Some Models and Their Extensions for Longitudinal Analyses

Tapio Nummi[1*][0000-0001-8682-6951]

[1] Tampere University, Faculty of Information Technology and Communication Sciences, Finland
tapio.nummi@tuni.fi

Abstract. In this article, I present some of my statistical research in the field of longitudinal data analysis along with applications of these methods to real data sets. The aim is not to cover the whole field; rather, the perspective is based on my own personal preferences. The presented methods are mainly based on growth curve and mixture regression models and their extensions, where the focus is on continuous longitudinal data. In addition, an example of the analysis of extensive register data for categorical longitudinal data is presented. Applications range from forestry and health sciences to social sciences.

1 Introduction

Longitudinal studies play an important role in many fields of science. The defining feature of these studies is that measurements of the same individual are taken repeatedly over time. The primary goal is to characterize the change in response over time as well as the factors that influence the change. Special statistical methods which address intra-individual correlation and inter-individual variation are needed. Fortunately, many statistical analysis tools developed for clustered data (e.g. mixed, multilevel, and mixture models) also apply to longitudinal data, since longitudinal data can be seen as a special case of clustered data. Here, these methods are divided into three main categories:

1. Regression and multivariate techniques,
2. Methods based on finite mixtures, and
3. Clustering techniques for categorical longitudinal data.

The main aim of this article is to briefly present some of these techniques with interesting real data applications. The purpose is not to give an overview of the topics. Instead, the perspective is based on my own research in these areas.

* Corresponding author: Tapio Nummi (Email: tapio.nummi@tuni.fi)

2 Regression and Multivariate Techniques

2.1 The Growth Curve Model

Perhaps the most important of the early models in this area is the generalized multivariate analysis of variance model (GMANOVA), which is often called the growth curve model (GCM). This model was first presented by Potthoff and Roy (1964). GCM is particularly useful in balanced experimental study designs where there are no missing data. This model can be presented as follows

$$\mathbf{Y} = \mathbf{TBA}' + \mathbf{E}\,,$$

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ is a matrix of n response vectors, $\mathbf{T}$ is the within individual design (model) matrix, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_m)$ is a matrix of growth curve parameters, $\mathbf{A}$ is the between individual design matrix and $\mathbf{E}$ is a matrix of random errors, where columns are independently normally distributed as $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, $i = 1, \ldots, n$.

Closed-form formulas for the estimation and testing of growth curve parameters $\mathbf{B}$ can be obtained using the Maximum Likelihood method (under unknown positive definite $\mathbf{\Sigma}$). Some basic results and model extensions are nicely summarized in the review papers by von Rosen (1991) and Zezula and Klein (2011).

3 Some Extensions of the Growth Curve Model

Various aspects of analysis under GCM are presented in the series of articles by myself and my co-authors. Some practical computational aspects are considered in Nummi (1989), a method for prediction is presented in Liski and Nummi (1990), an analysis under missing data with the EM algorithm is investigated in Liski and Nummi (1991), model selection for mean and covariance structure is considered in Nummi (1992), prediction and inverse estimation is investigated in Liski and Nummi (1995a), and a method of covariable selection for model parameter estimation is presented in Wang et al. (1999).

3.1 Random Effects Growth Curve Model

Perhaps one of the most important extensions is the so-called Random effects growth curve model. This model can be written as

$$\mathbf{Y} = \mathbf{TBA}' + \mathbf{T_c\Lambda} + \mathbf{E},$$

where $\mathbf{T}_c$ is given as $\mathbf{T} = (\mathbf{T}_c, \mathbf{T}_{\bar{c}})$ and $\mathbf{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n)$ is a matrix of random effects. Here we take $\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$ independent of $\boldsymbol{\lambda}_i \sim \boldsymbol{N}(\mathbf{0}, \mathbf{D})$, $\forall\ i = 1, \ldots, n$, where $\mathbf{D}$ is a positive define covariance matrix.

In the article by Nummi (1997), several topics were considered: model parameter estimation and hypothesis testing, estimation under parsimonious covariance structures (e.g. AR(1)) for random errors, estimation under incomplete data using the EM algorithm, and an extension to multivariate growth curves. The multivariate extension was further studied by Nummi and Möttönen (2000), where they studied ML and REML estimation and testing in this context. For example, it was shown that under certain situations, estimated variances of growth curve parameters are greater for REML. The basic Random effects GCM was further extended and applied in small area estimation by Ngarue et al. (2017).

3.2 Measurement Errors

In some cases also in an experimental situation, a measurement error may occur. This is especially the case when the planned measurement time is fixed in advance, but the exact attained measurement time does not match the planned time. An appropriate frame for this kind of analysis is found under Berkson-type measurement errors (see Berkson, J. (1950)). The basic (Berkson) model for the observations (y, x^*) is

$$y = g(x) + \epsilon$$
$$x = x^* + u$$

where ϵ and u are independent random variables with $E(\epsilon) = E(u) = 0$. Here, the exact value of the explanatory variable x is not directly observed, but instead another quantity, the planned measurement time $x^* = x - u$ is utilized. Actually, this form of measurement error is quite common in experimental situations where the values of the predictor variable is controlled by the experimenter (e.g. in agricultural or medical studies). For GCM, the Berkson type of measurement errors is studied in Nummi (2000) and later extended and applied to forest harvesting in Nummi and Möttönen (2004).

