The series of 10 measurements were repeated 100 times for each model. For these 200 series of measurements completely random dropouts were also generated with a dropout probability for a single measurements as $p_j = 0.2, j = 2, \ldots, 10$ (no dropouts in x_1).

For the simulated data mixture regression analysis was performed. First, the true semiparametric mixture model was fitted with $g(x)$ as the nonparametric term and z as the parametric term (method 1). This is then compared with the fit provided by approximating model, where first five eigenvectors $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_5$ and z are used as explanatory variables (method 2). For both methods 20 runs with different starting values were tested with $K = 1, 2, 3, 4$. The following BIC values were observed: method 1) 1646.629, 1559.574, 1612.512 and 1666.069; method 2) 1637.456, 1536.163, 1565.656 and 1604.579. Clearly $K = 2$ gives the minimum and this is therefore taken as the number of groups for both methods.

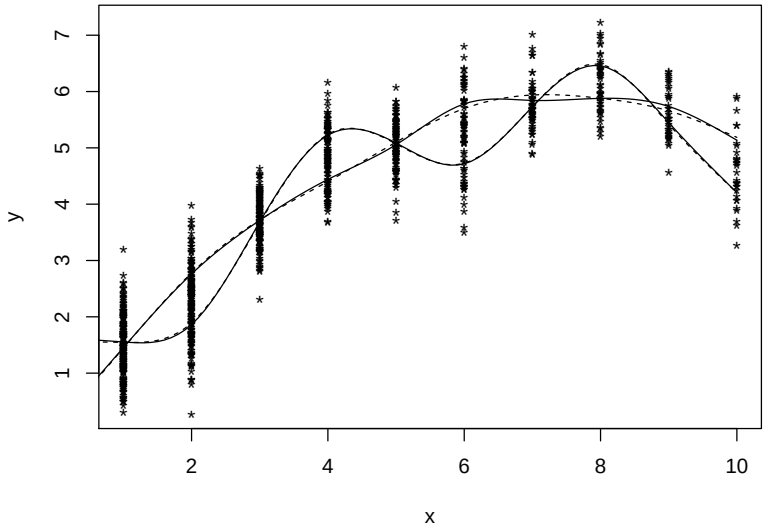

Figure 1: Plot of simulated data and conditional means. Solid line corresponds the true semiparametric model (method 1) and dotted curve corresponds the linear model approximation with $c = 5$ (method 2).

Figure 1 gives the plot of simulated data and the means in $x_j, j = 1, \ldots, 10$ for the identified groups.

The fit of these two methods were very close to each other. First the mixing proportion estimates were very close: $\hat{\pi}_{11} = 0.46; \hat{\pi}_{12} = 0.45$ (group 1) and $\hat{\pi}_{21} = 0.54; \hat{\pi}_{22} = 0.55$ (group 2). The conditional means at points $x_j, j = 1, \ldots, 10$ were also very close for both groups. For group 1 the fitted curves almost completely overlap and for the group 2 only a slight difference for the last points of x_j $(j > 5)$ is observed. This demonstrates that the approximation works very well when the semiparametric mixture regression model with one smooth term and parametric part is approximated by the proposed linear model.

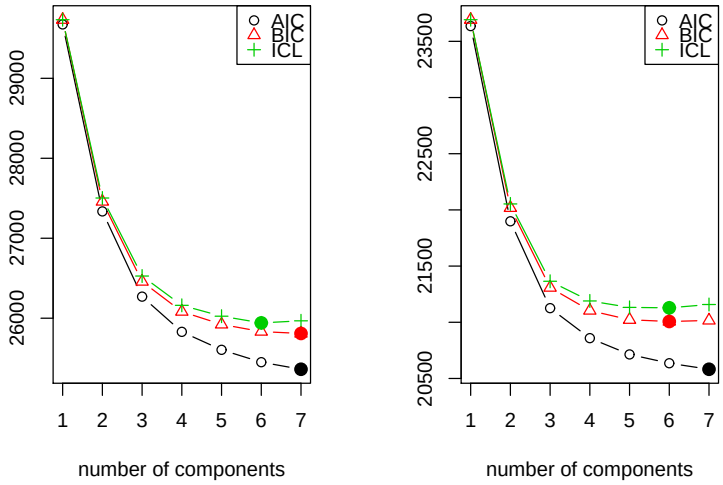

Figure 2: AIC, BIC and ICL values of the fitted models for $k = 1, \ldots, 7$ (males on the left-hand side and females on the right-hand side).

