LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Felix Heinzl & Gerhard Tutz

Clustering in linear mixed models with a group fused lasso penalty

Technical Report Number 123, 2012
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Clustering in Linear Mixed Models with a Group Fused Lasso Penalty

Felix Heinzl & Gerhard Tutz

Department of Statistics

Ludwig-Maximilians-University Munich

SUMMARY: A method is proposed that aims at identifying clusters of individuals that show similar patterns when observed repeatedly. We consider linear mixed models which are widely used for the modeling of longitudinal data. In contrast to the classical assumption of a normal distribution for the random effects a finite mixture of normal distributions is assumed. Typically, the number of mixture components is unknown and has to be chosen, ideally by data driven tools. For this purpose an EM algorithm-based approach is considered that uses a penalized normal mixture as random effects distribution. The penalty term shrinks the pairwise distances of cluster centers based on the group lasso and the fused lasso method. The effect is that individuals with similar time trends are merged into the same cluster. The strength of regularization is determined by one penalization parameter. For finding the optimal penalization parameter a new model choice criterion is proposed.

KEY WORDS: *EM algorithm; fused lasso; group lasso; linear mixed models; longitudinal data*

1 Introduction

Linear mixed models (LMM) which were proposed by Laird and Ware (1982) are a common tool for the modeling of longitudinal data. The model can be written as

$$\boldsymbol{y}_i|\boldsymbol{b}_i \overset{ind.}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,\, \sigma^2\boldsymbol{I}_{n_i}) \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{y}_i$ contains the response values observed for subject i at observation times $t_{i1}, \ldots, t_{in_i}$. Here $\boldsymbol{I}_{n_i}$ is the identity matrix with dimension n_i. Population effects are included in the parameter $\boldsymbol{\beta}$ whereas $\boldsymbol{b}_i$ represents the individual-specific effects. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ denote the corresponding individual design matrices. All observations y_{ij} are normally distributed conditional on the random effects and are regarded as independent with the same variance σ^2. The classical assumption in (1) is a Gaussian distribution for the random effects, i.e. $\boldsymbol{b}_i$ i.i.d. $N(\boldsymbol{0}, \boldsymbol{D})$, see, for example, Verbeke and Molenberghs (2000) and Ruppert et al. (2003). While this choice is mathematically convenient, it often is questionable in applications for several reasons. The normal distribution is symmetric, unimodal and has light tails. Since the distributional assumption is made on unobserved quantities, it is typically hard to validate these properties based on estimates. Possible skewness and multimodality (arising, for example, from an unconsidered grouping structure in the data) may be masked when checking the normal distribution in terms of estimated random effects. In contrast to this homogeneity model the heterogeneity model introduced by Verbeke and Lesaffre (1996) is much more flexible. It assumes

$$\boldsymbol{b}_i \sim \sum_{h=1}^{N} \pi_h\, N(\boldsymbol{\mu}_h, \boldsymbol{D}), \tag{2}$$

where $\pi_1, \ldots, \pi_N$ are mixture weights. Several extensions and alternatives to this heterogeneity model have been proposed in the following. For example, Gaffney and Smyth (2003) used random effects regression mixtures in the context of curve clustering. Approaches for clustering functional data were proposed by James and Sugar (2003) and Liu and Yang (2009). By contrast Celeux et al. (2005), Ng et al. (2006) and Scharl et al. (2010) dealt with mixtures of linear mixed effects models. While Booth et al. (2008) proposed an extension of this concept by including the partition as parameter, De la Cruz-Mesía et al. (2008) generalized the approach to a mixture of non-linear hierarchical models. Villarroel et al. (2009) extended the heterogeneity model to allow for a multivariate response variable. In addition, a heteroscedastic normal mixture in the random effect distribution for multiple longitudinal markers was considered by Komárek et al. (2010) for linear mixed models and by Komárek and Komárková (2012) for generalized linear mixed models. However, in all these ap-

proaches the number of mixture components has to be chosen. A data driven choice of this number can be achieved by penalization of pairwise distances of cluster centers by a group fused lasso penalty term. In contrast to approaches that aim at penalizing the reparameterized mixture weights (Komárek and Lesaffre (2008) or Heinzl and Tutz (2011)) the "penalized heterogeneity approach" introduced here reduces the number of clusters by penalizing the cluster centers in the form

$$\sqrt{N \cdot q} \sum_{h<l} \|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|. \tag{3}$$

The idea of the penalty term is the following: If two cluster locations are very similar in terms of the Euclidean distance $\|\cdot\|$, these clusters should be fused. Therefore only the relevant clusters are expected to remain in the model. Fusion methods in regression modeling, but with quite differing penalty terms, have been proposed by Tibshirani et al. (2005). Penalty terms that include vectors, as is needed here, have been considered by Yuan and Lin (2006) but not in a fusion context. It should be noted that the factor $\sqrt{N \cdot q}$, where q denotes the dimension of random effects, is used for incorporating the number of parameters to estimate. For inference, we extend the traditional Expectation-Maximization (EM) algorithm (Dempster et al., 1977) used in the heterogeneity model of Verbeke and Molenberghs (2000) by adding the penalty term (3) multiplied with a penalty parameter to the logarithm of the complete but not fully observed likelihood (see Section 2.1). To find the optimal penalty parameter we introduce a new model choice criterion which is based on the concept of Braun et al. (2012) (see Section 2.2). The usefulness of our approach is demonstrated by two applications (see Section 3) and a simulation study (see Section 4).

It will be shown that our penalized heterogeneity approach is much more flexible than the conventional homogeneity model and allows to determine the number of clusters automatically. Regularization allows to identify the underlying clusters and cluster individuals in longitudinal studies.

2 Linear Mixed Models with Group Fused Lasso Penalty

2.1 Estimation

For the model introduced in Section 1 we give an EM algorithm which is based on derivations by McLachlan and Peel (2000) and McLachlan and Krishnan (1997) and is similar to the algorithm used by Verbeke and Molenberghs (2000) but includes the penalty term (3). Let the parameters be collected in $\boldsymbol{\xi} = (\boldsymbol{\pi}, \boldsymbol{\gamma})^T$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ comprises the mixture weights and $\boldsymbol{\gamma}$ is the vector containing all the remaining parameters $\boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, \boldsymbol{D}, \sigma^2$. In the following the order of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ is determined by the corresponding weights in decreasing order under the restrictions $\sum_{h=1}^N \pi_h = 1$ and $\sum_{h=1}^N \pi_h \boldsymbol{\mu}_h = \boldsymbol{0}$. The latter ensures $E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. The cluster membership of each individual can be described by latent variables $\boldsymbol{z}_i := (z_{i1}, \ldots, z_{iN})^T$ where $z_{ih} = 1$ if subject i belongs to cluster h and 0 otherwise. Marginalization over the random effects yields the complete model with observed data $\boldsymbol{y}_i$ as well as unobserved data $\boldsymbol{z}_i$:

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{z}_i &\overset{ind.}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i), && i = 1, \ldots, n, \\ \boldsymbol{z}_i &\overset{i.i.d.}{\sim} M(1, \boldsymbol{\pi}), && i = 1, \ldots, n, \end{aligned} \tag{4}$$

with $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}_{n_i}$ and $M(\cdot)$ representing the multinomial distribution. The likelihood function corresponding to (4) is given by

$$L(\boldsymbol{\xi}) = \prod_{i=1}^n \prod_{h=1}^N [\pi_h\, f_{ih}(\boldsymbol{y}_i; \boldsymbol{\gamma})]^{z_{ih}},$$

where $f_{ih}(\cdot)$ denotes the density function of $N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_h,\, \boldsymbol{V}_i)$. The penalized log-likelihood we propose is

$$l_P(\boldsymbol{\xi}) = \sum_{i=1}^n \sum_{h=1}^N z_{ih}[\log \pi_h + \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\gamma})] - \lambda \sqrt{N \cdot q} \sum_{h<l} \|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|, \tag{5}$$

where λ indicates the penalty parameter. Obviously for $\lambda = 0$ the penalization term drops out. We will use an EM algorithm procedure which alternates between taking the expectation of $l_P(\boldsymbol{\xi})$ over all unobserved z_{ih} in the E-step and maximization of the expected value in the M-step instead of maximizing the penalized incomplete likelihood function based only on the observed data directly. The steps have the following form.

