LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Gerhard Tutz and Margret-Ruth Oelker

Modeling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures

Technical Report Number 156, 2014
Department of Statistics
University of Munich

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

Modeling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures

Gerhard Tutz & Margret-Ruth Oelker

Ludwig-Maximilians-Universität München

Akademiestraße 1, 80799 München

{gerhard.tutz}@stat.uni-muenchen.de

October 30, 2014

Abstract

Although each statistical unit on which measurements are taken is unique typically there is not enough information available to account totally for its uniqueness. Therefore heterogeneity among units has to be limited by structural assumptions. One classical approach is to use random effects models, which assume that heterogeneity can be described by distributional assumptions. However, inference may depend on the assumed mixing distribution and it is assumed that the random effects and the observed covariates are independent. An alternative considered here, are fixed effect models, which let each unit have its own parameter. They are quite flexible but suffer from the large number of parameters. The structural assumption made here is that there are clusters of units that share the same effects. It is shown how clusters can be identified by tailored regularized estimators. Moreover, it is shown that the regularized estimates compete well with estimates for the random effects model, even if the latter is the data generating model. They dominate if clusters are present.

Keywords: Fixed effects, random effects, mixture modeling, heterogeneity.

1 Introduction

The modeling of heterogeneity is an important topic when observations are grouped or clustered. The clustering can be due to repeated measurements over time, as in longitudinal studies, or due to sub sampling of the primary sampling units in cross-sectional studies. The assumption that units behave the same is usually much too restrictive and strategies have to be found that take the heterogeneity of effects into account.

Repeated measurements can be represented by $(y_{ij}, \boldsymbol{x}_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, where y_{ij} denotes the response of unit i at measurement occasion j, and $\boldsymbol{x}_{ij}$ is a vector of covariates that potentially varies across measurements. Data have the same form in cross-sectional studies if they are collected in clusters or groups. For example, in a multicenter treatment study, y_{ij} may denote the response of patient j in study center i. In the terminology of multilevel models, the patients are the so-called first level units and the study centers the second level units.

An application that will be considered in more detail later, deals with the effect of beta blockers on the mortality after myocardial infarction, see also Aitkin (1999), Grün and Leisch (2008a). In a 22-center clinical trial, for each center, the number of deceased/successfully treated patients in control/test groups was observed. The binary response (1 = deceased/0 = not deceased) suggests a logit model, which in its simplest form is given by

$$\text{logit } P(y_{ij} = 1) = \beta_0 + \beta_T \cdot \text{Treatment}_{ij}, \quad i = 1, \ldots, 22 \text{ Centers}, \tag{1}$$

where $\text{Treatment}_{ij} \in \{-1, 1\}$ codes the treatment in hospital i for patient j (1: Treatment, −1: Control). Model (1) does not account for the heterogeneity among the hospitals. The treatment effect β_T as well as the basic risk captured in β_0 are assumed to be the same for all hospitals. Of course, this is a very strong assumption that hardly holds.

The most popular model that incorporates heterogeneity is the *random effects model*. It replaces the intercept β_0 by $\beta_0 + b_{i0}$, yielding

$$\text{logit } P(y_{ij} = 1) = \beta_0 + b_{i0} + \beta_T \cdot \text{Treatment}_{ij}, \quad i = 1, \ldots, 22 \text{ Centers}, \tag{2}$$

where b_{i0} is a random effect, for which a distribution is assumed, typically a normal distribution, $b_{i0} \sim N(0, \sigma_b^2)$. Implicitly, the hospitals are considered as a random sample and the inference concerning the treatment effect should hold for the whole underlying set of hospitals. One can go one step further and replace the treatment effect β_T by $\beta_T + b_{iT}$, allowing for heterogeneity of treatments over hospitals. Random effects models are a strong tool to model heterogeneity and a wide body of literature is available (see, for example, Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Tutz, 2012). However, the approach has several drawbacks. One is that inference on the unknown distributional assumption is hard to obtain and the choice of the distribution may affect the results, see, for example, Heagerty and Kurland (2001), Agresti et al. (2004), Litière et al. (2007). In a more recent paper, McCulloch and Neuhaus (2011) give an overview on results and discuss the impact of the random effects distribution. Another drawback of the random effects model is that it is assumed that random effects and covariates are uncorrelated with the effect that estimation accuracy suffers. In special cases, one can use alternative estimators that are consistent, for example, conditional likelihood methods (Diggle et al., 2002, Section 9.2.1) can be used for for canonical links. Also mixed

effects models that decompose covariates into between- and within-cluster components (Neuhaus and McCulloch, 2006; Grilli and Rampichini, 2011) can alleviate the problem of poor estimates in specific settings.

Moreover, the widely used assumption of normally distributed random effects implicitly assumes that all hospitals differ with respect to the basic risk and/or treatment effect. In other words: hospitals with the same basic risk are not designated and the hospitals themselves are not clustered. If one wants to investigate which hospitals show the same characteristics concerning the risk of mortality while accounting for heterogeneity, random effects models are not the best choice. Of course one can fit a random effects model, then look for similar effects and refit, but this involves several steps of model fitting. If one intends to detect clusters, this can be obtained more elegantly by allowing clusters from the beginning.

An alternative that is considered here, are *group-specific models* which belong to the class of *fixed effects models*. In this class of models, the effects are considered as unknown but fixed. In the beta blocker data, the intercept β_0 is replaced by the parameter β_{i0} and the treatment effect is (potentially) replaced by β_{iT}. The obvious disadvantage is that the number of parameters increases. But, as will be shown, carefully tailored regularization methods allow to reduce the number of parameters and to identify clusters of hospitals with identical performance. In contrast to random effects models, in fixed effects models, the inference refers to the given sample; that is, to the hospitals in the data set. The second level units are not considered as representatives of an underlying population. Thus, fixed effects models are especially useful when one is interested in the performance of specific units.

The objective of the paper is to show that group-specific approaches in combination with regularized estimates are an attractive alternative to existing approaches, in particular in cases where the units themselves are of interest. They allow to identify clusters with identical effects on the response, which is one of the major topics considered.

The paper is organized as follows: in Section 2, three approaches to model heterogeneity are shortly sketched. In Section 3, we propose regularized estimates for group-specific models. In Section 4, the performance of penalized group-specific models is investigated. Section 5 gives some extensions on the simultaneous fusion of group-specific effects related to several covariates. In Section 6 and 7, data on beta blockers and on the math performance of pupils in ten different schools in the U.S. are analyzed.

2 Modeling Heterogeneity

In the following, we consider methods that model heterogeneity. We start with methods that are based on random effects; afterwards, we consider group-specific models, which are most flexible but call for regularized estimation procedures.

2.1 Random Effects Models

Let the observations be given as y_{ij}, where j denotes an observation in the second level unit i, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. In addition, let $\boldsymbol{x}_{ij}^T = (1, x_{ij1}, \ldots, x_{ijp})$ be a covariate vector associated with fixed effects and $\boldsymbol{z}_{ij}^T = (z_{ij1}, \ldots, z_{ijq})$ be a covariate vector associated with random effects.

The structural assumption in a generalized linear mixed effects model (GLMM), specifies that the conditional means $\mu_{ij} = \mathrm{E}(y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$ have the form

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i = \eta_{ij}^{par} + \eta_{ij}^{rand}, \tag{3}$$

where g is a monotonic and continuously differentiable link function and $\eta_{ij}^{par} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$, which includes an intercept. The second term, $\eta_{ij}^{rand} = \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$, contains the random effects that model the heterogeneity of the second level units. For the random effects, one assumes a distributional form, typically a normal distribution, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{Q})$, with covariance matrix $\boldsymbol{Q}$.

In a GLMM, the distributional assumption for $y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}$ is of the exponential family type $f(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i) = \exp\{(y_{ij}\theta_{ij} - \kappa(\theta_{ij}))/\phi + c(y_{ij}, \phi)\}$, where $\theta_{ij} = \theta(\mu_{ij})$ denotes the natural parameter, $\kappa(\theta_{ij})$ is a specific function corresponding to the type of the exponential family, $c(\cdot)$ is the log-normalization constant and ϕ the dispersion parameter (compare Fahrmeir and Tutz, 2001). Moreover, it is assumed that the observations y_{ij} are conditionally independent with means $\mu_{ij} = \mathbb{E}(y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$ and variances $\mathrm{var}(y_{ij}|\boldsymbol{b}_i) = \phi\upsilon(\mu_{ij})$, where $\upsilon(\cdot)$ is a known variance function.

The focus of the random effects models is on the fixed effects; the distribution of the random effects is mainly used to account for the heterogeneity of the second level units. Although it is the most popular model that accounts for heterogeneity, it has some drawbacks. The assumption of a specific distribution for the random effects may affect the inference. In particular, if the distributional assumption is far from the data generating distribution, inference can be strongly biased. Moreover, assuming a continuous distribution prevents that the effects of units can be the same. Therefore, by assumption, no clustering of units is available. One further aspect is that it is assumed that the random effects and the covariates observed per second level unit are independent; a criticism that has a long tradition, in particular in the econometric literature, see, for example, Mundlak (1978).