3.2.1 Example: Forest Harvesting

The forestry harvesting technique in the Nordic countries converts tree stems into smaller logs at the point of harvest. Modern harvesters are equipped with computer systems capable of continuously measuring the length and diameter of the stem and also predicting the profile of an unknown stem section. The harvester feeds the tree top-first through the measuring and delimbing device for a given length, then the computer predicts the rest of the stem profile and calculates the optimal cross-cutting points for the whole stem (see e.g. Uusitalo et al. (2006)). In forestry, stem curve models are often presented for relative heights (e.g. Laasasenaho, 1982; Kozak, 1988). However, height is unknown for a harvester, and therefore these relative mean curve models are quite difficult to apply in practice. Low degree polynomial models were tested, e.g. in Liski and Nummi (1995b, 1996a, b).

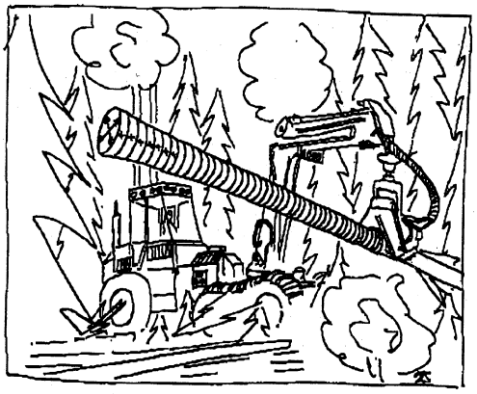

Fig. 1. A forest harvester at work. In the figure, the harvester has cut down a tree and started pruning the branches. At the same time, the stem diameter and length are measured and the measurements are transferred to the harvester's computer. The harvester is now at the first possible cutting point, at which the decision must be made as to whether to cut at that point or at another point along the stem. (Source: Ponsse company, Finland)

Assume now that x is a sum of sub-intervals $\delta_i^* = \delta_i + \xi_i$

$$x = \sum \delta_i^* = \sum \delta_i + \sum \xi_i = x^* + u$$

where random error u is a sum random terms $u = \sum \xi_i$, where ξ_i are independent with $E(\xi_i) = 0$ and $\mathrm{Var}(\xi_i) = \sigma^2_{\xi_i}$. Random errors u are now dependent and the variance σ_u^2 increases with x^*. A special model for the covariance structure Var(**y**) is now needed.

The general least squares methods provide unbiased parameter estimates only in the most simple first-degree model $g(x^*) = \beta_0 + \beta_1 x^*$ For more complex models for $g(x^*)$, the least squares estimates of the model parameters are biased. However, as shown in Nummi and Möttönen (2004), predictions may still be unbiased for low-degree polynomial models. Estimation and prediction for an extended Berkson model (with dependent measurement errors) are considered in Nummi and Möttönen (2004).

3.3 Spline Growth Model

A more general formulation of the basic GMANOVA can be written as

$$\mathbf{Y} = \mathbf{G}\mathbf{A}' + \mathbf{E}$$

where $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)$ is a matrix of smooth mean curves (Spline Growth Model, SGM; Nummi and Koskela, 2008; Nummi and Mesue, 2013; Mesue and Nummi, 2013; Nummi et al., 2017). Here we assume that

$$\mathbf{\Sigma} = \sigma^2 \mathbf{R}(\boldsymbol{\theta})$$

where **R** takes a certain kind of parsimonious covariance structure with covariance parameters $\boldsymbol{\theta}$. A smooth solution for **G** can be obtained by minimizing the penalized least squares criterion (see Nummi and Koskela, 2008)

$$\text{PLS} = \text{tr}\left[\left(\mathbf{Y} - \dot{\mathbf{G}}\right)\mathbf{H}\left(\mathbf{Y} - \dot{\mathbf{G}}\right) + \alpha\dot{\mathbf{G}}'\mathbf{K}\dot{\mathbf{G}}\right],$$

where α (> 0) is a fixed smoothing parameter, $\dot{\mathbf{G}} = \mathbf{G}\mathbf{A}'$, $\mathbf{H} = \mathbf{R}^{-1}$ and the roughness matrix **K** (from $RP = \int g''^2$) is defined as $\mathbf{K} = \boldsymbol{\nabla}\Delta^{-1}\boldsymbol{\nabla}'$ where $\boldsymbol{\nabla}$ and Δ are banded $q \times (q-2)$ and $(q-2) \times (q-2)$ matrices defined as (non-zero elements)

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

and

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

where $h_j = t_{j+1} - t_j$, $j = 1, 2, \ldots, (q-1)$ and $l = 1, 2, \ldots, (q-2)$ (see e.g. Green and Silverman (1994)).