E-step

Collecting all observed data in $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ the E-step we get

$$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)}) = E\left(l_P(\boldsymbol{\xi})|\boldsymbol{y}, \boldsymbol{\xi}^{(t)}\right) =$$

$$= \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}(\boldsymbol{\xi}^{(t)})[\log \pi_h + \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\gamma})] - \lambda\sqrt{N \cdot q}\sum_{h<l}\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|,$$

where $\pi_{ih}(\boldsymbol{\xi}^{(t)})$ is the probability at iteration t that subject i belongs to cluster h and is given by

$$\pi_{ih}(\boldsymbol{\xi}^{(t)}) = \frac{f_{ih}(\boldsymbol{y}_i; \boldsymbol{\gamma}^{(t)})\pi_h^{(t)}}{\sum_{l=1}^{N} f_{il}(\boldsymbol{y}_i; \boldsymbol{\gamma}^{(t)})\pi_l^{(t)}}.$$

M-step

For simplicity, in the following we write $\pi_{ih} := \pi_{ih}(\boldsymbol{\xi}^{(t)})$ but it should be noted that for the M-step it is essential that π_{ih} is fixed from the last iteration t because then one can use that $Q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)})$ is the sum of two components, $Q(\boldsymbol{\pi}|\boldsymbol{\xi}^{(t)})$ and $Q(\boldsymbol{\gamma}|\boldsymbol{\xi}^{(t)})$, and the optimization problem in the M-step can be separated into two parts: The maximization of

$$Q(\boldsymbol{\pi}|\boldsymbol{\xi}^{(t)}) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}\log \pi_h$$

with respect to $\boldsymbol{\pi}$ and the maximization of

$$Q(\boldsymbol{\gamma}|\boldsymbol{\xi}^{(t)}) = \sum_{i=1}^{n}\sum_{h=1}^{N} \pi_{ih}\ \log f_{ih}(\boldsymbol{y}_i; \boldsymbol{\gamma}) - \lambda\sqrt{N \cdot q}\sum_{h<l}\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_l\|$$

with respect to $\boldsymbol{\gamma}$. The first optimization problem yields

$$\pi_h = \frac{1}{n}\sum_{i=1}^{n}\pi_{ih}, \qquad h = 1, \ldots, N.$$

In the second part of the M-step one obtains the current state for $\boldsymbol{\gamma}$ by alternating between the maximization of $Q(\boldsymbol{\gamma}|\boldsymbol{\xi}^{(t)})$ with respect to $\boldsymbol{\beta}$, to $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ and to the variance parameters $\boldsymbol{D}$ and σ^2. Conditional on the current state of the other parameters the maximization of $\boldsymbol{\beta}$ results in

$$\boldsymbol{\beta} = \left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\left(\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{y}_i - \sum_{h=1}^{N}\pi_{ih}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\boldsymbol{\mu}_h\right)\right).$$

For the maximization of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ given $\boldsymbol{\beta}$ and the variance parameters as well as for the maximization of the variance parameters given $\boldsymbol{\beta}$ and $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ a numerical procedure like the Nelder-Mead method is necessary.

Choice of starting values

For EM algorithms it is essential how to choose the starting values because the (penalized) incomplete log-likelihood is ascending at each step and the algorithm can converge to a local maximum. Because in each M-step the fusion of clusters is investigated it is sensible to choose starting values for an agglomerative clustering method. Therefore each subject starts in its own cluster. Thus, in the beginning there are $N = n$ clusters with weights $\pi_h = 1/N$, $h = 1, \ldots, N$. As starting values for the cluster locations $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ we consider the predicted random effects $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ of the previously fitted LMM with Gaussian random effect distribution. This fit yields starting values for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{D}$, too. To reduce computation time it is sometimes advisable to choose $N < n$ if the number of individuals is high. Then one obtains starting values for the cluster centers by a k-means clustering of predicted random effects of the former fitted LMM. However, the algorithm starts with N clusters and successively merges clusters until there is no further ascent of the penalized incomplete log-likelihood. If two clusters centers $\boldsymbol{\mu}_h$ and $\boldsymbol{\mu}_l$ are fused only one of these parameters is kept and the other one is deleted with the effect that the number of clusters N is reduced by one. In general, our penalized heterogeneity approach can be seen as an agglomerative cluster analysis but based on a regression model. After convergence we get the cluster membership by the matrix of estimated π_{ih}. Individual i is assigned to that cluster h for which $\hat{\pi}_{ih}$ is maximal. Based on the weights of all clusters the prediction of the random effects has the form

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + (\boldsymbol{I}_q - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^T\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{Z}_i)\sum_{h=1}^{N}\hat{\pi}_{ih}\hat{\boldsymbol{\mu}}_h,$$

which can be shown by using derivations from Lindley and Smith (1972).

Implementation

All computations are implemented in C++ to allow for an efficient treatment of loop-intensive calculations and to reduce the typically slow convergence of the EM algorithm. They are made easily accessible by a wrapper function within an `R`-package which will be soon provided. All variables are standardized internally for calculations. For updating the cluster centers and the variance parameters we use an implementation of the Nelder-Mead algorithm in C++ (library ASA047) which was used by Papageorgiou and Hinde (2012) for similar tasks. For reflection, extension and con-

traction coefficients we choose the common settings 1.0, 2.0 and 0.5 respectively. See Nelder and Mead (1965) and O'Neill (1971) for more technical details of the algorithm. Note that for ensuring that the covariance matrix $\boldsymbol{D}$ is nonnegative-definite we parameterize the concerning variance parameters by the entries of a lower triangular matrix $\boldsymbol{L}$ according to the Cholesky decomposition $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^T$. Then $\boldsymbol{D}$ is nonnegative-definite for each $\boldsymbol{L}$ and positive-definite (and so invertible, too) if $\boldsymbol{L}$ is a matrix with exclusively nonzero diagonal entries (Lindstrom and Bates, 1988).