2.2 Group-Specific Models

An alternative to random effects models are fixed effects model. They model heterogeneity by using a parameter $\boldsymbol{\beta}_i$ instead of the random effect. In repeated measurements studies, they are also called *subject-specific* models; in cross-sectional studies, one might prefer the term *group-specific*, which is used in the following. For the link between explanatory

variables and the mean $\mu_{ij} = E(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$, the group-specific model assumes

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_i. \tag{4}$$

The model specifies that each group or second-level unit has its own vector of coefficients $\boldsymbol{\beta}_i^T = (\beta_{i0}, \ldots, \beta_{iq}), i = 1, \ldots, n$, which represents weights on the vector $\boldsymbol{z}_{ij}^T = (1, z_{ij1}, \ldots, z_{ijq})$. The problem with these models is that the large number of parameters can render the estimates unstable and encourage overfitting. Typically, there is not enough information available to distinguish among all the units; but under the assumption that observations form clusters with respect to their effect on the response, the number of parameters can be reduced and estimates are available. In contrast to common approaches, we assume that $\boldsymbol{z}_{ij}$ is not a subset of $\boldsymbol{x}_{ij}$ in order to avoid identifiability problems.

The group-specific term in the model (4) can also be seen as a varying-coefficient term. It represents the interaction between the variables in $\boldsymbol{z}_{ij}$ and the groups. Let us consider the model with group-specific intercepts,

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \beta_{i0},$$

where $\boldsymbol{z}_{ij} = 1$, in more detail. Let the groups in $\{1, \ldots, n\}$ be coded by the dummy variables $x_{C(1)}, \ldots, x_{C(n)}$, where $x_{C(i)} = 1$ if $C = i$, $x_{C(i)} = 0$, otherwise. Then the model can be written as

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + x_{C(1)}\beta_{10} + \ldots + x_{C(n)}\beta_{n0}.$$

Since the intercept depends on the groups, it is a model where the effect modifier is a factor. In this case, only the variable $\boldsymbol{z}_{ij} = 1$ is modified. In the general case, the model has the form

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + x_{C(1)}\boldsymbol{z}_{i1}^T\boldsymbol{\beta}_1 + \cdots + x_{C(n)}\boldsymbol{z}_{in}^T\boldsymbol{\beta}_n,$$

where the products of z-variables and the dummies for the groups represent the interaction terms. Then, the factor "group" modifies the effects of all the the z-variables. It should be noted that the dummy variables used to denote the group are given as 0-1 variables without a reference category.

2.3 Random Effects Models Versus Group-Specific Models

The comparison of random effects models and group-specific model, which are also called fixed effects models, has a long tradition. More recently, Townsend et al. (2013) summarized much of the work that has been done concerning the choice between random and fixed effects. There are various criteria that can be used when comparing the two approaches.

One advantage of fixed effects models refers to the underlying assumptions. The assumptions in fixed effects models are weaker because in contrast to random effects models,

conditional independence between the covariates and the groups has not to be postulated. Although this does not mean that the model is more robust to other violations of the model (see Townsend et al., 2013), it should suffer less from the violation of conditional independence between the covariates and the groups.

What is often considered as a drawback of fixed effects models is the reduced efficiency of estimates. The problem is that for a large number of groups, the number of degrees of freedom is consumed by the fixed effects. With 60 groups, the fixed effects model with a group-specific intercept and one explanatory variable has 61 parameters, whereas the random intercept model contains one intercept, one slope parameter and requires only one parameter for the heterogeneity, namely $\sigma_b = \text{var}(b_i)$. But the effective degrees of freedom is typically larger. Ruppert et al. (2003) considered the linear random effects model

$$y_{ij} = \beta_0 + b_i + \beta x_{ij} + \varepsilon_{ij},$$

with $\sigma_b = \text{var}(b_i)$ and $\sigma_\varepsilon = \text{var}(\varepsilon_{ij})$. Then, the vector of fitted values can be written as $\hat{\boldsymbol{y}} = \boldsymbol{H}_0\boldsymbol{y} + \boldsymbol{H}_b\boldsymbol{y} + \boldsymbol{H}_x\boldsymbol{y}$, where $\boldsymbol{H}_0$ refers to the intercept, $\boldsymbol{H}_b$ to the random effects and $\boldsymbol{H}_x$ to the predictors x_{ij}. The hat matrices yield the effective degrees of freedom for the components of the model as $df_0 = tr(\boldsymbol{H}_0)$, $df_b = tr(\boldsymbol{H}_b)$, $df_x = tr(\boldsymbol{H}_x)$. One obtains $df_0 = df_x = 1$; for balanced designs with $n_i = m$ for all i, it holds that

$$df_b = \frac{(n-1)m}{m + \sigma_\varepsilon^2/\sigma_b^2}.$$

Thus, the effective number of parameters depends on the ration $\sigma_\varepsilon^2/\sigma_b^2$. In the extreme case $\sigma_b^2 = 0$, one obtains a model with two parameters, namely the intercept and the slope; in the case $\sigma_b^2 \to \infty$, one obtains the fixed effects model with $n + 1$ parameters. Therefore, the random effects model can be seen as a compromise between these extreme cases and the fixed model itself represents an extreme case of the random effects model. The closeness to the fixed effects model is determined by the ratio of within-group and between-group variance components. The possibly large number of parameters of the fixed effects model has led to several recommendations to use the fixed effects model, in particular when there are few groups and moderately large numbers of observations in each, see, for example, Goldstein (2011). However, this restriction does not hold for the approach advocated here. An advantage of the approach is that the number of parameters of the fixed model is implicitly reduced by assuming that the groups form clusters. With the methods considered in Section 3, the cluster-effects can be efficiently estimated.

Moreover, the potential loss of efficiency has to be weighted against the bias reduction obtained by the fixed effects model. By adding group-specific indicators as explanatory variables, the fixed effects model controls for all sorts of confounders. That means, in the clinical trial example, it controls for all the confounding variables such as different sizes and different patient populations of the centers.

One further issue in the comparison of fixed effects and random effects models is that the former postulate that the x-variables have to vary across first-level units. If the x-variables are split into $(\boldsymbol{x}_i^T, \boldsymbol{x}_{ij}^T)$ with the first component representing explanatory variables on the group level, in the corresponding model $g(\mu_{ij}) = \beta_{i0} + \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_i$ the term $\boldsymbol{x}_i^T\boldsymbol{\beta}_1$ can be absorbed in the fixed effect β_{i0}. Thus, in the classical fixed effects model, group-specific explanatory variables cannot be included. This is considered a disadvantage of fixed effects models. However, if regularized estimates as considered in Section 3 are used, also the effects of group-specific explanatory variables can be estimated. We will not consider this in detail here, but refer to Tutz and Schauberger (2014) for an example.

2.4 Finite Mixture Models

An alternative approach to identify clusters in multilevel models is based on finite mixtures. As a competing approach, it is sketched briefly and will be included in our simulation settings. In finite mixtures of generalized linear models, it is assumed that the density or mass function of observation y given $\boldsymbol{x}$ is a mixture

$$f(y|\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f_k(y|\boldsymbol{x}, \boldsymbol{\beta}_k, \phi_k), \tag{5}$$

where $f_k(y|\boldsymbol{x}, \boldsymbol{\beta}_k, \phi_k)$ represents the k-th component of the mixture that follows a simple exponential family parameterized by the parameter vector from the model $\mu_k = \mathrm{E}(y|\boldsymbol{x}, k) = h(\boldsymbol{x}^T\boldsymbol{\beta}_k)$ with response function $h(\cdot)$ and the dispersion parameter ϕ_k. The unknown component weights follow $\sum_{k=1}^{K} \pi_k = 1, \pi_k > 0, k = 1, \ldots, K$.

For hierarchical settings, the components can be linked to the second level units. Let $C = \{1, \ldots, n\}$ denote the set of units that are observed. Then, one specifies one model for the k-th component

$$g(\mu_{ij}) = \beta_{k(i)} + \boldsymbol{x}_{ij}^T\boldsymbol{\beta},$$

where $\beta_{k(i)}$ denotes that the component membership is fixed for each second level unit, that is, $\beta_{k(i)} = \beta_k$ for all $i \in C_k$, where $C_1, \ldots, C_K$ is a disjunct partition of C. Therefore, the units are clustered into subsets with identical intercepts with the total vector of coefficients being given by $\boldsymbol{\alpha}^T = (\beta_1, \ldots, \beta_K, \boldsymbol{\beta}^T)$.

Mixture models were, for example, considered by Follmann and Lambert (1989), and Aitkin (1999). An extensive treatment was given by Fruehwirth-Schnatter (2006). Follmann and Lambert (1989) investigated the identifiability of finite mixtures of binomial regression models and gave sufficient identifiability conditions for mixing at the binary and the binomial level. Grün and Leisch (2008b) consider identifiability for mixtures of multinomial logit models and provide the `R` package *flexmix* with various applications (Grün and Leisch, 2008a).