As shown in Nummi and Koskela (2008), the minimizer is easily seen by rewriting the PLS-function in a slightly different form. Then given α and **H**, the spline estimator becomes

$$\widetilde{\mathbf{G}} = (\mathbf{H} + \alpha\mathbf{K})^{-1}\mathbf{H}\mathbf{Y}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1},$$

where the fitted growth curves $\widetilde{\mathbf{G}}$ are natural cubic smoothing splines. It is further easily seen that if **K = KH** (or **KR = K**), the spline estimator simplifies as

$$\widehat{\mathbf{G}} = \mathbf{S}\mathbf{Y}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1},$$

where the so-called smoother matrix is denoted as $\mathbf{S} = (\mathbf{I} + \alpha\mathbf{K})^{-1}$. Note that this is an important simplification, since estimates $\widehat{\mathbf{G}}$ are now simple linear functions of the observations (α fixed). In a sense, this can be compared to the results of linear models, where OLSE = BLUE. It is quite easy to see that certain important structures (e.g. uniform, random effects, etc.) used for the analysis of longitudinal data meet the simplifying condition. The smoothing parameter α can then be chosen using the Generalized Cross-Validation (GCV) criteria, for example.

3.3.1 Testing with an Application for Behavioral Cardiology

Note that the smoother matrix **S** is not a projection matrix and therefore certain results developed for linear models are not directly applicable for SGM. Our approach is to approximate **S** with the following decomposition (Nummi and Mesue, 2013; Mesue and Nummi, 2013; Nummi et al., 2017 and Nummi et al., 2018)

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

where **M** is the matrix of q orthogonal eigenvectors of **K** and **Λ** is a diagonal matrix of corresponding eigenvalues obtained from the Spectral decomposition. Here we assume that eigenvectors are ordered according to eigenvalues $\gamma = 1/(1 + \alpha\lambda)$ of **S**, where λ is an eigenvalue of **K**. Note that the sequence of eigenvectors $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_q$

increases in complexity. The first two eigenvectors $\mathbf{m}_1$, $\mathbf{m}_2$ span a straight line model and the corresponding eigenvalues are 1.

Using the approximation $\mathbf{S} \approx \mathbf{M_c M'_c}$ the set of fitted curves with SGM are

$$\bar{\mathbf{Y}} = \mathbf{M_c M'_c Y A}(\mathbf{A'A})^{-1}\mathbf{A'} = \mathbf{M_c}\hat{\mathbf{\Omega}}\mathbf{A'},$$

where $\mathbf{M_c}$ is a matrix of c first eigenvectors of $\mathbf{S}$ that can be chosen using GCV criteria, and $\hat{\mathbf{\Omega}} = \mathbf{M'_c Y A}(\mathbf{A'A})^{-1}$. All the relevant information for testing is now in $\hat{\mathbf{\Omega}}$, which can be seen to be unbiased estimates of the parameters of the growth curve model

$$E(\mathbf{Y}) = \mathbf{M}_c\mathbf{\Omega A'}.$$

Testing can be based on the linear hypothesis of the form H_0: $\mathbf{C\Omega D = O}$– where $\mathbf{C}$ and $\mathbf{D}$ are appropriate given matrices. It is easy to construct an F-test for this H_0. It is further easy to show that for some important special cases, the distribution of this F-test does not depend on the estimated covariance structure.

Example. Tampere Ambulatory Hypertension Study: In these data, 95 men (aged 35-45 years with the same ethnic and cultural background) were selected and their body functions were accurately monitored for one day (see Nummi et al., 2017). Inclusion criteria: healthy according to conventional health criteria and not on medication. For this study, we investigated the hourly means of systolic blood pressure (SBP), diastolic blood pressure (DBP) and heart rate (HR). The participants were classified before the experiment into two groups:

Group 1: Normotensive (NT), 33 participants, and
Group 2: Borderline hypertensive (BHT) and hypertensive (HT), 62 participants

In our example, we were especially interested in testing whether blood pressure variables behaved the same during the day in these two groups. In particular, by definition, there is a level difference in these two groups. If we define the roughness matrix as $\mathbf{K}_s = \mathbf{W}\otimes\mathbf{K}$, where $\mathbf{W}$ = diag (α_1, …, α_s) the multivariate uniform structure

$$\mathbf{R} = \left(\mathbf{I}_s \otimes \mathbf{1}_q\right)\mathbf{D}\left(\mathbf{I}_s \otimes \mathbf{1}_q\right)' + \mathbf{I}_{qs}$$

where $\mathbf{D}$ is a covariance matrix, satisfies the multivariate version of the simplifying condition $\mathbf{RK}_s = \mathbf{K}_s$ and the unweighted spline estimator becomes (Mesue and Nummi, 2013)

$$\hat{\mathbf{G}} = \left(\boldsymbol{I}_{qs} + \boldsymbol{W} \otimes \boldsymbol{K}\right)^{-1}\mathbf{Y A}(\mathbf{A'A})^{-1}$$