2.2 Model Choice: Predictive Cross-Validation

In general, optimal penalization parameters can be chosen by cross-validation or information criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC). In normal linear mixed models the AIC is not as straightforward as in normal linear models (compare Vaida and Blanchard (2005) and Greven and Kneib (2010)). For the penalized heterogeneity approach, the evaluation of the marginal or conditional AIC is even more complicated. Hence we prefer a cross-validation approach. Braun et al. (2012) introduced a new predictive cross-validation approach for model choice in linear mixed models with Gaussian distributed random effects that is based on the "mixed" cross-validation approach proposed by Marshall and Spiegelhalter (2003). An advantage of this approach is that in contrast to full cross-validation the model must be fitted only once which saves computing time. In general, each observed response value y_{obs} is compared to the corresponding predictive distribution, for example, by the continuous ranked probability score (CRPS)

$$CRPS(y_{obs}) = -\int_{-\infty}^{\infty} \left(P(Y_{obs} \leq r) - \mathbb{1}(y_{obs} \leq r)\right)^2 dr\,,$$

where P symbolizes the predictive distribution of the random variable Y_{obs}. If the predictive distribution is a normal distribution with estimated mean μ and estimated standard deviation σ, the continuous ranked probability score will take the form

$$CRPS(y_{obs}) = \hat{\sigma}\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{y_{obs}-\hat{\mu}}{\hat{\sigma}}\right) - \frac{y_{obs}-\hat{\mu}}{\hat{\sigma}}\left(2\Phi\left(\frac{y_{obs}-\hat{\mu}}{\hat{\sigma}}\right) - 1\right)\right]. \quad (6)$$

Here $\varphi(\cdot)$ denotes the density function and $\Phi(\cdot)$ the distribution function of the standard normal distribution. For linear mixed models Braun et al. (2012) consider the predictive distribution of the random variable y_{ij} conditional on the other given response values $\boldsymbol{y}_{i,-j} := (y_{i1}, \ldots, y_{i,j-1}, y_{i,j+1}, \ldots, y_{in_i})^T$ of the same subject for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. They argue that there is only a low danger of conservatism due to ignoring the individual random effect as well as the real response value even though the model choice criterion is based on full data. When assuming

normally distributed random effects one also obtains for the distribution of $y_{ij}|\boldsymbol{y}_{i,-j}$ a normal distribution. Unfortunately in our case this distribution is not normal. Thus we extend the approach of Braun et al. (2012) to work in our scenario. We exploit that in the case of known cluster membership the conditional distribution is normal. Because the cluster membership is not known the continuous ranked probability score is weighted by the estimated weights

$$WCRPS(y_{ij}) = \sum_{h=1}^{N} \hat{\pi}_h \, CRPS_h(y_{ij}),$$

where $CRPS_h(y_{ij})$ is given by formula (6) with $y_{obs} := y_{ij}$ as well as

$$\begin{aligned}
\hat{\mu} &:= \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T\hat{\boldsymbol{\mu}}_h + \boldsymbol{z}_{ij}\hat{\boldsymbol{D}}\boldsymbol{Z}_{i,-j}^T\left(\hat{\sigma}^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\hat{\boldsymbol{D}}\boldsymbol{Z}_{i,-j}^T\right)^{-1} \cdot \\
&\quad \cdot(\boldsymbol{y}_{i,-j} - \boldsymbol{X}_{i,-j}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_{i,-j}\hat{\boldsymbol{\mu}}_h), \\
\hat{\sigma} &:= \left(\boldsymbol{z}_{ij}\hat{\boldsymbol{D}}\boldsymbol{z}_{ij}^T - \boldsymbol{z}_{ij}\hat{\boldsymbol{D}}\boldsymbol{Z}_{i,-j}^T\left(\hat{\sigma}^2\boldsymbol{I}_{n_i-1} + \boldsymbol{Z}_{i,-j}\hat{\boldsymbol{D}}\boldsymbol{Z}_{i,-j}^T\right)^{-1}\boldsymbol{Z}_{i,-j}\hat{\boldsymbol{D}}\boldsymbol{z}_{ij}^T + \hat{\sigma}^2\right)^{1/2}.
\end{aligned}$$

This can be shown by derivations from Braun et al. (2012). Here $\boldsymbol{x}_{ij}$ is the jth row of $\boldsymbol{X}_i$ while $\boldsymbol{X}_{i,-j}$ symbolizes the matrix $\boldsymbol{X}_i$ without row j (analog for $\boldsymbol{z}_{ij}$ and $\boldsymbol{Z}_{i,-j}$). Thus $\hat{\mu}$ and $\hat{\sigma}$ are the parameters of the distribution of $y_{ij}|\boldsymbol{y}_{i,-j}, z_{ih} = 1$. Finally, the mean of the weighted continuous ranked probability score is taken over to all measurement points. The best value for the penalization parameter λ is that where the mean of the weighted continuous ranked probability score is maximal.

3 Applications

3.1 Hormonotherapy

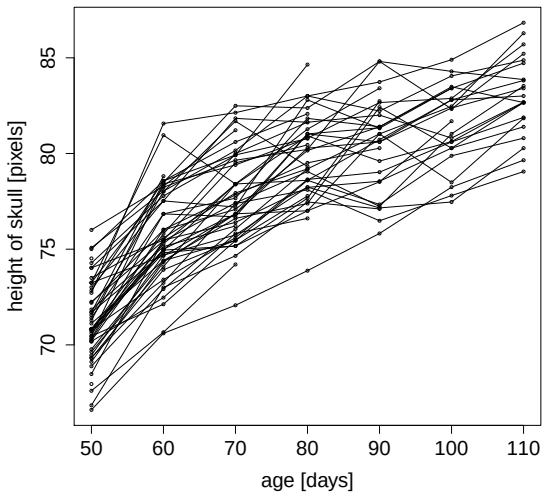
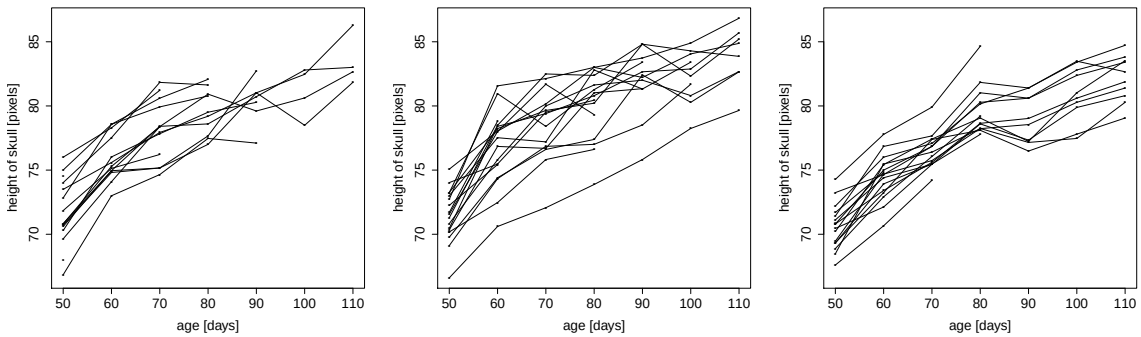


Fig. 1: Heights of rat skulls across age.

In the following the practical use of our model is illustrated by considering the craniofacial growth of male rats. The data were collected in an experiment at the Catholic University of Leuven with the aim to analyze the effect of testosterone on the growth of rats (Verdonck et al., 1998). Therefore 50 male rats have been randomized to either a control group or to one of the two treatment groups that differ in the dose of the drug Decapeptyl, which inhibits the testosterone production. The response of interest is the distance (in pixels) between well-defined points of the skull that characterize the height of skull. These heights have been measured for each rat every 10 days starting at the age of 50 days and the treatment began at the age of 45 days, see Verbeke and Molenberghs (2000) for more information about the data. Figure 1 shows different levels of heights of the skulls and a positive time trend which varies from rat to rat. According to Figure 2 there seems to be a negative effect of the drug Decapeptyl on the growth of rats but the three groups are relatively mixed and can not be clearly separated.