For the estimation of mixture models with a fixed number of mixture components, typically, the EM-algorithm is employed. The number of mixture components is chosen in

a second step, for example, by information criteria. Regularization methods for mixture models are in its infancy and focus on the selection of variables (Khalili and Chen, 2007; Städler et al., 2010). However, regularization techniques for the selection of the mixture components and therefore the clustering of second level units with respect to their effects, seem not yet available.

3 Regularized Estimation for Group-Specific Models

The basic concept to enforce the clustering of second-level units according to their effect strengths, is penalized maximum likelihood (ML) estimation. Let all the parameters be collected in $\boldsymbol{\alpha}^T = (\boldsymbol{\beta}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_n^T)$, with $\boldsymbol{\beta}_i$, $i = 1, \ldots, n$, denoting the group-specific parameters. Instead of maximizing the log-likelihood, one maximizes the penalized log-likelihood

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}),$$

where $l(\boldsymbol{\alpha})$ denotes the familiar unpenalized log-likelihood, the parameter λ is a tuning parameter, and $J(\boldsymbol{\alpha})$ is a penalty term that enforces clustering of second-level units. The choice of the penalty is crucial because it determines the clusters to be found.

For simplicity, we first assume group-specific intercepts only; that is, the model is given by $g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \beta_{i0}$, $i = 1, \ldots, n$. Then, a specific penalty term that enforces clustering is given by the pairwise differences of group-specific coefficients:

$$J(\boldsymbol{\alpha}) = \sum_{r>m} |\beta_{r0} - \beta_{m0}|. \tag{6}$$

The effect of the penalty is seen by looking at the extreme values of the tuning parameter λ. If $\lambda = 0$, one obtains the unpenalized estimates of $\boldsymbol{\alpha}$ and each second-level unit has its own intercept. If $\lambda \to \infty$, the penalty enforces that the estimates of all group-specific intercepts are the same. Then, the second level units form one cluster with the same intercept. Penalty (6) is a specific version of the fused lasso, which has been considered by Tibshirani et al. (2005) for ordered predictors. The use for categorical predictors has been propagated by Bondell and Reich (2009) for factorial designs and as a selection tool by Gertheiss and Tutz (2010).

In the general case with q covariates $\boldsymbol{z}_{ij}^T = (1, z_{ij1}, \ldots, z_{ijq})$, one uses the pairwise differences of all group-specific coefficients

$$J(\boldsymbol{\alpha}) = \sum_{s=0}^{q} \sum_{r>m} |\beta_{rs} - \beta_{ms}|. \tag{7}$$

The penalty enforces that for $\lambda \to \infty$, all the estimated group-specific parameters of a covariate s are the same, that is, $\beta_{1s} = \ldots = \beta_{ns} = \beta_s$. Hence, for $\lambda \to \infty$, there is one global parameter β_s per covariate.

If $\boldsymbol{z}_{ij}$ is a subset of $\boldsymbol{x}_{ij}$, the model is not identifiable. For this reason in our representation $\boldsymbol{z}_{ij}$ is not a subset of $\boldsymbol{x}_{ij}$. A representation of this form can always be obtained. Let $\boldsymbol{x}_{ij}$ be

partitioned into $\boldsymbol{x}_{ij}^T = (\boldsymbol{z}_{ij}^T, \boldsymbol{w}_{ij}^T)$ and accordingly $\boldsymbol{\beta}$ into $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_z^T, \boldsymbol{\beta}_w^T)$. Then the model with group-specific effect on $\boldsymbol{z}_{ij}$ only is

$$g(\mu_{ij}) = \boldsymbol{z}_{ij}\boldsymbol{\beta}_z + \boldsymbol{w}_{ij}^T\boldsymbol{\beta}_w + \boldsymbol{z}_{ij}^T\tilde{\boldsymbol{\beta}}_i,$$

where, for identifiability, some constraint on the vectors $\tilde{\boldsymbol{\beta}}_i$ is needed, for example, $\sum_i \tilde{\boldsymbol{\beta}}_i = 0$. But the model can also be given as

$$g(\mu_{ij}) = \boldsymbol{w}_{ij}^T\boldsymbol{\beta}_w + \boldsymbol{z}_{ij}^T(\boldsymbol{\beta}_z + \tilde{\boldsymbol{\beta}}_i) = \boldsymbol{w}_{ij}^T\boldsymbol{\beta}_w + \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_i,$$

where $\boldsymbol{z}_{ij}$ is not a subset of $\boldsymbol{w}_{ij}$ and the parameters $\boldsymbol{\beta}_i = \boldsymbol{\beta}_z + \tilde{\boldsymbol{\beta}}_i$ are not restricted.
Adaptive versions of Lasso-type penalties have been shown to have better properties in terms of variable selection. For the basic Lasso, this has been demonstrated by Zou (2006); for categorical predictors, similar results were obtained by Bondell and Reich (2009) and Gertheiss and Tutz (2010). With $w_{rms} = |\tilde{\beta}_{rs} - \tilde{\beta}_{ms}|^{-1}$ where $\tilde{\beta}_{rs}$ denotes an $\sqrt{n}$-consistent estimate as the ML estimate, one obtains an adaptive version of the penalty given by

$$J(\boldsymbol{\alpha}) = \sum_{s=0}^{q}\sum_{r>m} w_{rms}|\beta_{rs} - \beta_{ms}|. \tag{8}$$

The effect of the adaptive weights w_{rms} is that for a very small value $|\tilde{\beta}_{rs} - \tilde{\beta}_{ms}|$, the weights become very large, such that estimates of β_{rs} and β_{ms} have to be similar because otherwise the penalty term itself becomes huge. As group-specific models can be seen as varying coefficient models, the adaptive weighting allows to prove asymptotically normal estimates and asymptotically consistent variable selection (for details, see, Oelker et al., 2014). It should be noted that asymptotic properties are available for a fixed number of second level units whereas for mixed models, large sample theory requires an increasing number of second level units.

3.1 Computational Issues

The proposed penalties are L_1-type penalties on differences of parameters. A general scheme to obtain penalized estimates that are built from linear combinations of parameters was proposed by Oelker and Tutz (2013). It is based on a quadratic approximation of the penalty as proposed earlier by Fan and Li (2001). The basic idea is to approximate the absolute values such that the approximated objective is differentiable. Then for the approximated penalty a penalized iteratively re-weighted least squares (PIRLS) is derived. Penalty (7) does perfectly fit into this framework. The results presented here are obtained with the corresponding `R` package `gvcm.cat` (Oelker, 2013).

3.2 Choice of Tuning Parameters

A common way to obtain data driven tuning parameters is cross-validation. However, cross-validation based on omitting vectors of observations $\boldsymbol{y}_i^T = (y_{i1}, \ldots, y_{in_i})$, will not

work as excluding second level observations changes the model. Assume, for example, a simple model with group-specific intercepts, $g(\mu_{ij}) = \beta_{i0} + \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$. If the vector $\boldsymbol{y}_i$ is excluded from the data set, the parameters β_{i0} change their values, and, more severe, when predicting the outcome for the omitted observation $\boldsymbol{y}_i$, no estimate of β_{i0} is available.
Therefore, it is preferable to use cross-validation methods that allow to estimate the group-specific effects of all observations. One strategy is to exclude only parts of the measurements observed for unit i. When using the observation $\boldsymbol{y}_i$ for validation, one randomly selects components from the vector $\boldsymbol{y}_i^T = (y_{i1}, \ldots, y_{in_i})$ to obtain sub-vectors $\boldsymbol{y}_{i_1}$ and $\boldsymbol{y}_{i_2}$. The first one is kept in the learning sample while the last one is used in the validation sample. In k-fold cross-validation, all the observations that are used for validation are split into sub vectors, and only the first one is used in the learning sample. In order to obtain stable estimates, the first sub vectors to be used in the learning sample have to be sufficiently long.
Alternatively, the tuning parameters can be estimated by a generalized cross-validation (GCV) criterion, as proposed, for example, by O'Sullivan et al. (1986). The criterion avoids that the data have to be split into a learning and a test data set. However, it requires to estimate the degrees of freedom of the model.

4 Numerical Experiments

There are basically two situations in which the proposed fixed effects approach is of special interest: when one can assume that the second level units build clusters that shall be detected or when the assumptions for the random effects model are not fulfilled. Hence, in this section, there are two basic distinctions: whether the assumptions for a mixed model do hold and whether there are clusters or not.