3.2. Analysis of height growth

The data used for this study is a part of the data of growth measurements of 4,223 children collected in Finland (Vuorela 2011 and Nummi et al. 2014). Birth cohorts from five years were examined in original data: 1974 (n=1,108), 1981 (n=987), 1991 (n=586), 1995 (n=786) and 2001 (n=766). However, for our study we considered only the birth cohort 1974. The children were measured in well-baby clinics, schools and health care centers from birth up to age 15. The data included anthropometric measurements at birth and seven routine health checkup times: at six months and, 1, 2, 5, 7, 12, and 15 years. In addition, the gender, the area of residence (urban/rural), and the mother's pregnancy weeks were also included.

Understanding human growth during childhood and adolescence has been of special interest for pediatricians, health scientists, and the clothing industry,

among others. Statistical models for growth have been investigated by Gasser et al. (1984), Poortema (1989), and Karlberg (1987), for example. A recent overview of analytical strategies of human growth is presented in Johnson (2015). In statistical models, growth is often divided into age periods. For example, Karlberg (1987) applied the following models:

1. Infancy: $y = a + b\{1 - exp(-ct)\} + \epsilon$,
2. Childhood: $y = a + bt + ct^2 + \epsilon$
3. Puberty: $y = a/[1 + exp\{-b(t - t_*)\}] + \epsilon$,

where y is height, t is the age, a, b and c are parameters to be estimated, and t_* is the peak velocity age. Naturally, the age period in which each of the models applies varies from individual to individual. It is also well known that infant birth weights influence further childhood development, including mortality and morbidity. As a result, it could be interesting to use the birth weight as a parametric term and evaluate its effects on different mean developmental curves. The basic model for the ith individual in the kth group takes the form

$$y_{ij} = g_k(t_{ij}) + \beta_k u_i + \epsilon_{ij},$$

where u_i is the birth weight a child and ϵ_{ij} is independent and identically normally distributed random error term with $Var(\epsilon_{ij}) = \sigma_k^2$.

The data were first divided into two parts by gender, because it is well known that the growth curves differ. The actual analysis started by fitting the cubic smoothing spline over both data sets when $K = 1$ and the smoothing parameter was then estimated using the method of generalized cross-validation. The estimated degrees of freedom (EDF) for a smoother were ≈ 7.998 for both data sets. Therefore, a natural choice for the approximation model dimension is $c = 7$. This gives us seven first eigenvectors of $\boldsymbol{S}$ that are used in approximation models.