To examine how many and which clusters can be found in these data the penalized heterogeneity approach with a group fused lasso penalty is considered. As suggested by Verbeke and Lesaffre (1999) and also used in the analyzes of Verbeke and Molenberghs (2000) and Fahrmeir et al. (2007) the age of rat i at measurement j is transformed by $t_{ij} = \log(1 + \text{age})_{ij}$ to get a linear time trend. In analogy to Verbeke and Molenberghs (2000) and Fahrmeir et al. (2007) the time trends in each group are



(a) Control group (b) Low dose (c) High dose

Fig. 2: Heights of rat skulls across age depending on treatment group.

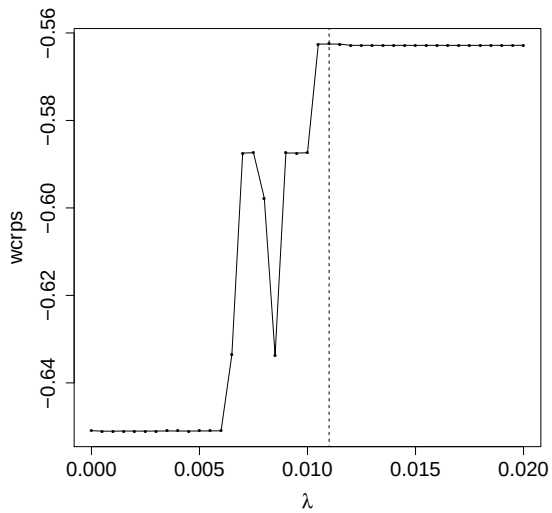


Fig. 3: Weighted continuous ranked probability score depending on λ.

modeled as fixed effects and a random intercept is included. We additionally use a random slope to incorporate individual deviations of the time trend.

In summary we consider the following model for the height y of the skull of rat i at measurement j

$$y_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + \beta_2 L_i + \beta_3 H_i + b_{i1})t_{ij},\, \sigma^2), \quad i = 1, \ldots, 50,\; j = 1, \ldots, n_i,$$

with effect-coded variables L_i and H_i for a low respectively for a high dose of drug. For the centered i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ we assume a mixture distribution of Gaussian components with penalized cluster centers (see Section 1). The four rats for which only one measurement was available were excluded because for these no reasonable random slope can be predicted. For faster computations the algorithm

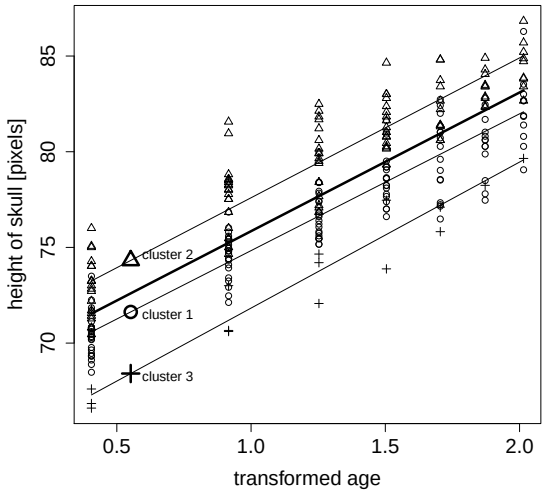
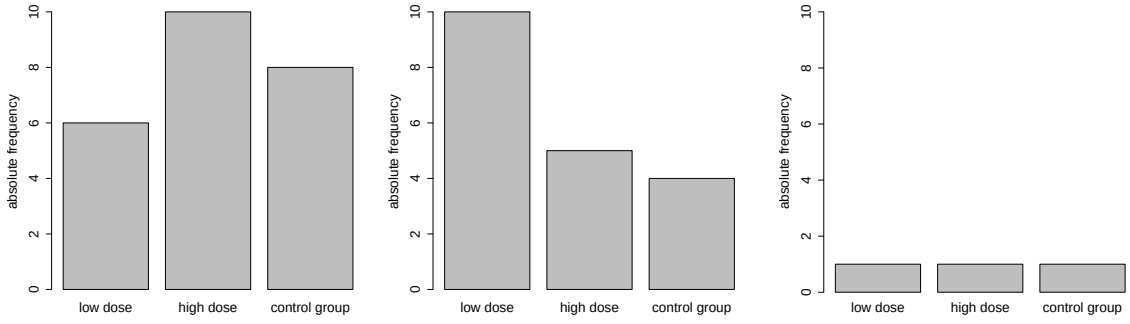


Fig. 4: Clustering of rats by penalized heterogeneity approach with $\lambda = 0.011$.

starts with 20 clusters. Figure 3 suggests to choose the penalization parameter $\lambda = 0.011$. The resulting fit can be seen in Figure 4.

The thick line symbolizes the population effect whereas the thin lines display the cluster centers. Observations belonging to the same cluster are marked with the same symbol. To each thin line the corresponding symbol is added to visualize which cluster center belongs to which cluster. The global intercept $\hat{\beta}_0 = 68.658$ can be interpreted as the mean height at the beginning of the treatment while the global slope $\hat{\beta}_1 = 7.248$ forms the mean growth of rat skulls in the considered time period. The expected negative effect of the drug Decapeptyl can be seen from the estimates $\hat{\beta}_2 = 0.082$ and $\hat{\beta}_3 = -0.459$ which can be interpreted as deviations from the overall time tend. For rats which had been exposed to a low dose of the drug (0.082) the growth is considerably less than in the control group (0.376). For rats in the high dose group the growth is even lower (-0.459). These results are more intuitive than the results obtained by Verbeke and Molenberghs (2000). In their analysis the rats which had been exposed to a low dose show a higher growth than these in the control group though the drug has a negative effect on the growth for a high dose. Obviously our penalized mixture of normal distributions as random effects distribution is much more adequate than a simple normal distribution for these data with a underlying grouping structure.

Three clusters are detected by our model. While there are only low discrepancies in the random slopes ($\hat{\mu}_{11} = -0.100$, $\hat{\mu}_{21} = 0.061$, $\hat{\mu}_{31} = 0.382$) the base levels are quite different. Cluster 2 ($\hat{\pi}_2 = 0.435$) shows the highest intercept which is about $\hat{\mu}_{20} = 1.706$ higher than the overall intercept. By comparison in Cluster 1 ($\hat{\pi}_1 = 0.503$) the base level is considerably lower ($\hat{\mu}_{10} = -0.912$). Cluster 3 ($\hat{\pi}_3 = 0.062$) contains



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Fig. 5: Distribution of the two treatment groups respectively the control group in the three clusters corresponding to a penalized heterogeneity approach with $\lambda = 0.011$.

the three rats with the lowest base level ($\hat{\mu}_{30} = -4.578$). As can be seen from Figure 5 response types collected in the clusters come from all groups. In cluster 1 rats of the high dose group are in the majority followed by rats of the control group. In cluster 2 in particular rats which had been exposed to a low dose of the drug are found.