Data with Correlation Between Predictor and Group-Specific Effects

To break the mixed model assumptions, we consider data with so-called level 2 endogeneity, which means that a correlation between the group-specific effect β_{i0}, the random effects b_{i0} respectively, and the coefficients x_{ij} is present. Correlations of this type cause biased estimates when the mixed model is fitted; see Grilli and Rampichini (2011), for examples with Gaussian responses.
To generate the data, we consider the joint distribution of $(\beta_{i0}, \boldsymbol{x}_i^T)$. Assume a multivariate normal distribution for $(\beta_{i0}, \boldsymbol{x}_i^T)$ with $\rho = corr(\beta_{0i}, x_{ij}) \neq 0$. Unfortunately, the covariance matrix of this distribution is not necessarily positive definite for arbitrary values of $corr(x_{ij}, x_{ik})$, $j \neq k$. Therefore, we apply a sequential procedure that is based on two-dimensional distributions: In a first step, β_{i0} is generated by $\beta_{i0} \sim N(\mu_0, \sigma_0^2)$. In a second step, n_i univariate standard normal variables x_{ij} are drawn and transformed according to the bivariate normal distribution of β_{i0} and x_{ij}. In an exemplary setting

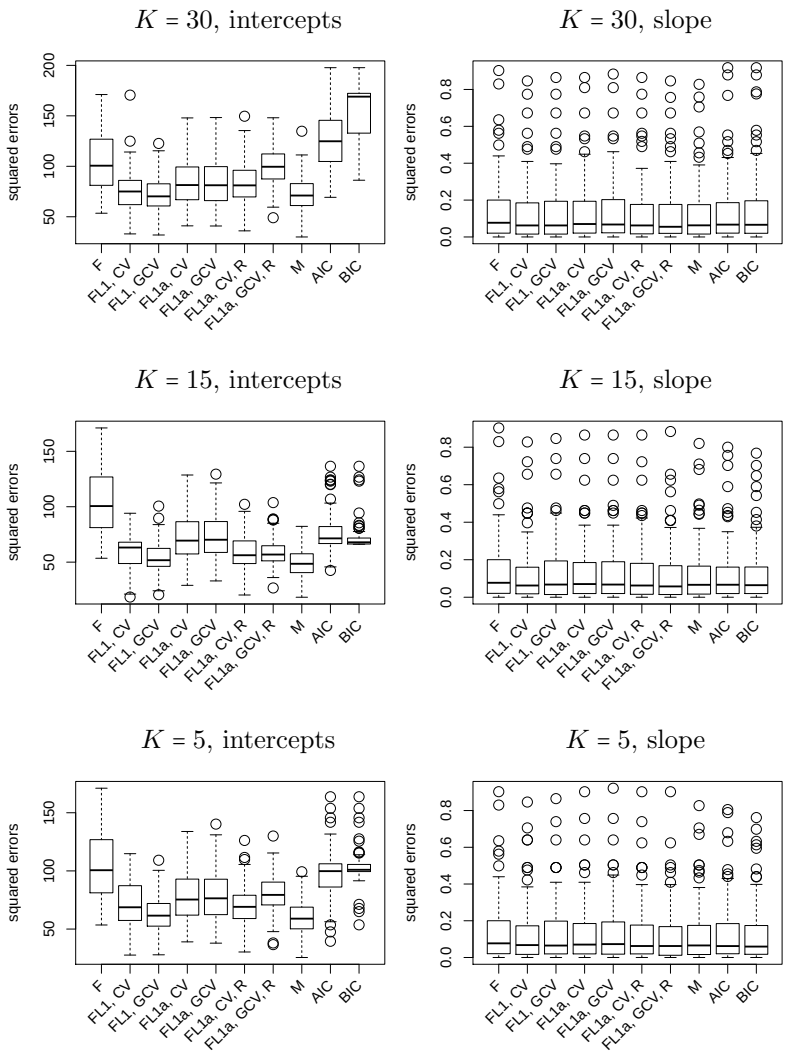

FIGURE 1: *Squared errors for the settings with Gaussian response and* $\beta_{i0} \sim N(1, 4)$ *(GN). The number of clusters* K *varies with the rows. The left panel relates to the intercepts, the right panel to the slopes.* $\rho = 0.0$, $n_i = 10$.

for $\rho = 0.8$, the average empirical correlation of β_{i0}, x_{ij} based on 1000 replications was 0.8016 ($n = 30$, $n_i = 10$, $\mu_0 = 1$, $\sigma_0^2 = 4$, $\mu_x = 0$, $\sigma_x^2 = 1$). The average range of $corr(x_{ij}, x_{ik})$, $j \neq k$, was $(0.4255, 0.8105)$. In an alternative setting, we consider skewed distributions for β_{i0}. In this case, the joint distribution of $(\beta_{i0}, \boldsymbol{x}_i^T)$ is not multivariate normal but the sequential procedure can be applied with small modifications. Let, for example, β_{i0} be drawn from a χ^2-distribution. The transformations to obtain x_{ij} are the same as before but refer to the empirical counterparts of μ_0 and σ_0^2. With $\beta_{i0} \sim \chi_3^2$, β_{i0} centered such that $\mu_0 = 1$, and the same parameters as in the exemplary setting above, the average empirical correlations behave the same as for $\beta_{i0} \sim N(1, 4)$.

Clustered Second Level Units

To construct clustered second level units, the group-specific intercepts β_{i0} are ordered by size and assigned to clusters $C_1, \ldots, C_K$. If one considers $n = 30$ second level units

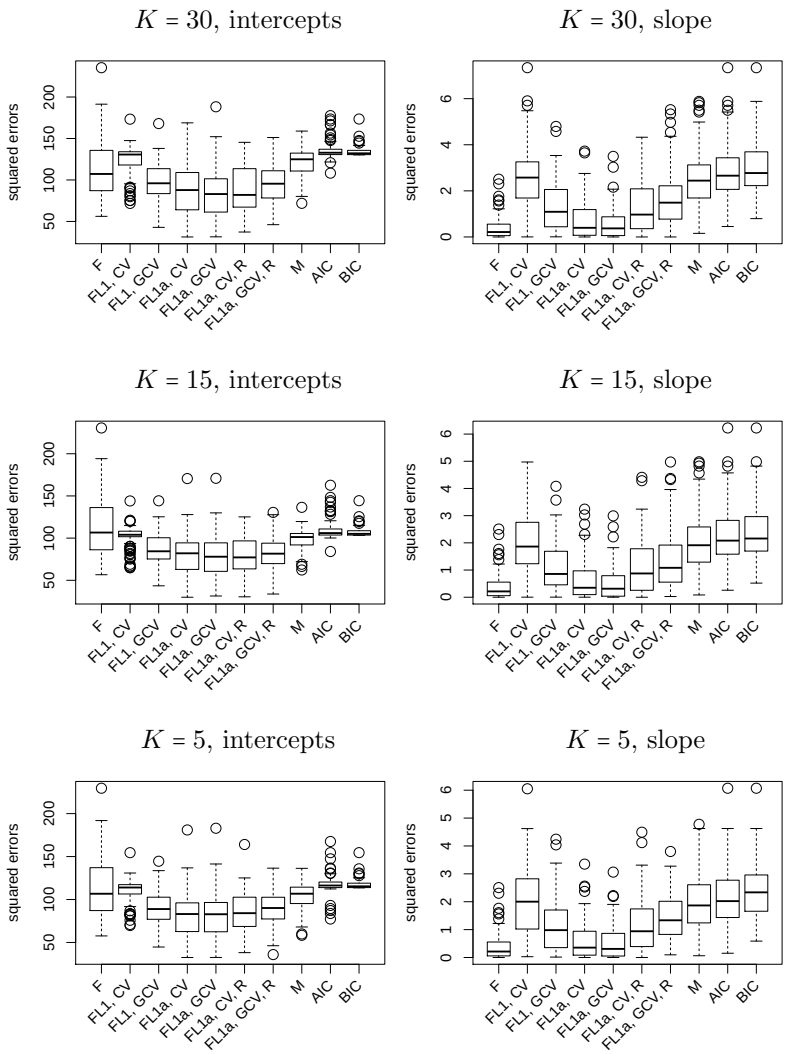

FIGURE 2: *Squared errors for the settings with Gaussian response and* $\beta_{i0} \sim N(1, 4)$ *(GN). The number of clusters* K *varies with the rows. The left panel relates to the intercepts, the right panel to the slopes.* $\rho = 0.8$, $n_i = 10$.

in which $K = 5$ clusters are to be generated, each cluster contains six group-specific intercepts; the six smallest in the first cluster, and so forth. Then, the mean of a cluster yields the cluster specific intercept β_{k0}, $k = 1, \ldots, K$. If one wants level 2 endogeneity as well as clustered second level units, the second level units are generated as described and clustered afterwards.