$$= \begin{pmatrix} \mathbf{S}(\alpha_1) & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{S}(\alpha_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{S}(\alpha_s) \end{pmatrix}\mathbf{Y A}(\mathbf{A'A})^{-1} \qquad (1)$$

where $\mathbf{S}(\alpha_j) = (\mathbf{I}_q + \alpha_j \mathbf{K})^{-1}$, where α_j is a smoothing constant for j = 1, …, s. A straightforward generalization of the earlier considerations gives us an estimator

$$\widehat{\mathbf{\Omega}} = \mathbf{M}'_{\bullet}\mathbf{Y}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1},$$

where $\boldsymbol{M}_{\bullet} = \mathrm{diag}(\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_s)$ and the corresponding multivariate growth curve model is

$$\mathbf{Y} = \mathbf{M}_{\bullet}\mathbf{\Omega}\mathbf{A}'.$$

Testing (see Nummi et al. 2017) can be based on the linear hypothesis H_0: $\mathbf{C\Omega D} = \mathbf{0}$, where $\mathbf{C}$ and $\mathbf{D}$ are known $v \times c$ and $m \times g$ matrices with ranks v and g, respectively, with

$$F = \frac{Q_*/vg}{\hat{\sigma}^2} \sim F\,[vg, \quad n(sq - c_{\mathrm{tot}})],$$

where $c_{\mathrm{tot}} = c_1 + \cdots + c_s$ and

$$Q_* = \mathrm{tr}[\mathbf{D}'(\mathbf{A}'\mathbf{A})^{-1}\mathbf{D}]^{-1}\left[\mathbf{C}\widehat{\mathbf{\Omega}}\mathbf{D}\right]'[\mathbf{C}\mathbf{M}'_{\bullet}\mathbf{R}\mathbf{M}_{\bullet}\mathbf{C}']^{-1}\left[\mathbf{C}\widehat{\mathbf{\Omega}}\mathbf{D}\right]$$

and

$$\hat{\sigma}^2 = \sum_{l=1}^{s} \frac{1}{n(q - c_l)} \mathrm{tr}\mathbf{Y}'_l\left(\mathbf{I}_q - \mathbf{P}_l\right)\mathbf{Y}_l .$$

Often $\mathbf{R}$ may not be known and need to be estimated. In this case, the distribution of F is only approximate. However, with the multivariate uniform covariance model, when investigating the progression only we can take $\mathbf{C} = [\mathbf{I}_s \otimes (\mathbf{0}, \mathbf{I})]$. It can then be shown that the test statistics have an exact F-distribution.

For the example data, c_1 = 12 (SBP), c_2 = 10 (DPB) and c_3 = 12 (HR). To test if the progression is the same in both groups, we attained

$$F = \frac{1811.041/31}{66.26201} = 0.88166,$$

which gives the P-value $P(F_{31,2470} \geq 0.88166) \approx 0.654967$. Therefore, the null hypothesis of equal progression for each of the variables in these groups is not rejected. The fitted mean curves are shown in figure 2.

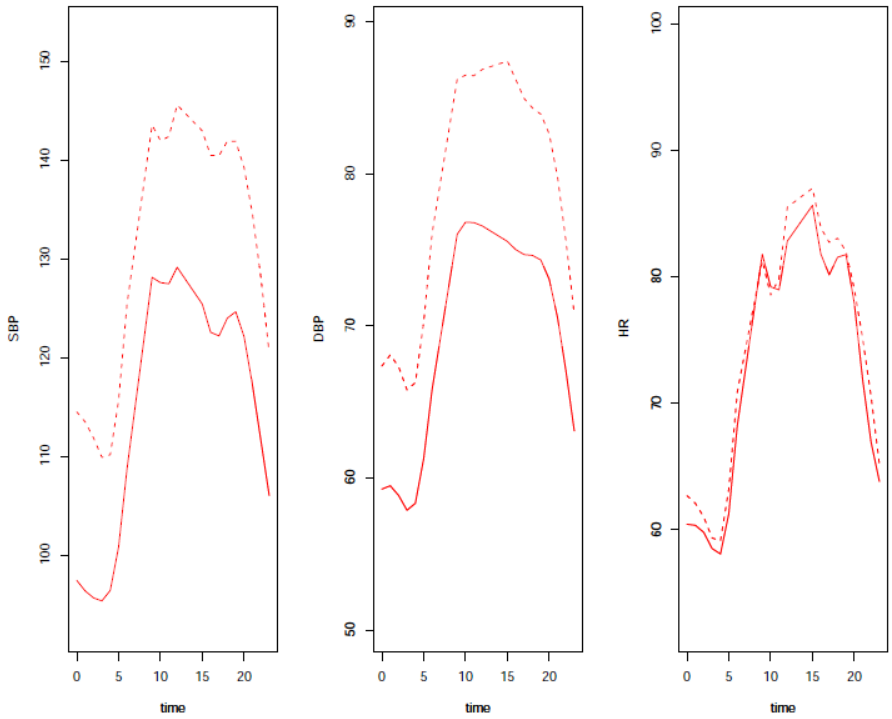

Fig. 2. Fitted mean curves for systolic blood pressure (SBP), diastolic blood pressure (DBP) and heart rate (HR) during the test day. Solid line is for Group 1 (normotensive) and dotted line for Group 2 (Borderline hypertensive and hypertensive) (Source: created by the authors).