3.2 Lung Function Growth

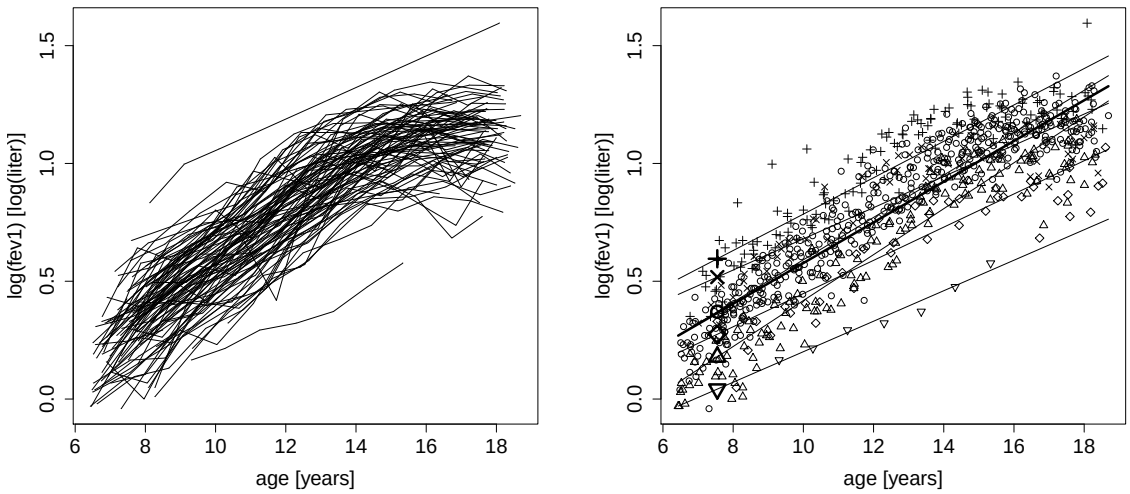


Fig. 6: Logarithmic forced expiratory volume in one second of girls across age: raw data (left) and clustering by penalized heterogeneity approach with $\lambda = 0.0175$ (right).

The second data example deals with lung function growth of girls in Topeka (USA). These data are a subsample from the six cities study of air pollution and health in Dockery et al. (1983). Our sample consists of 100 girls, with a minimum of two and a maximum of twelve observations over time. Although a cluster structure is not

evident from looking at the raw data (Figure 6, left) our approach is able to identify clusters in the data. Again we consider a random slope model

$$\log(fev1)_{ij}|\boldsymbol{b}_i \overset{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})age_{ij},\, \sigma^2), \quad i = 1, \ldots, 100, \quad j = 1, \ldots, n_i,$$

for modeling the logarithmic forced expiratory volume in one second (fev1) subject to age and use a finite mixture as random effects distribution with a group fused lasso penalty. Because of the comparably large number of individuals we start with $N = 30$ clusters instead of 100.

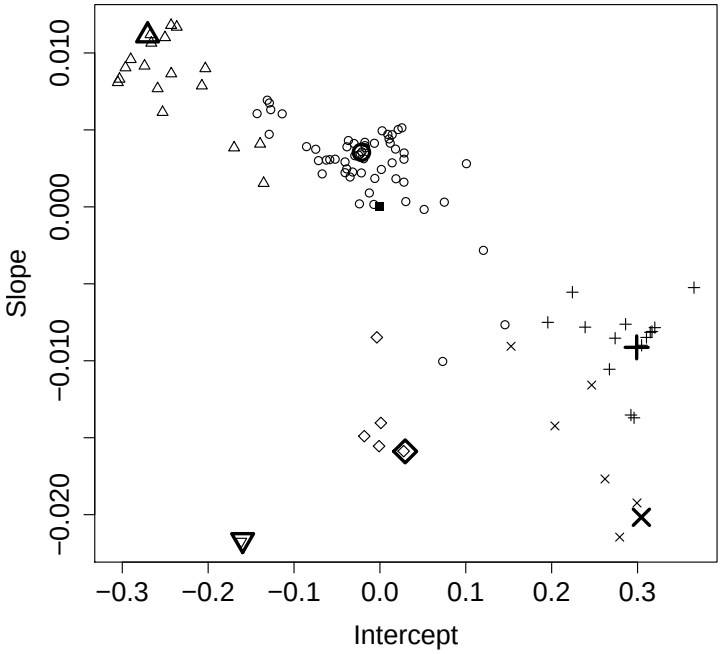


Fig. 7: Cluster locations and corresponding random effects of penalized heterogeneity approach with $\lambda = 0.0175$ for lung function growth data.

In Figure 6 (right) the clustering structure is visualized but it is hard to see which girls are merged to the same cluster. Figure 7 makes clear how the clustering of our approach works. Here on the axes the intercepts and the slopes are drawn. The filled square at coordinates (0,0) symbolizes the population effect. All other icons represent deviations from the population effect. The big bold ones represent the cluster locations $\boldsymbol{\mu}_h$ and the thin small ones the random effects $\boldsymbol{b}_i$. Girls that are assigned to the same cluster are marked with the same symbol and are arranged around the three cluster locations in the form of ellipses. It is easily seen that subjects with random effects that are similar in the meaning of a low Euclidean distance belong to the same cluster.

4 Simulation Study

4.1 Setting

In the following simulation study the performance of our penalized heterogeneity approach is evaluated. The study aims at clarifying in which data situations our approach improves estimation compared to the commonly used LMM with Gaussian random effects distribution and to the heterogeneity model by Verbeke and Lesaffre (1996). Note that the estimated number of clusters and the estimated clustering in general have an essential impact on the prediction accuracy of the random effects. Of course for the prediction of $\boldsymbol{b}_i$ it is reasonable to borrow information from other subjects which show a similar behavior and so belong to the same cluster while incorporating dissimilar individuals impairs the prediction accuracy. For examining this trade-off we compare the usual LMM with normal random effects distribution (one cluster model) using the `R`-function `lmer()` from the `lme4` package by Bates et al. (2011) to our penalized heterogeneity approach at which the penalization parameter λ is determined by predictive cross-validation (see section 2.2). Furthermore, the heterogeneity model by Verbeke and Lesaffre (1996) with a finite unpenalized mixture of normal distributions as random effects distribution is considered, too, where the number of mixture components is identified by the same predictive cross-validation criterion. More precisely, in the simulation study we investigate the impact of the number of observations per unit and the separation between clusters. We generated data sets assuming a simple linear trend model

$$y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})t_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i$$

with i.i.d. errors $\varepsilon_{ij} \sim N(0, \sigma^2)$. The centered i.i.d. random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ follow a mixture distribution with three Gaussian components:

$$\boldsymbol{b}_i \sim 0.4 \cdot N(\boldsymbol{\mu}_1, \boldsymbol{D}) + 0.3 \cdot N(\boldsymbol{\mu}_2, \boldsymbol{D}) + 0.3 \cdot N(\boldsymbol{\mu}_3, \boldsymbol{D}), \quad i = 1, \ldots, n,$$

imitating a population consisting of three clusters of overlapping subpopulations. Throughout the simulations, we set $n = 20$ and

$$\sigma^2 = 0.25, \qquad \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} \sigma^2_{b_0} & \sigma_{b_{01}} \\ \sigma_{b_{01}} & \sigma^2_{b_1} \end{pmatrix} = \begin{pmatrix} 0.02 & 0.01 \\ 0.01 & 0.02 \end{pmatrix}.$$