4.1 Settings

To illustrate that the proposed method works well, simulation settings are varied systematically. In the first set of settings, the response is Gaussian. The model contains the group-specific intercepts β_{i0} or random intercepts b_{i0}, respectively, and only one covariate $x_{ij} \sim N(0, 1)$ with impact $\beta_1 = 2$. The distribution of the subject specific intercepts is either symmetric or skewed: $\beta_{i0} \sim N(1, 4)$ or $\beta_{i0} \sim \chi^2_3$, where the χ^2-distribution is centered such that $\mu_b = 1$. In all settings, $n = 30$; the number of clusters K in the second level

Random Intercepts				F	FL1, CV	FL1, GCV	FL1a, CV	FL1a, GCV	FL1a, CV, R	FL1a, GCV, R	M	AIC	BIC
Gaussian	$K = 30$	$\rho = 0$	FP	-	-	-	-	-	-	-	-	-	-
			FN	0.00	0.06	0.04	0.12	0.11	0.23	0.25	0.00	0.60	0.81
		$\rho = 0.8$	FP	-	-	-	-	-	-	-	-	-	-
			FN	0.00	0.73	0.07	0.22	0.14	0.40	0.29	0.00	0.95	1.00
	$K = 15$	$\rho = 0$	FP	1.00	0.52	0.93	0.79	0.83	0.61	0.62	1.00	0.11	0.02
			FN	0.00	0.47	0.05	0.17	0.13	0.35	0.31	0.00	0.84	0.98
		$\rho = 0.8$	FP	1.00	0.25	0.92	0.74	0.83	0.55	0.61	1.00	0.04	0.00
			FN	0.00	0.74	0.07	0.21	0.14	0.39	0.30	0.00	0.95	1.00
	$K = 5$	$\rho = 0$	FP	1.00	0.70	0.94	0.80	0.84	0.64	0.61	1.00	0.19	0.04
			FN	0.00	0.27	0.04	0.13	0.11	0.27	0.27	0.00	0.68	0.92
		$\rho = 0.8$	FP	1.00	0.27	0.93	0.79	0.82	0.56	0.61	1.00	0.05	0.00
			FN	0.00	0.73	0.06	0.15	0.12	0.37	0.27	0.00	0.93	1.00
χ^2_3	$K = 30$	$\rho = 0$	FP	-	-	-	-	-	-	-	-	-	-
			FN	0.00	0.13	0.04	0.17	0.13	0.32	0.29	0.00	0.71	0.87
		$\rho = 0.8$	FP	-	-	-	-	-	-	-	-	-	-
			FN	0.00	0.60	0.06	0.15	0.12	0.36	0.29	0.00	0.93	1.00
	$K = 15$	$\rho = 0$	FP	1.00	0.73	0.94	0.81	0.84	0.61	0.63	1.00	0.15	0.04
			FN	0.00	0.25	0.05	0.15	0.13	0.31	0.27	0.00	0.75	0.93
		$\rho = 0.8$	FP	1.00	0.47	0.94	0.79	0.82	0.59	0.62	1.00	0.07	0.00
			FN	0.00	0.52	0.05	0.15	0.12	0.33	0.27	0.00	0.90	1.00
	$K = 5$	$\rho = 0$	FP	1.00	0.75	0.93	0.79	0.83	0.63	0.61	1.00	0.16	0.04
			FN	0.00	0.23	0.04	0.15	0.11	0.28	0.28	0.00	0.70	0.91
		$\rho = 0.8$	FP	1.00	0.39	0.93	0.78	0.83	0.59	0.62	1.00	0.09	0.00
			FN	0.00	0.61	0.06	0.18	0.12	0.35	0.29	0.00	0.88	1.00

TABLE 1: *Estimates of FP and FN rates for the settings with Gaussian response,* $n_i = 10$.

units varies: $K \in \{30, 15, 5\}$. Settings with and without level 2 endogeneity are considered ($\rho = 0.8$ vs. $\rho = 0.0$). Moreover, the number of first level observations is varied; it is either $n_i = 10$ or $n_i = 5$. Since the variance of the responses determines the effective degrees of freedom, it has to be chosen carefully. We used the standard deviation $\sigma_\varepsilon = 6$, which yields effective degrees of freedom equal to 15.20 in the mixed model with $n_i = 10$, and equal to 10.35 in the mixed model with $n_i = 5$. Thus, one is not too close to the fixed model ($\sigma_0^2 \to \infty$) but far away from the case without variation of the intercept.

As discrete distribution, we use the binomial distribution. The generation of the predictors is roughly the same, but some parameters are changed. The slope parameter is chosen as $\beta_1 = 0.3$ and $\beta_{i0} \sim N(\mu_0 = -0.3, \sigma_0^2 = 4)$ or $b_{i0} \sim \chi^2_3$; in the latter case, β_{i0} is centered such that $\mu_0 = 1$. Because for a binomial model, $n = 30$ is huge, estimates for the unrestricted fixed effects model are quite unstable or do not exist; therefore, they are

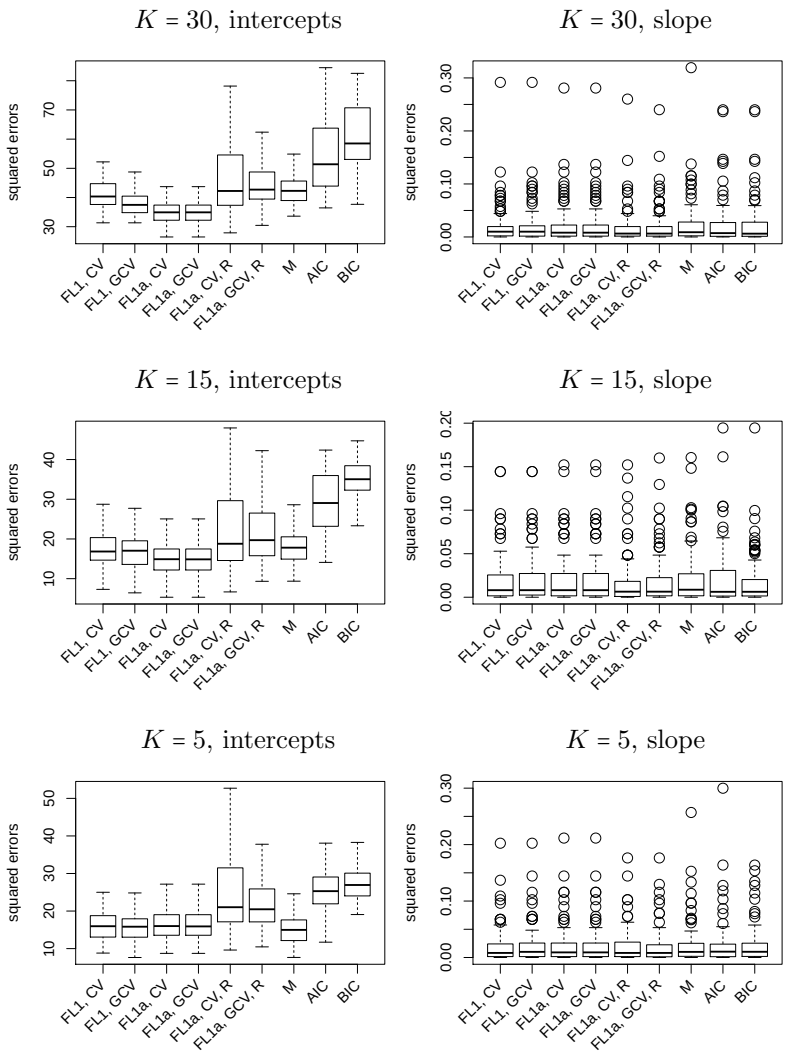

FIGURE 3: *Squared errors for the settings with binomial response and* $\beta_{i0} \sim \chi^2_3$ *(B*χ^2*). The number of clusters K varies with the rows. The left panel relates to the intercepts, the right panel to the slopes.* $\rho = 0.0$, $n_i = 10$.

omitted. Accordingly, if adaptive weights are used, they do not rely on the unrestricted estimates but on an estimate obtained with a small ridge penalty.

For each setting, the mixed model approach, the finite mixture approach of Section 2.4 and the proposed penalized group-specific models are compared. For the penalized approaches, the tuning parameter λ is chosen by 5-fold cross-validation with the deviance (dev) as loss criterion. As binomial settings are usually more sensitive to the selection of folds, the GCV criterion is considered as an alternative. The random effects model are estimated by a restricted maximum likelihood approach (REML) implemented in the `R` package `lme4` (function `lmer`; R Core Team, 2014; Bates et al., 2013). The finite mixture models are estimated by `R` package `flexmix` (Grün and Leisch, 2008a).