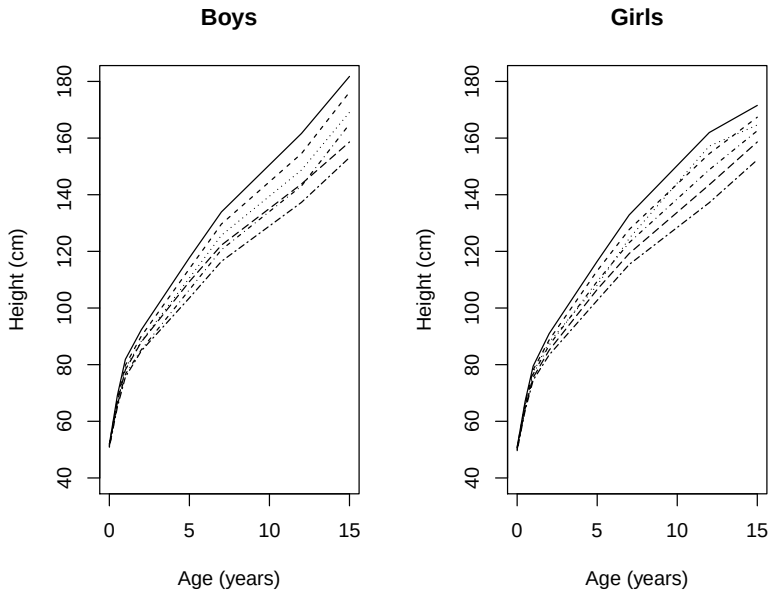

Figure 3: Fitted trajectory curves $\hat{\boldsymbol{\mu}}_{*k} = \boldsymbol{M}_5\hat{\boldsymbol{\gamma}}_k + \boldsymbol{u}\hat{b}_k$ of the final models when birth weight u_i is set to the mean value (males on the left and females on the right hand side).

The approximation model was fitted for $k = 1, \ldots, 7$ and the corresponding criterion values are plotted in Figure 2. It is clear from Figure 2, that for both genders, the decrease in criterion values when $k > 6$ is relatively small. Therefore, we took $k = 6$ and $k = 7$ as possible candidate models. However, the graphical investigation of the fitted trajectory curves revealed that $k = 7$ may not provide any new relevant information from the interpretation point of view. Therefore, our choice was $K = 6$ for both genders. The fitted curves are presented in Figure 3 with model covariates fitted to their mean values.

The parameter estimates of each of the groups are given in Table 1. Clearly, birth weight has some effect and the effects are not similar for genders. For boys the estimates $\hat{\beta}_{km}$ does not vary much over the groups. However, the smallest estimate $\hat{\beta}_{3m} = 2.042$ was obtained for the largest group 3. For girls the estimates vary depending on the group. Interestingly, the largest estimate $\hat{\beta}_{6m} = 4.043$ is

Table 1: Model parameter estimates for both genders. The groups are set to decreasing order according to the level of the mean curve at the end of the follow-up period.

Group	$\hat{\pi}_M$	$\hat{\pi}_F$	$\hat{\beta}_{1M}$	SE($\hat{\beta}_{1M}$)	$\hat{\beta}_{1F}$	SE($\hat{\beta}_{1F}$)
1:	0.0842	0.1095	2.986	0.3178	2.035	0.259
2:	0.1969	0.2329	2.285	0.1960	1.796	0.223
3:	0.3497	0.0697	2.042	0.1405	3.954	0.353
4:	0.1292	0.3196	2.520	0.2586	2.411	0.174
5:	0.1431	0.1910	2.647	0.2207	2.801	0.264
6:	0.0970	0.0774	2.668	0.2570	4.043	0.790

obtained for the group 6 where the level of the mean curve is the lowest (Figure 3). It seems possible that birth weight is an important factor in the development of further height growth. Especially, this finding is very interesting for girls. However, further analysis of this connection is a topic of further research work.

4. CONCLUDING REMARKS

The aim of this study was to apply nonparametric regression techniques for mean modeling of normal mixtures. Here, the mean consisted of one time-dependent smooth term and a set of linear predictors that may or may not depend on time. It was also shown how to obtain a computationally simple approximate solution. We believe that our approach provides a new, more flexible method, for the analysis of normal mixtures. Modeling the within-trajectory covariance matrix remains an interesting challenge for further research. Further analysis of height or weight growth data with different statistical methods using more background covariates also remains a topic of a future study.

ACKNOWLEDGMENTS

Authors like to thank the referees who gave valuable comments that led to improvements in the manuscript.

REFERENCES

[1] Basford K.E. and McLahlan G.J. (1985). Likelihood Estimation with Normal Mixture Models, *Applied Statistics*, **34**, 3, 282–289.

[2] Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood estimation for incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, B*, **39**, 1-38.

[3] Diggle, P., Heagerty, P., Liang, K–Y. and Zeger, S. (2013). *Analysis of Longitudinal Data*, Oxford: Oxford University Press, 2nd ed.

[4] Fariaa S. and Soromenho G. (2010). Fitting mixtures of linear regressions, *Journal of Statistical Computation and Simulation*, Vol. 80, No. 2, 201–225.