We vary, however, the number of individual observations n_i, the centers $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ of the clusters and the locations of observation times t_{ij}. To produce longitudinal data with varying numbers of repeated observations per unit i, we set $n_i = 2 + X_i$, where X_i follows a Poisson distribution with rate ν. Setting $\nu = 1$ corresponds to longitudinal data with only few (3 on average) repeated observations per unit,

$\nu = 3$ to a moderate number and $\nu = 5$ to (comparably) large numbers of repeated observations. For given n_i, observation times are generated from

$$\begin{aligned} t_{i1} &\sim U(0,1), \quad i = 1, \ldots, n, \\ t_{ij} &\sim U(t_{i,j-1} + 0.5, t_{i,j-1} + 1.5), \quad i = 1, \ldots, n, \quad j = 2, \ldots, n_i, \end{aligned}$$

where $U(\cdot)$ denotes the uniform distribution. In this way, different numbers $n_i(s)$ and measuring times $t_{ij}(s)$ are generated in each simulation run $s = 1, \ldots, 100$. Similarly, different "true" random effects $\boldsymbol{b}_i(s)$ are drawn from the Gaussian mixture distribution in each simulation run. For the cluster locations, we chose

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -2.25 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.75 \\ -1.2 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 2.25 \\ -2/15 \end{pmatrix}$$

corresponding to *clearly separated clusters*,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1.5 \\ 0.75 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.5 \\ -0.9 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 1.5 \\ -0.1 \end{pmatrix}$$

corresponding to *moderately separated clusters*,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -0.75 \\ 0.5 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.25 \\ -0.6 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 0.75 \\ -1/15 \end{pmatrix}$$

corresponding to *substantially overlapping clusters*.

Combining these different settings for observations times and clusters results in nine different scenarios. For each of them, the prediction accuracy of the random effects as well as the estimation results of the fixed effects are compared for all considered models. More concretely, in each simulation run s, we calculate the average prediction error

$$PE_k(s) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{b}^*_{ik}(s) - b^*_{ik}(s) \right)^2, \quad k = 0, 1$$

for uncentered random intercepts $b^*_{i0} = \beta_0 + b_{i0}$ and random slopes $b^*_{i1} = \beta_1 + b_{i1}$. In addition, the estimation accuracy of the fixed effects is investigated by the relative bias $RB_k = 100 \cdot (\hat{\beta}_k - \beta_k)/\beta_k$, $k = 0, 1$.

4.2 Results

In the following, we summarize results of the nine combinations. For all scenarios we illustrate the empirical distribution of $PE_0(s)$ values obtained from simulation run $s = 1, \ldots, 100$ by box plots. The corresponding figures of the random slopes are not shown because these are very similar to those of the random intercepts.

Clearly separated clusters

Figure 8 (top) displays trace plots of typical longitudinal data generated in the setting of clearly separated clusters, which show that cluster effects can easily be detected visually. On the left, there are only a few observations for each subject while on the right the mean of the number of repeated measurements is 5 corresponding to several observations. Figure 8 (bottom) demonstrates that in both cases the penalized heterogeneity approach detects three clusters. Again, in this type of plot the thick line shows the overall effect and the thin lines visualize the means of the resulting clusters. Which observation is assigned to which cluster is marked by the same symbol.

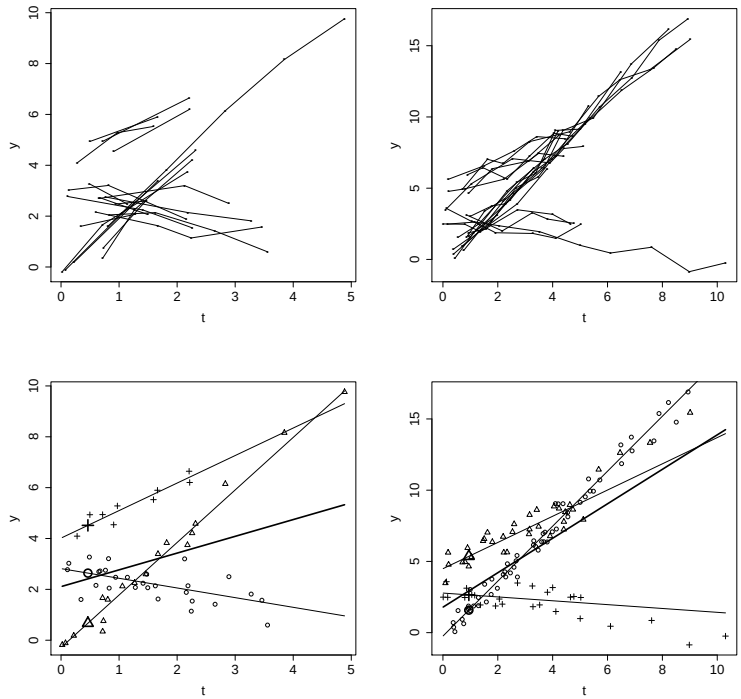
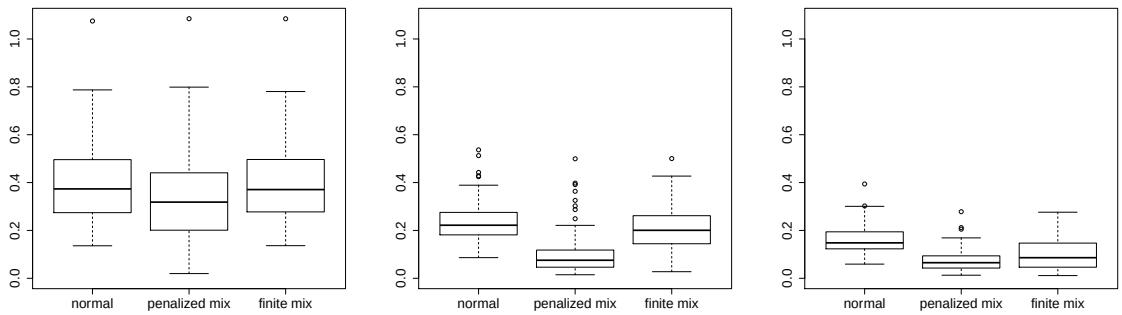


Fig. 8: Trace plots (top) and clustering by penalized heterogeneity approach with clearly separated clusters for few individual observations ($\nu = 1$) (left) and several individual observations ($\nu = 3$) (right).

Table 1 and Figure 9 show the simulation results in the setting of clearly separated clusters. The denotation "normal" labels the homogeneity model with normally distributed random effects. In the heterogeneity model the random effects follow a

"finite mixture" as specified in equation (2) where the number of mixture components has been determined by predictive cross-validation. In contrast to this discrete optimization the approach proposed here uses a penalty term which is determined by a smoothing parameter. It is seen that the penalization approach outperforms the homogeneity model and the heterogeneity model for few observations as well as for several and many observations. It is especially remarkable that the "penalized mixture" yields a better prediction accuracy than in the "finite mixture" although in both cases the same criterion for finding the best number of clusters is used. The reason for that is that for optimization in our penalized heterogeneity approach a closer grid is applied. This is the main justification for our model. Apart from that it can be seen that the more repeated measurements per unit are in the data the better is the prediction accuracy of the penalized heterogeneity approach. Overall there is only a small bias concerning the estimation of fixed effects.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Fig. 9: Box plots of PE_0 with clearly separated clusters for few individual observations (left), several individual observations (middle) and many individual observations (right).