Accuracy of the estimation of parameters is measured in terms of the mean squared error (MSE) of coefficients of all $n_{rep} = 100$ replications. In the case of mixed models, the MSEs relate to the sum of fixed and random intercepts. If second level units are clustered, the

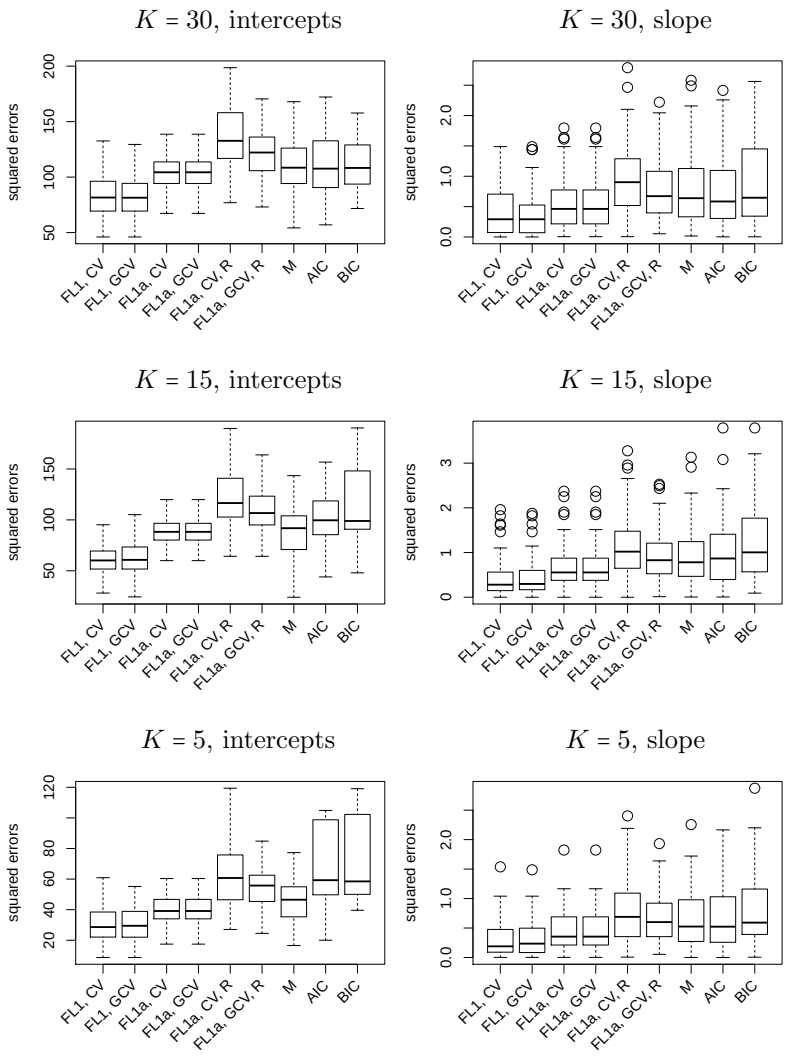

FIGURE 4: *Squared errors for the settings with binomial response and* $\beta_{i0} \sim \chi^2_3$ *(B*χ^2*). The number of clusters K varies with the rows. The left panel relates to the intercepts, the right panel to the slopes.* $\rho = 0.8$, $n_i = 10$.

"right" units should be merged; that is, the rate of falsely fused units should be low (false negatives/FN). The rate of units that should be in one cluster but are not (false positives/FP) should be small likewise; however, high FP are assessed less severe than FN. Of course, when the second level units are not clustered, FP rates are not defined.

4.2 Results

In this section, we present the results of selected scenarios; the results of other scenarios are available in the Appendix. In the Figures to follow, "F" stands for the fixed effects model, "FL1" denotes the L_1-penalized estimates and "FL1a" the adaptive L_1-penalized estimates. "M" stands for the mixed model, "AIC" and "BIC" for the finite mixture models with the respective model selection criterion. "CV" indicates that the penalty parameter is chosen by the 5-fold cross-validation; "GCV" denotes the use of the GCV criterion. If there is an additional "R", the penalty parameter is chosen with an additional

Random Intercepts				FL1, CV	FL1, GCV	FL1a, CV	FL1a, GCV	FL1a, CV, R	FL1a, GCV, R	M	AIC	BIC
Gaussian	$K = 30$	$\rho = 0$	FP	-	-	-	-	-	-	-	-	-
			FN	0.04	0.04	0.08	0.08	0.17	0.13	0.00	0.39	0.47
		$\rho = 0.8$	FP	-	-	-	-	-	-	-	-	-
			FN	0.14	0.03	0.15	0.15	0.47	0.22	0.00	0.71	0.88
	$K = 15$	$\rho = 0$	FP	0.75	0.77	0.65	0.65	0.49	0.60	1.00	0.19	0.13
			FN	0.08	0.07	0.11	0.11	0.22	0.14	0.00	0.39	0.48
		$\rho = 0.8$	FP	0.87	0.89	0.67	0.67	0.44	0.58	1.00	0.20	0.12
			FN	0.04	0.02	0.13	0.13	0.32	0.18	0.00	0.49	0.66
	$K = 5$	$\rho = 0$	FP	0.82	0.83	0.70	0.71	0.57	0.65	1.00	0.21	0.12
			FN	0.03	0.02	0.05	0.05	0.10	0.06	0.00	0.27	0.36
		$\rho = 0.8$	FP	0.88	0.88	0.68	0.68	0.45	0.61	1.00	0.18	0.14
			FN	0.02	0.02	0.12	0.12	0.25	0.14	0.00	0.37	0.45
χ^2_3	$K = 30$	$\rho = 0$	FP	-	-	-	-	-	-	-	-	-
			FN	0.07	0.06	0.11	0.11	0.23	0.14	0.00	0.46	0.56
		$\rho = 0.8$	FP	-	-	-	-	-	-	-	-	-
			FN	0.03	0.03	0.14	0.14	0.33	0.20	0.00	0.50	0.59
	$K = 15$	$\rho = 0$	FP	0.83	0.85	0.72	0.72	0.56	0.66	1.00	0.19	0.14
			FN	0.06	0.05	0.10	0.10	0.20	0.13	0.00	0.42	0.50
		$\rho = 0.8$	FP	0.87	0.87	0.67	0.67	0.46	0.60	1.00	0.19	0.13
			FN	0.03	0.02	0.13	0.13	0.32	0.17	0.00	0.48	0.61
	$K = 5$	$\rho = 0$	FP	0.82	0.85	0.74	0.74	0.53	0.67	1.00	0.19	0.14
			FN	0.05	0.04	0.07	0.07	0.20	0.10	0.00	0.38	0.46
		$\rho = 0.8$	FP	0.89	0.89	0.71	0.71	0.51	0.64	1.00	0.19	0.13
			FN	0.02	0.02	0.11	0.11	0.26	0.16	0.00	0.45	0.58

TABLE 2: *Estimates of FP and FN rates for the settings with binomial response,* $n_i = 10$.

refit in the cross-validation procedure. When evaluating the accuracy of estimation it is important to distinguish between the parameters β_{i0}, which represent the heterogeneity, and the coefficients on covariates. The first ones are referred to as "intercepts" whereas the second ones are referred to as "slope".

Figures 1 and 2 show the boxplots of the mean squared errors for Gaussian response and symmetrically distributed group-specific effects β_{i0} with $n_i = 10$ (GN). In Figure 1 no correlation between the covariates and the heterogeneity parameters is assumed whereas in Figure 2 the correlation is 0.8. It is immediately seen from the right column of Figure 1 that in the case of no correlation all the methods show comparable results as far as the estimation of the slope parameter is concerned. However, there is some variation in the MSEs for the heterogeneity parameters (left column). It is remarkable that the mixed model performs very well, also in the case where clusters of parameters are present (sec-

ond and third row). The fixed model without regularization performs poorly but with regularization the performance is comparable to the mixed model; in particular the approaches without adaptive weights perform rather well. Moreover, the mixed model is not affected by truly skew distributed random intercepts (for illustration, see, Appendix A). In contrast the finite mixture model performs badly for all settings.

The picture changes dramatically if correlation is present. It is seen from Figure 2 that in this case, the performance varies strongly over methods in terms of the MSEs for the slope (right column). Now, the mixed model performs as bad as the finite mixture models whereas the fixed effects model and the regularized versions perform very well. A similar picture results for the intercepts (left column). The mixed model and the finite mixture models show poor performance, but the regularized fixed effects model perform well. The simple fixed model without regularization, although being competitive for the slopes, does not perform well for the intercepts.

The FP and the FN rates for all settings with Gaussian responses with $n_i = 10$ are shown in Table 1. By construction the mixed model and the fixed effects model without regularization have false positive rate 1 and false negative rate 0 when clusters are present. Regularized estimates without refit tend to relatively large false positive rates and very small false negative rates. The methods with refit aim at a different compromise with smaller false positive rates and larger false negative rates. The mixed model (AIC and BIC) tends to build very few but large clusters entailing very small positive rates but very large false negative rates.

The results for the settings with binomial responses are shown in Figures 3–4 and in Table 2. We focus on the setting with $n_i = 10$ and skew random effects (Bχ^2). In Figure 3, no correlation between the covariates and the heterogeneity parameters is assumed. As in the case of a metric response the slopes are estimated equally well by all approaches and the best estimates of intercepts are found for the adaptively penalized approaches. The simple fixed effects model is not given because it performs too poorly. The results with correlation are shown in Figure 4. In contrast to the case of metric responses, for binary responses the L1 penalized approaches without weights and refit perform best – followed by the adaptively weighted penalized approaches, the random effects models and the finite mixture models. The same pattern is observed for the slopes. Overall the differences between methods are less distinct than for metric responses.. The results for the settings with symmetrically distributed random intercepts are given in the Appendix: The results are similar but in contrast to the settings with Gaussian responses, the results for binomial responses seem to be affected by both, correlations and skewly distributed random intercepts.