4 Models Based on Finite Mixtures

4.1 Introduction

Denote random vectors of longitudinal measurements as $\mathbf{y}_i = \left(y_{i1},\ y_{i2}, \ldots, y_{ip_i}\right)'$ and the marginal probability density of $\mathbf{y}_i$ with possible time-dependent covariates $\mathbf{X}_i$ as $f(\mathbf{y}_i \mid \mathbf{X}_i)$ for $i = 1, \ldots, N$. It is assumed that $f(\mathbf{y}_i \mid \mathbf{X}_i)$ follows a mixture of K densities

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

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

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

where $\boldsymbol{\mu}_{ik}(\boldsymbol{\theta}_k, \boldsymbol{X}_i)$ is a function of covariates $\mathbf{X}_i$ with parameters $\boldsymbol{\theta}_k$ and $\boldsymbol{\Sigma}_{ik}(\boldsymbol{\sigma}_k)$ is a variance-covariance matrix within the *k*th cluster, involving a vector of unique covariance parameters $\boldsymbol{\sigma}_k$. In the most general case, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the unstructured mean and covariance matrices. However, often some more parsimonious structures are imposed either on $\boldsymbol{\mu}_k$, or $\boldsymbol{\Sigma}_k$ or on both.
In this article, we focus on the normal GLM case (so-called trajectory analysis; see e.g. Nagin 1999, 2005). It is then simply assumed that

$$\boldsymbol{\mu}_k = \mathbf{X}_i\boldsymbol{\theta}_k \quad \text{with } \boldsymbol{\Sigma}_k = \sigma_k^2\mathbf{I}$$

Note that this conditional independence assumption does not mean independence over the whole sample. One important aim is to identify and estimate the possible subpopulations as well as possible. Therefore, often a natural interpretation of the identified groups is emphasized jointly with model selection and statistical fit criteria.

4.2 Data Analysis: Analysis of Drinking Profiles

The Northern Swedish Cohort Study covers all pupils who in 1981 attended the final year of compulsory school (at age 16) in Luleå. The number of participants in all follow-up surveys (1983, 1986, 1995 and 2007) was 1,005. For this study, we used the alcohol consumption (converted to absolute alcohol in centiliters) of male participants at the age of 16, 18, 21, 35 and 42 years (for more details, see Virtanen et al. (2015)).

Table 1. Results of the fits of the mixture models with $k = 1, 2, \ldots, 7$ and $\lambda \in [-2, 2]$

K	$\hat{\lambda}$	*L*	BIC	AIC
1	0.16	-16367.23	32770.40	32744.47
2	0.04	-16065.51	32206.09	32151.02
3	0.00	-15961.67	32037.55	31953.35
4	0.08	-15885.47	31924.28	31810.94
5	0.08	-15852.29	31897.07	31754.58
6	0.04	-15818.75	31869.12	31697.49
7	0.04	-15802.01	31874.97	31674.20

Source: Authors' own processing.