	$\nu = 1$				$\nu = 3$				$\nu = 5$			
	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1
normal	0.373	0.185	-4.091	2.068	0.222	0.054	-1.048	4.710	0.148	0.015	-2.127	0.957
penalized mix	0.318	0.161	-3.530	5.267	0.075	0.015	-3.113	3.938	0.065	0.006	-0.452	0.987
finite mix	0.371	0.186	-4.065	2.241	0.201	0.042	-1.312	5.463	0.086	0.008	-0.453	1.743

Table 1: Medians of PE_k and RB_k with $k = 0, 1$ for clearly separated clusters.

Moderately separated clusters

When the differences between clusters get smaller the penalized heterogeneity approach still outperforms the homogeneity model and the heterogeneity model, especially in the case of several and many individual observations (Figure 11 and Table 2), although in the trace plots (Figure 10) the underlying cluster structure is hard to see.

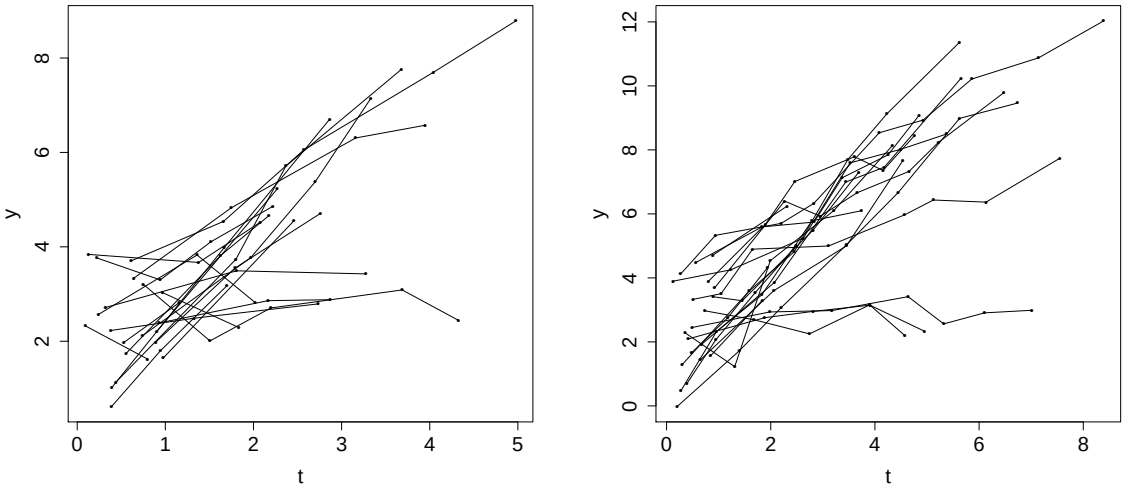
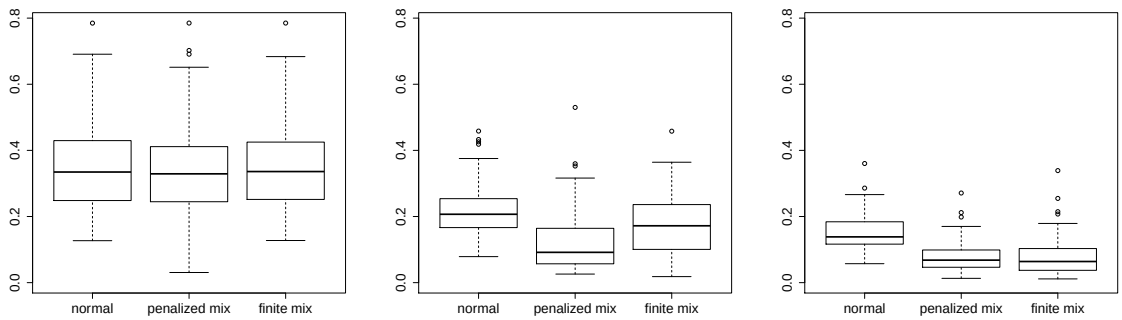


Fig. 10: Trace plots with moderate separated clusters for few individual observations ($\nu = 1$) (left) respectively several individual observations ($\nu = 3$) (right).

	$\nu = 1$				$\nu = 3$				$\nu = 5$			
	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1
normal	0.335	0.164	-2.112	1.912	0.207	0.046	-0.751	2.204	0.138	0.015	-1.122	0.750
penalized mix	0.329	0.155	-2.538	1.982	0.092	0.019	-0.116	2.962	0.068	0.007	-0.870	0.295
finite mix	0.336	0.164	-2.084	1.832	0.172	0.036	-0.481	2.756	0.064	0.007	-1.231	0.887

Table 2: Medians of PE_k and RB_k with $k = 0, 1$ for moderately separated clusters.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Fig. 11: Box plots of PE_0 with moderately separated clusters for few individual observations (left), several individual observations (middle) and many individual observations (right).