Table 2 shows the FP and the FN rates for symmetric and skew random intercepts, $n_i = 10$. Again, the adaptively weighted approaches perform best in the sense that both the FN and the FP rates are relatively low. Overall, the clustering performance is better

than for Gaussian responses.

If one is especially interested in the detection of second level units, the clustering performance for both Gaussian and binomial responses is considerably improved by an additional refit in the cross-validation procedure. However, with an additional refit, the estimation accuracy may suffer – it is only recommended in combination with adaptive weights and when the focus is on the clustering performance. Detailed results for all settings can be found in the Appendix.

5 Extension: Group Specific Models with Vector Fused Penalties

Before looking at applications, an extended version of fusion penalties is considered that is helpful if more than one parameter is expected to be group-specific. The penalty for group-specific models allows for different clusters of second level units in different components of the predictor $\boldsymbol{z}_{ij}$. Thus, for each component of the predictor, one obtains a disjunct partition $C_1^{(s)}, \ldots, C_K^{(s)}$, where s refers to the component in $\boldsymbol{z}_{ij}$. It depends on the application, if this is desirable. If one wants one consistent partition that is based on the whole vector $\boldsymbol{z}_{ij}$, a modified penalty has to be used. Let again variables that are in $\boldsymbol{z}_{ij}$, be excluded in $\boldsymbol{x}_{ij}$. Then, a penalty that fuses the second level units *simultaneously* is

$$J(\boldsymbol{\alpha}) = \sum_{r>m} \|\boldsymbol{\beta}_r - \boldsymbol{\beta}_m\|_2 \,, \tag{9}$$

where $\|\boldsymbol{\xi}\|_2 = \{\xi_1^2 + \cdots + \xi_q^2\}^{1/2}$ denotes the L_2-norm of a q-dimensional vector $\boldsymbol{\xi}$.

The penalty enforces that the whole vectors of parameters $\boldsymbol{\beta}_r$ and $\boldsymbol{\beta}_m$ are fused. In contrast to the componentwise penalty (7), the penalty yields uniform clusters. The effect of the penalty is that the group-specific coefficients $\boldsymbol{\beta}_r$ are shrunk towards each other and one obtains only one partition $C_1, \ldots, C_K$ of C that is based on the whole vector $\boldsymbol{z}_{ij}$. The approach works in a similar way as the group Lasso proposed by Yuan and Lin (2006). However, the group Lasso refers to the simultaneous selection of a group of parameters, whereas penalty (9) refers to the fusion of a set of coefficients.

As for componentwise fusion penalties, one can use an adaptive version of the penalty, see, for example Wang and Leng (2008). It is given by $J(\boldsymbol{\alpha}) = \sum_{r>m} w_{rm} \|\boldsymbol{\beta}_r - \boldsymbol{\beta}_m\|_2$, where $w_{rm} = ||\tilde{\boldsymbol{\beta}}_r - \tilde{\boldsymbol{\beta}}_m||_2^{-1}$ with $\sqrt{n}$-consistent estimates $\tilde{\boldsymbol{\beta}}_r$, $\tilde{\boldsymbol{\beta}}_m$.

Interestingly, the component-wise fusion penalty (7) can be written as

$$J(\boldsymbol{\alpha}) = \sum_{r>m} \|\boldsymbol{\beta}_r - \boldsymbol{\beta}_m\|_1 \,,$$

where $\|\boldsymbol{\xi}\|_1 = |\xi_1| + \cdots + |\xi_q|$ denotes the L_1-norm of a q-dimensional vector $\boldsymbol{\xi}$. Thus, penalties (7) and (9) basically differ in the applied norm. Of course, it is possible to combine the penalties in specific applications. In the simplest case, let $\boldsymbol{\beta}_i$ be partitioned

Coefficients		Fixed Effects	Random Intercept Model	Fixed Effects Pen. (8)	Discrete Mixture Model AIC	Discrete Mixture Model BIC
Center-specific Intercept	$\beta_{15,0}$	-1.4782	-1.5520	-1.70	-1.5688	-1.7388
	$\beta_{12,0}$	-1.5644	-1.6053			
	$\beta_{16,0}$	-1.5999	-1.6494			
	$\beta_{20,0}$	-1.6038	-1.6524			
	$\beta_{7,0}$	-1.8832	-1.8917	-1.92	-1.9171	
	$\beta_{17,0}$	-2.0801	-2.1065	-2.35	-2.3873	-2.3793
	$\beta_{9,0}$	-2.0910	-2.1079			
	$\beta_{8,0}$	-2.2083	-2.2133	-2.36		
	$\beta_{3,0}$	-2.2370	-2.2575	-2.37		
	$\beta_{21,0}$	-2.2832	-2.2859			
	$\beta_{2,0}$	-2.3059	-2.3097			
	$\beta_{6,0}$	-2.3113	-2.3162			
	$\beta_{10,0}$	-2.3840	-2.3832			
	$\beta_{11,0}$	-2.4278	-2.4239			
	$\beta_{1,0}$	-2.4798	-2.4144	-2.38		
	$\beta_{5,0}$	-2.5015	-2.4881			
	$\beta_{4,0}$	-2.5189	-2.5151			
	$\beta_{14,0}$	-2.7862	-2.7670	-2.72	-2.9626	-2.9628
	$\beta_{18,0}$	-3.0433	-2.8803	-2.87		
	$\beta_{22,0}$	-3.0610	-3.0122			
	$\beta_{13,0}$	-3.1155	-3.0020			
	$\beta_{19,0}$	-3.4942	-3.1536	-2.89		
Treatment	β_T	-0.1305	-0.1305	-0.13	-0.1292	-0.1291

TABLE 3: *Estimates for the beta blocker data. Intercept-coefficients are ordered such that their structure becomes obvious. Presented intercept-coefficients of the mixed model are the sum of the fixed and the random effects. Horizontal lines denote clusters of coefficients.*

into $\boldsymbol{\beta}_i^T = (\boldsymbol{\beta}_{i1}^T, \boldsymbol{\beta}_{i2}^T)$ and use

$$J(\boldsymbol{\alpha}) = \sum_{r>m} \|\boldsymbol{\beta}_{r1} - \boldsymbol{\beta}_{m1}\|_s + \|\boldsymbol{\beta}_{r2} - \boldsymbol{\beta}_{m2}\|_t ,$$

where $s, t \in \{1, 2\}$. One can, for example, employ the L_1-norm (or equivalently the L_2-norm) for the intercept and the L_2-norm for the remaining predictors. Again, computational issues are met by local quadratic approximations as proposed by Oelker and Tutz (2013).

6 Mortality after Myocardial Infarction

In the beta blocker example considered in the introduction, one models the mortality rate after myocardial infarction depending on the study center and the treatment group

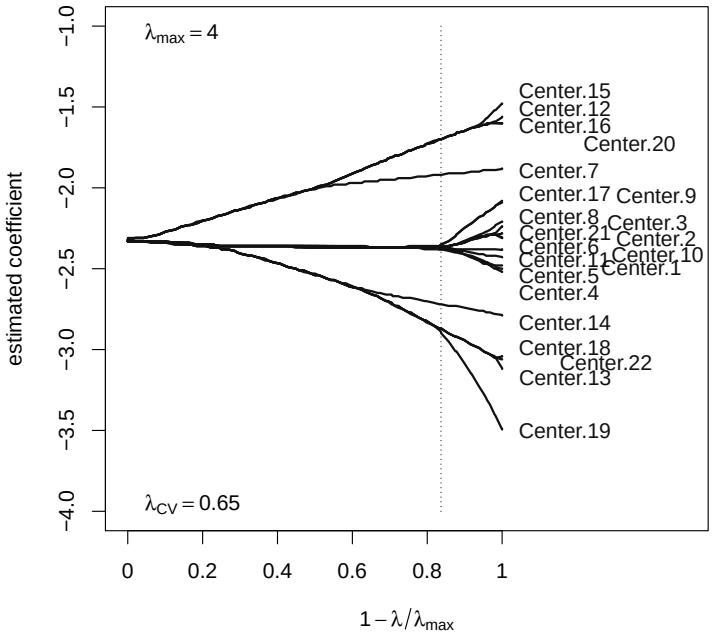

FIGURE 5: *Coefficient paths for the beta blocker data. The very right end of the figure relates to ML estimates; that is, to* $\lambda = 0$*. The left end relates to the minimal value of* λ *giving maximal penalization; in this case* $\lambda = 4$*.*

assigned to a patient. A classical model that has been used on this data set is the random intercept model, which has the form $\text{logit}\,\mathbb{P}(y_{ij} = 1) = \beta_0 + b_{i0} + \beta_T \cdot \text{Treatment}_{ij}$, where the random effects b_{i0} follow a normal distribution, and where $\text{Treatment}_{ij} \in \{-1, 1\}$ codes the treatment in hospital i for patient j. The model with group-specific intercepts has the form

$$\text{logit}\,\mathbb{P}(y_{ij} = 1) = \beta_{i0} + \beta_T \cdot \text{Treatment}_{ij}, \quad i = 1, \ldots, 22 \text{ Centers}, \tag{10}$$

where β_{i0} are fixed unknown parameters. Table 3 shows the results for the random intercept model, the ML estimates of the (unpenalized) group-specific model and the estimates of the group-specific model with the adaptively weighted penalty (8). One can see that the estimates of the random effects model and the fixed effects model are quite similar. The assumption of a normal distribution does not strongly affect the estimates. Figure 5 shows the corresponding coefficient build-ups of the penalized approach that enforces clustering of hospitals; the dotted line denotes the tuning parameter selected by the GCV criterion with an additional refit ($\lambda_{CV} = 0.65$). There are basically five clusters of hospitals that are to be distinguished in terms of the basic risk captured by the intercepts, although numerically one obtains more clusters. But clusters with effects -2.35 and -2.36 can hardly be considered as having different effects.