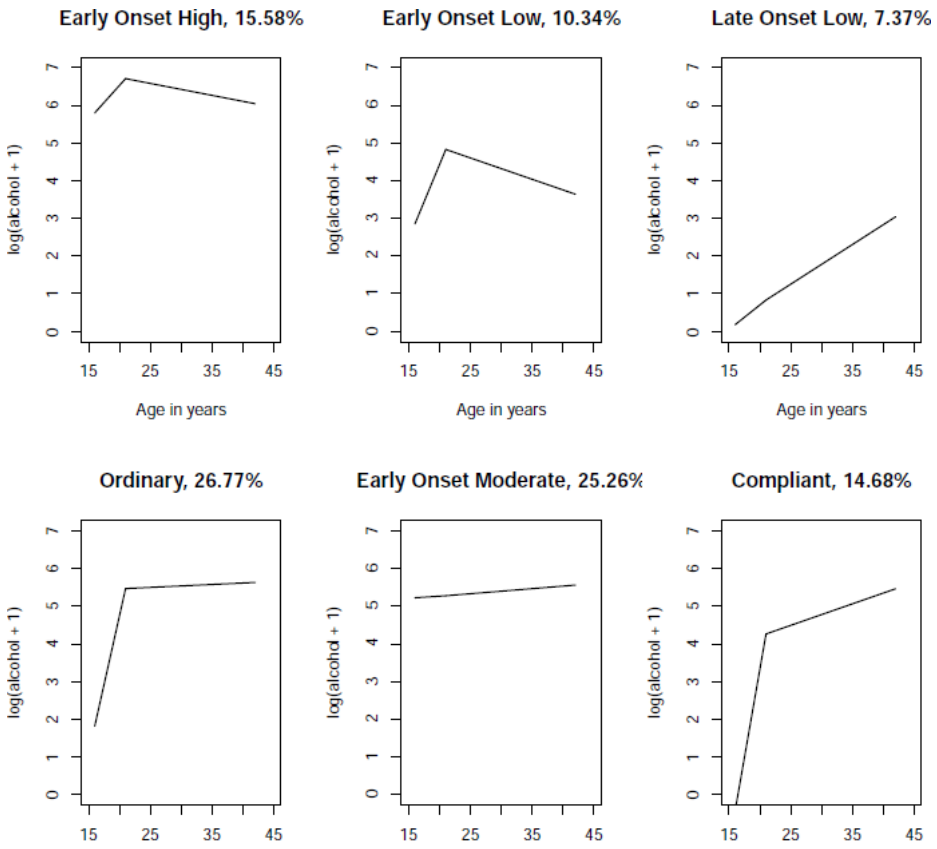

Fig. 3. Fitted trajectory curves for alcohol consumption of males (Source: created by the authors).

The distribution of alcohol consumption is highly skewed, but there is a relatively high probability of zero observations (semicontinuous data). This kind of data is quite common in many areas (cf. zero inflation). One possible solution to the problem is to apply mixture modeling, where one component of the mixture functions is degenerated near to zero. The advantage is that the mixture approach allows additional heterogeneity by avoiding a sharp dichotomy between zero and near-zero observations. A brief summary of the methods for semicontinuous and zero-inflated data is presented in Min and Agresti (2005).

Here the so-called "broken stick" model was applied as the basic model for transformed observations:

$$Y^{(\lambda)} = \beta_0 + \beta_1 t + \beta_2 (t - K_1)_+ + \epsilon \,,$$

K_1 = 21, $(.)_+$ equals (.) if (.) $\geq$ 0 and 0 otherwise. The model consists of two combined straight lines at the age of 21.

The number of clusters *K* and the transformation parameter of the Box-Cox transformation was jointly estimated. Depending on the criterion, a different *K* is selected.

With AIC, the minimum is obtained for $K = 7$ and with BIC, the minimum is obtained for $K = 6$ (Table 1). Our choice was $K = 6$, which was also in line with the earlier studies with these data (e.g. Virtanen et al. (2015)). The estimate of the transformation parameter (Box-Cox) $\hat{\lambda} = 0.04$ suggested the log transformation.

The estimated mixture proportions are $\pi_1 = 0.1558$; $\pi_2 = 0.1034$; $\pi_3 = 0.0737$; $\pi_4 = 0.2677$; $\pi_5 = 0.2526$ and $\pi_7 = 0.1468$. Group 3 is the zero or near-zero cluster. Interestingly, those who had a high consumption level at the earlier ages tended to maintain a high consumption level also at the later ages.

4.3 Extension: Semiparametric Mean Model

The set of explanatory variables in $\mathbf{X}_i$ is divided into the parametric part $\mathbf{U}_i$ and the non-parametric part $\mathbf{t}_i$, where $\mathbf{t}_i$ is the vector of measuring times $t_1, \ldots, t_{p_i}$. For the ith individual within the kth cluster, we assume the semiparametric model

$$\mathbf{y}_{ik} = \mathbf{g}_{ik}(\boldsymbol{t}_i) + \mathbf{U}_i\mathbf{b}_k + \boldsymbol{\epsilon}_i \,,$$

where $g_{ik}(\mathbf{t}_i)$ is a smooth vector of twice differentiable functions evaluated at time points $\mathbf{t}_i$, $\mathbf{U}_i$ is a matrix of h covariates (constant term not included), $\mathbf{b}_k$ is a parameter vector to be estimated and Var $(\boldsymbol{\epsilon}_i) = \sigma_k^2\mathbf{I}_i$.

When using the EM algorithm, the estimation problem can be seen as a missing data problem, where $\mathbf{y}_i$ are observed but "group indicators" $\mathbf{z}'_i$ are missing. We denote

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

where $z_{ik} = 1$ if $\mathbf{y}_i$ stemmed from the kth component; otherwise, $z_{ik} = 0$. The vectors $\mathbf{z}_1, \ldots, \mathbf{z}_N$ can now be seen as realized values of random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ from the multinomial distribution. The complete-data, joint penalized log-likelihood function is (see Nummi et al. (2018) for details)

$$l_c(\phi) = \sum_{i=1}^{N}\sum_{k=1}^{K}\left\{z_{ik}[\log(\pi_k) + \log(f_k)] - \frac{\alpha_k}{2N}\mathbf{g}'_k\mathbf{K}\mathbf{g}_k\right\} .$$

The E step is to calculate

$$E\left(Z_{ik}\middle|\hat{\boldsymbol{\phi}}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right) = \frac{\hat{\pi}_k f_k\left(\mathbf{y}_i\middle|\mathbf{X}_i, \hat{\boldsymbol{\xi}}_k\right)}{\sum_{l=1}^{K}\hat{\pi}_l f_l\left(\mathbf{y}_i\middle|\mathbf{X}_i, \hat{\boldsymbol{\xi}}_l\right)} = \hat{z}_{ik}$$

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

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

Table 2. Results of the sequence analysis of combined employee-employer data.