[5] Fitzmaurize, G. M., Laird, N. M. and Ware, J. H. (2011). *Applied Longitudinal Analysis*, Hoboken, N. J.: Wiley, 2nd ed.

[6] Gasser T., Muller H.G., Kohler W., Molinari L., Prader A. Nonparametric Regression Analysis of Growth Curves, *The Annals of Statistics* 1984; 12:210–229.

[7] Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach*. Monographs on Statistics and Applied Probability 58, Chapman Hall/CRC.

[8] Huang, M. and Yao, W. (2012). *Mixture Regression Models With Varying Mixing Proportions: A semiparametric Approach*, *Journal of the American Statistical Association*, **107**, 711–724.

[9] Huang, M., Li, R. and Wang, S. (2013). *Nonparametric Mixture Regression Models*, *Journal of the American Statistical Association*, **108**, 929–941.

[10] Johnson, W. (2015). Human biology toolkit: Analytical Strategies in Human Growth Research, *American Journal of Human Biology*, **27**, 69–83.

[11] Jones, B., Nagin, D. and Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories, *Sociological Methods & Research*, **29**, p. 374-393.

[12] Karlberg J. (1987). On the modeling human growth, *Statistics in Medicine*, **6**, 185–192.

[13] Leisch, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R, *Journal of Statistical Software*, Vol. **11**, Issue 8, p. 1–18.

[14] Leng, C., Zhang, W. and Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data, *Journal of the American Statistical Association*, **105**, No. 489, 181-193. DOI: 10.1198/jasa.2009.tm08485

[15] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, John Wiley and Sons Inc, New York.

[16] McLachlan, G. and Rathnayake (2014). *WIREs Data Mining Knowl Discov 2014*, doi: 10.1002/widm.1135, John Wiley & Sons.

[17] Muthen, B. and Khoo, S.T. (1998). Longitudinal studies of achievement growth using latent variable modeling. *Learning and Individual Differences*, Special issue: latent growth curve analysis, **10**, 73–101.

[18] Muthen, L. and Muthen, B. (2007). *Mplus User's Guide*, Sixth Edition, Los Angeles, CA: Muthen & Muthen.

[19] Nagin, D. (1999). Analyzing developmental trajectories: semiparametric, group-based approach, *Psychological Methods*, **4**, p. 39-177.

[20] Nagin, D. (2005). *Group-based modeling of development*, Cambridge, MA: Harvard University Press.

[21] Nummi, T., Pan, J., Siren, T. and Liu, K. (2011). Testing for Cubic Smoothing Splines under Dependent Data, *Biometrics*, Vol. **67**, Issue 3, p. 871-875.

[22] Nummi T., Pan J. and Mesue N. (2013). Testing linearity in semiparametric regression models, *Statistics and Its Interface*, Vol. **6**, 3–8.

[23] Nummi T., Hakanen T., Lipiäinen L., Harjunmaa U., Salo M., Saha M.-T. and Vuorela N. (2014). A trajectory analysis of body mass index for Finnish children, *Journal of Applied Statistics*, **41** (7), 1422-1435.

[24] Poortema K. (1984). On the statistical analysis of growth. *PhD Thesis, Groningen University.*

[25] Ruppert D., Wand M.P. and Carrol R.J. (2005). *Semiparametric Regression*, Cambridge University Press, New York, USA.

[26] Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical analysis of finite mixture distribution*, Wiley, UK.

[27] Verbeke, G. and Lesaffre E. (1996). A Linear Mixed-Effects Model with Heterogeneity in the Random-Effects Population, *Journal of the American Statistical Association*, Vol. **91**, No. 433, 217–221.

[28] Vuorela N. (2011). Body Mass Index, Overweight and Obesity Among Children in Finland - A Retrospective Epidemilogical Study in Pirkanmaa District Spanning Over Four Decades, *Acta Universitatis Tamperensis.*

[29] Ye, H. and Pan, J. (2006). Modelling covariance structures in generalized estimating equations for longitudinal data, *Biometrika*, **93**, 927–941.