Substantially overlapping clusters

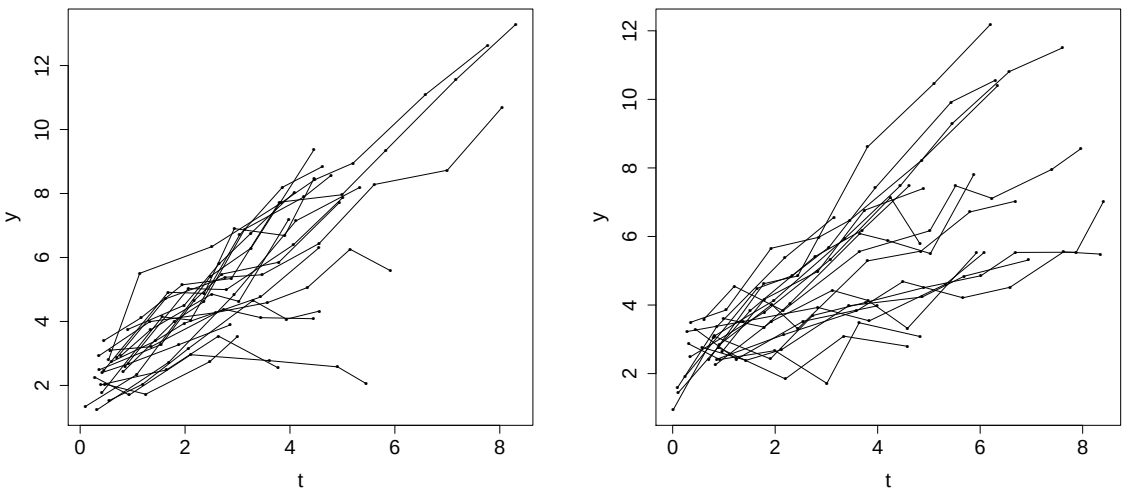


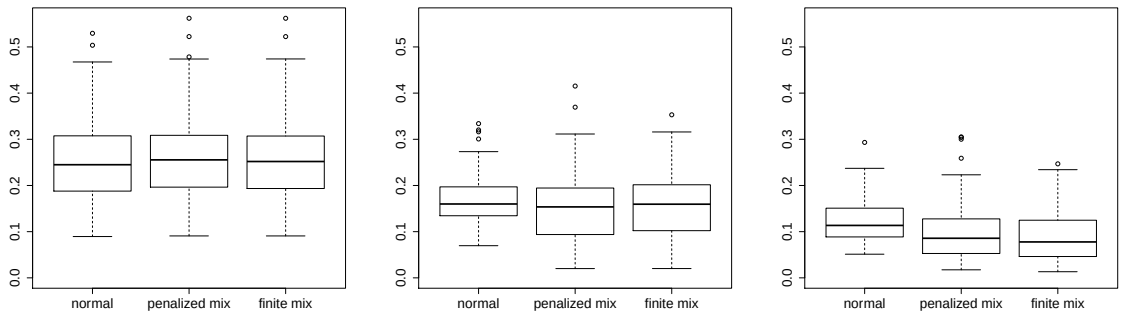
Fig. 12: Trace plots with substantially overlapping clusters for several individual observations ($\nu = 3$) (left) respectively many individual observations ($\nu = 5$) (right).

For data sets like in Figure 12 it would be tempting to use a LMM with normally distributed random effects. Nevertheless even in these settings for penalized heterogeneity approaches prediction errors are significantly lower for several and many observations (Figure 13 and Table 3). Only for few observations the classical LMM with normal random effects distribution outperforms the penalized heterogeneity approach. Here, different patterns in the data are taken seriously. Thus there is a low risk of overfitting the data in the case of few individual observations. Overall the accuracy of estimates of the heterogeneity model and the penalized heterogeneity approach are quite similar.

	$\nu = 1$				$\nu = 3$				$\nu = 5$			
	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1	PE_0	PE_1	RB_0	RB_1
Normal	0.245	0.111	-1.247	1.563	0.160	0.037	0.036	2.304	0.114	0.013	-0.207	1.004
penalized mix	0.255	0.112	-1.906	1.228	0.154	0.032	0.512	1.622	0.086	0.009	0.181	0.854
finite mix	0.252	0.110	-1.356	1.497	0.159	0.033	0.050	1.690	0.078	0.008	-0.095	1.126

Table 3: Medians of PE_k and RB_k with $k = 0, 1$ for substantially overlapping clusters.

In summary, we draw the following conclusion: The penalized heterogeneity approach yields the better predictions for random effects in terms of prediction errors the clearer the clusters differ and the more observations are in the data. Except for substantially overlapping clusters with few observations the prediction error is considerably reduced by using penalization methods.



(a) $\nu = 1$ (b) $\nu = 3$ (c) $\nu = 5$

Fig. 13: Box plots of PE_0 with substantially overlapping clusters for few individual observations (left), several individual observations (middle) and many individual observations (right).

5 Concluding Remarks

We introduced a penalized heterogeneity approach for linear mixed models which assumes a finite mixture of normal distributions for the random effects distribution and which penalizes the number of mixture components by fusing the cluster centers via a group fused lasso penalty term. The approach aims at clustering individuals for longitudinal data. We presented an EM algorithm for estimating all parameters in detail. A simulation study showed that our approach basically outperforms the classical linear mixed model with normal random effects distribution and the heterogeneity model. Furthermore, the usefulness of our model is demonstrated in two data examples: We identified similarities in the development of growth of rats depending on the treatment group and showed that our model is able to detect a underlying cluster structure in the lung function growth data which can not be seen easily in the raw data.

References

Bates, D., M. Maechler, and B. Bolker (2011). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 0.999375-42.

Booth, J. G., G. Casella, and J. P. Hobert (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society B 70*, 119–139.

Braun, J., L. Held, and B. Ledergerber (2012). Predictive cross-validation for the choice of linear mixed-effects models with application to data from the swiss HIV cohort study. *Biometrics 68*, 53–61.

Celeux, G., O. Martin, and C. Lavergne (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling 5*, 243–267.

De la Cruz-Mesía, R., F. A. Quintana, and G. Marshall (2008). Model-based clustering for longitudinal data. *Computational Statistics & Data Analysis 52*, 1441–1457.

Dempster, A. P., N. M. Laired, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Dockery, D. W., C. S. Berkey, J. H. Ware, F. E. Speizer, and B. G. Ferris (1983). Distribution of fvc and fev1 in children 6 to 11 years old. *American Review of Respiratory Disease 128*, 405–412.

Fahrmeir, L., T. Kneib, and S. Lang (2007). *Regression - Modelle, Methoden und Anwendungen*. Berlin: Springer.

Gaffney, S. J. and P. Smyth (2003). Curve clustering with random effects regression mixtures. In C. M. Bishop and B. J. Frey (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Key West, FL.

Greven, S. and T. Kneib (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika 97*, 773–789.

Heinzl, F. and G. Tutz (2011). Clustering in linear mixed models with Dirichlet process mixtures using EM algorithm. Technical Report 115, Ludwig-Maximilians-University Munich.

James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*, 397–408.

Komárek, A., B. E. Hansen, E. M. M. Kuiper, H. R. van Buuren, and E. Lesaffre (2010). Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine 29*, 3267–3283.

Komárek, A. and L. Komárková (2012). Clustering for multivariate continuous and discrete longitudinal data. *To appear*.

Komárek, A. and E. Lesaffre (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics and Data Analysis 52*, 3441–3458.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*, 963–974.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B 34*, 1–41.

Lindstrom, M. J. and D. M. Bates (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association 83*, 1014–1022.

Liu, X. and M. C. K. Yang (2009). Simultaneous curve registration and clustering for functional data. *Computational Statistics & Data Analysis 53*, 1361–1376.

Marshall, E. C. and D. J. Spiegelhalter (2003). Approximate crossvalidatory predictive checks in disease mapping models. *Statistics in Medicine 22*, 1649–1660.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. New York: Wiley.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *Computer Journal 7*, 308–313.

Ng, S. K., G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S.-W. Ng (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics 22*, 1745–1752.

O'Neill, R. (1971). Algorithms AS 47: Function minimization using a simplex procedure. *Journal of the Royal Statistical Society C 20*, 338–345.

Papageorgiou, G. and J. Hinde (2012). Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. *Statistics and Computing 22*, 79–92.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression.* Cambridge: Cambridge University Press.

Scharl, T., B. Grün, and F. Leisch (2010). Mixtures of regression models for time course gene expression data: evaluation of initialization and random effects. *Bioinformatics 26*, 370–377.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Kneight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B 67*, 91–108.

Vaida, F. and S. Blanchard (2005). Conditional Akaike information for mixed-effects models. *Biometrika 92*, 351–370.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*, 217–221.

Verbeke, G. and E. Lesaffre (1999). The effect of drop-out on the effiency of longitudinal experiments. *Applied Statistics 48*, 363–375.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Verdonck, A., L. de Ridder, G. Verbeke, J. P. Bourguignon, C. Carels, E. R. Kuhn, V. Darras, and F. de Zegher (1998). Comparative effects of neonatal and prepurbetal castration on craniofacial growth in rats. *Archives of Oral Biology 43*, 861–871.

Villarroel, L., G. Marshall, and A. E. Barón (2009). Cluster analysis using multivariate mixed effects models. *Statistics in Medicine 28*, 2552–2565.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.