Main activity	Combined (%)	Study group (%)	Reference (%)
Employed	52	49	55
Disability pension	12	13	11
Retired	12	10	13
Part-time retired	10	9	11
Unemployed	9	11	7
Unemp. pension	5	8	3

Source: Authors' own processing.

is maximized. These two steps are iterated until convergence. The method gives closed-form formulas for $\mathbf{g}_k$ and $\mathbf{b}_k$ with estimates for $\boldsymbol{\pi}_k$. Here each of the k group can be smoothed independently, and thus this provides a very flexible model within each of the k clusters. In Nummi et al. (2018), a technique providing an approximate solution is also introduced. This makes semiparametric mixture analysis possible in general statistical software developed for mixture regression.

5 Clustering Techniques for Categorical Longitudinal Data: Factory Downsizing

Example of sequence analysis is based on Statistics Finland's combined employee-employer data (FLEED), which includes data for all 15–70 year old of those who lived in Finland in 1988–2014. For research purposes a random sample of the size of one–third was taken. The starting point of the study group is those enterprices that reduced more than 30 % of staff or were dismissed in the year 2005. The actual study group taken then consisted a sample of 7,730 people (aged 45-60) who lost their job in 2005 (followed until 2014). A reference group of matched (Propensity score) 7,844 people from the same register who did not lose their job in 2005 was also taken. Since the data are categorical (employment status), so-called sequence analysis was applied to the combined data. Sequence analysis was performed with R software using the Weighted Cluster library Studer (2013). The number of clusters was evaluated using the Average Silhouette Width. For further details on the data, methods and results, we refer to Kurvinen et al. (2018).

Finally, six clusters were identified that were named according to the main activity prevailing in the group. The results are presented in Table 2. It is observed that in the study group about half of the sample still continued in employment. It can be seen as an indication of an effective labor market policy in Finland. However, there is clearly an elevated risk (compared to those who continued as employed) for those who lost their job entering the unemployment group and the unemployment pension group even after controlling for covariates gender, age, sector of employment, education, socio-economic status, type of residence area, employment and unemployment in 2004 and sickness allowance paid in 2003–2004.

Although sequence analysis is mainly descriptive in nature, a suitable experimental study design can also provide a framework for further estimation and testing of important statistical quantities.

References

1. Berkson, J.: *Are there two regressions*? J. Am. Statist. Ass., 45, 164?180 (1950).
2. Green, P. and Silverman, B:. *Nonparametric regression and generalized linear models*. A roughness penalty approach. Monographs on Statistics and Applied Probability, 58. Boca Raton, FL: Chapman Hall/CRC (1994).
3. Kozak, A.: *A variable-exponent taper equation*. Can. J. For. Res., 18, 1363-1368 (1988).
4. Kurvinen, A., Jolkkonen, A.. Koistinen, P., Lipiäinen, L., Nummi, T. ja Virtanen, P.: *Työn menetys työuran loppuvaiheessa - Tutkimus 45–60-vuotiaana rakennemuutoksessa työnsä menettäneiden työllisyysurista ja riskistä päätyä työttömäksi tai työvoiman ulkopuolelle*, 83, 5-6 (2018). (in Finnish with English abstract).
5. Laasasenaho, J.: *Taper curve and volume functions for pine, spruce and birch.* Communicationes Instituti Forestalis Fenniae, 108 (1982). http://urn.fi/URN:ISBN:951-40-0589-9
6. Liski E. P. and Nummi T.: *Prediction in Growth Curve Models Using the EM Algorithm*, Computational Statistics & Data Analysis, Vol 10, No. 2, 99-108 (1990).
7. Liski E. P. and Nummi T.: *Missing Data under the GMANOVA Model*. The Frontiers of Statistical Scientific Theory - Industrial Applications. (Volume II of the Proceedings of ICOSCO - I, The First International Conference on Statistical Computing, ed. Özturk and van der Meulen), American Science Press Inc, Columbus, Ohio, 391–404 (1991).
8. Liski E. P. and Nummi T.: *Prediction and Inverse Estimation in Repeated-Measures Models*, Journal of Statistical Planning and Inference, 47, pp. 141-151 (1995).
9. Liski, E.P. and Nummi, T.: *Prediction of tree stems to improve efficiency in automatized harvesting of forests*. Scand. J. Stat. 22(2), 255–269 (1995).
10. Liski, E.P., Nummi, T.: *Prediction in repeated-measures models with engineering applications*. Technometrics 38(1), 25–36 (1996). https://doi.org/10.1080/00401706.1996.10484413
11. Liski E. P. and Nummi T.: *The Marking for Bucking under Uncertainty in Automatic Harvesting of Forest*, International Journal of Production Economics, 46-47, pp. 373-385 (1996).
12. Mesue, N. and Nummi, T.: *Testing of growth curves using smoothing spline: a multivariate approach.* Proceedings of the 28th International Workshop on Statistical Modelling. Muggeo, VMR., Capusi, V., Boscaino, G. and Lovison, G. (eds.) Palermo, Italia: Statistical Modelling Society, p. 281–288 (2013).
13. Min, Y. and Agresti, A.: *Random effect models for repeated measures of zero-inflated count data*. Statistical Modelling, 5, 1?19 (2005).
14. Nagin, D.S.: *Analyzing developmental trajectories: a semiparametric, group-based approach.* Psychol. Methods 4(2), 139–157 (1999). https://doi.org/10.1037/1082-989X.4.2.139
15. Nagin, D. S.: *Group-Based Modeling of Development.* Harvard University Press, Cambridge, MA (2005).

16. Ngaruye I., Nzabanita J., von Rosen D. & Singull M.: *Small area estimation under a multivariate linear model for repeated measures data.* Communications in Statistics - Theory and Methods, 46:21, 10835-10850 (2017). DOI: 10.1080/03610926.2016.1248784
17. Nummi T.: *APL as a Tool for Computations in Growth Studies*, Adamm Kertesz, Lynne C. Shaw (Eds.). APL Quote-Quad (Conference Proceedings APL89 - APL as a Tool of Thought, August 7-10, New York City),Volume 19, Number 4, 293–298 (1989).
18. Nummi, T.: *On model selection under the GMANOVA model*, Statistical Modelling (P.G.M van der Heyden, W. Jansen, B. Francis and G.U.H. Seeber, eds.), Elsevier Science Publishers B.V., Amsterdam, 283-292 (1992).
19. Nummi T.: *Estimation in a random effects growth curve model*. J. Appl. Stat. 24(2), 157–168 (1997). https://doi.org/10.1080/02664769723774
20. Nummi, T.: *Analysis of growth curves under measurement errors.* J. Appl. Stat. 27(2), 235–243 (2000). https://doi.org/10.1080/02664760021763
21. Nummi, T. and Koskela, L.: *Analysis of growth curve data using cubic smoothing splines*. J. Appl. Stat. 35(6), 681–691 (2008). https://doi.org/10.1080/02664760801923964
22. Nummi, T. and Mesue, N.: *Testing of growth curves with cubic smoothing splines*. In: Dasgupta, R. (ed.) *Advances in Growth Curve Models*. Springer Proceedings in Mathematics & Statistics, vol. 46, pp. 49–59. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6862-2_3
23. Nummi T. and Möttönen, J.: *On the analysis of multivariate growth curves.* Metrika, 52, 77–89 (2000). https://doi.org/10.1007/s001840000063
24. Nummi, T. and Möttönen, J.: *Estimation and prediction for low-degree polynomial models under measurement errors with an application to forest harvester.* J. R. Stat. Soc. Ser. C Appl. Stat. 53, 495–505 (2004). https://doi.org/10.1111/j.1467-9876.2004.05138.x
25. Nummi, T., Möttönen, J. and Tuomisto, M.T.: *Testing of multivariate spline growth model.* In: Chen, D.G., Jin, Z., Li, G., Li, Y., Liu, A., Zhao, Y. (eds.) New Advances in Statistics and Data Science, pp. 75–85. ICSA Book Series in Statistics. Springer, Cham (2017). https://doi.org/10. 1007/978-3-319-69416-0_5
26. Nummi T., Salonen J., Koskinen L. & Pan J: *A semiparametric mixture regression model for longitudinal data*, Journal of Statistical Theory and Practice, vol. 12:1, 12–22 (2018) https://doi.org/10.1080/15598608. 2017.1298062
27. Potthoff, R.F. and Roy, S.N.: *A generalized multivariate analysis of variance model useful especially for growth curve problems*. Biometrika 51(3–4), 313–326 (1964). https://doi.org/10.1093/biomet/51.3-4.313
28. Studer, M.:*Weighted Cluster Library Manual:Apractical guide to creating typologies of trajectories in the social sciences with* R. LIVES Working Papers, 24, 2013. http://dx.doi.org/10.12682/lives.2296-1658.2013.24
29. Virtanen, P., Nummi, T., Lintonen, T.,Westerlund, H., Hägglöf, B., Hammarström, A.: *Mental health in adolescence as determinant of alcohol consumption trajectories in the Northern Swedish Cohort*. Int. J. Public Health 60, 335–342 (2015). https://doi.org/10.1007/s00038-015-0651-5
30. Uusitalo, J., Puustelli, A, Kivinen, V.-P., Nummi, T.&Sinha, B.K.: *Bayesian estimation of diameter distribution during harvesting.* Silva Fennica 40(4): 663–671 (2006).
31. Wang S.G., Liski E. P. and Nummi T.: *Two-way selection of covariables in multivariate growth curve models*, Linear Algebra and Its Applications, 289, 333–342 (1999).
32. von Rosen D. (1991): *The growth curve model: a review*, Communications in Statistics – Theory and Methods, 20:9, 2791-2822, https://doi.org/10.1080/03610929108830668
33. Zezula I. and Klein D. (2011): *Overview of Recent Results in Growth-curve-type Multivariate Linear Models*, Mathematica, 50, 2, 137–146.