The finite mixture model of Grün and Leisch (2008b), which in the simulation study had a tendency to identify a too small number of clusters, found three (BIC) and four (AIC) clusters. The last two columns of Table 3 present the respective estimates and clusters. The estimates are similar and the structure found in the hospitals, corresponds to those

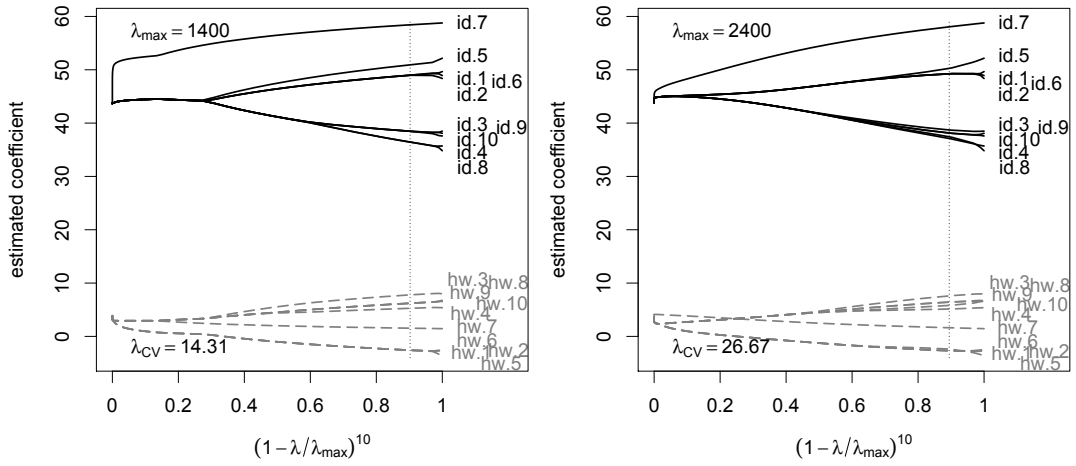

FIGURE 6: *Coefficient paths for the NELS:88 data with group-specific intercepts and group-specific covariate homework; left panel: componentwise penalty (8); right panel: simultaneous clusters penalty (9) with adaptive weights.*

of the penalized fixed effects model.

The fitted treatment effect has approximately the same size in all models. But only the regularization approach combines data driven clustering with stable results.

Note that the predictor in model (10) corresponds to a generalized linear model with the nominal covariates Center and Treatment. However, when the Centers are coded as a nominal covariate, the choice of the reference category affects the estimate of a penalized nominal predictor crucially. With group-specific intercepts for the hospitals, there is no need for a reference category; all hospitals are penalized in the same way. Moreover, there is an intrinsic interpretation for the group-specific intercepts, whereas a nominal predictor always draws comparisons with the reference category which is arbitrary.

7 Analysis of Selected Schools in the National Education Longitudinal Study of 1988

In a second example, we analyze $n = 10$ selected schools out of the National Education Longitudinal Study of 1988 (NELS:88, Curtin et al. (2002)). The data contains information on eighth-graders surveyed in 1988 in different schools in the United States and is a subsample of a nation-wide longitudinal study. The number of pupils per school varies. We use the standardized mathematics score (y_{ij}, measured between 0 and 100, the higher the better), the time in hours spent on math homework weekly (hw) and the school (id) of 260 pupils. We investigate if the math skills and the impact of the duration of the homework do differ over the schools when explaining the mathematics score. The mathematics score is assumed to be Gaussian and we fit the linear model

$$y_{ij} = \beta_{i0} + \beta_{i1} \cdot \text{hw}_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, 10 \text{ Schools}. \tag{11}$$

Penalty	Coeff.	School (id) 1	2	3	4	5	6	7	8	9	10
none	β_{i0}	49.66	48.36	38.49	35.70	52.17	48.91	58.78	34.81	38.09	37.59
	β_{i1}	-2.80	-2.60	7.99	5.38	-3.60	-2.37	1.45	6.80	6.68	6.54
(8)	β_{i0}	48.98	48.97	38.47	36.44	50.92	48.98	58.43	36.44	38.47	38.47
	β_{i1}	-2.59	-2.59	7.79	5.27	-2.59	-2.59	1.54	6.21	6.21	6.21
(9)	β_{i0}	49.24	49.24	38.71	37.11	50.26	49.24	58.05	37.43	38.14	38.14
	β_{i1}	-2.69	-2.69	7.59	5.13	-2.38	-2.69	1.62	5.81	6.43	6.43

TABLE 4: *Estimated coefficients for the NELS:88 data obtained with model (11) and different penalization strategies.*

In contrast to the example in Section 6, we assume group-specific intercepts β_{i0} and group-specific slopes β_{i1}. Hence, we can compare the componentwise penalty (8) and penalty (9) for simultaneous clusters, both adaptively weighted. Figure 6 shows the resulting coefficient paths. The left panel refers to penalty (8); that is, the pairwise differences of the group-specific intercepts are penalized by the adaptive Lasso and so are the differences of the group-specific slopes. In the right panel, the pairwise difference of the group-specific intercepts and slopes are penalized simultaneously by penalty (9) with adaptive weights. In both panels, a dotted line marks the optimal results according to 5-fold cross-validation with the MSE as loss criterion. As the range of the coefficients is large, in both panels, the minimal tuning parameter giving maximal penalization is considerably higher as in Figure 5. However, for the optimal tuning parameters, with both penalties, some schools are clustered. With penalty (8), one obtains two separate partitions of schools. Regarding the intercepts, schools $\{4, 8\}$, $\{3, 9, 10\}$ and $\{1, 6\}$ are merged to clusters; regarding the slopes, schools $\{1, 2, 5, 6\}$ and $\{8, 9, 10\}$ are merged while the other schools have individual effects. With penalty (9), one obtains one uniform partition: schools $\{1, 2, 6\}$ and $\{9, 10\}$ are found to have similar effects for the intercepts and the slopes while the other schools form its own clusters. Table 4 gives the exact results. With both penalties, an interesting effect is seen: schools with relatively low intercepts tend to have larger effects regarding the homework and vice versa. The effect is seen, for example for the schools 3 and 5. In school 3, the average math skills without homework are relatively low while the effect of the homework is high. In school 5, the group-specific intercept is high; the impact of the homework is actually negative. A possible explanation is that in schools with higher average math skills, the time required for the homework might be an indicator for inertial pupils; while in schools with lower initial skills, the time spent on math homework might be a surrogate for the extent of the homework. Comparing the two approaches by the residual deviances, the model with penalty (8) performs slightly better.

8 Concluding Remarks

We compared three different approaches that take the heterogeneity in hierarchical data into account: random effects models, finite mixture models and fixed effects models with regularization. We were especially interested in situations where the second level units are clustered.

Mixed models do not allow for clustered second level units and the observed second level units are assumed to be a sample of an underlying population. In particular when the data show level 2 endogeneity the performance of the estimates suffers because the estimates are biased. Finite mixture models assume that the effects of the second level units are drawn from an unknown finite set of effects. The performance is acceptable only for settings with very few clusters in the underlying data structure.

In contrast, fixed effects models rely only on the observed second level units. With the proposed fused Lasso penalty on group-specific coefficients, one obtains clusters. Although identification of clusters is not perfect, the estimation accuracy is much better than for the maximum likelihood estimate – even when there are no clusters in the data. In situations where the focus is on the second level units, the proposed Lasso-type penalty has substantial advantages.

A Numerical Results

The Appendix presents the detailed results of the numerical experiments conducted for Section 4. Tables 5–8 show the results for Gaussian responses. Tables 9–12 the results for binomial responses. If there is an additional "R", the penalty parameter is chosen with an additional refit in the cross-validation procedure.

$\beta_{i0} \sim N(1,4)$		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	Fixed	100.59	0.08		0.00	107.23	0.21		0.00
	Fixed L1, CV	75.03	0.06		0.06	130.65	2.58		0.73
	Fixed L1, GCV	70.21	0.06		0.04	96.00	1.09		0.07
	Fixed L1, adapt., CV	81.42	0.07		0.12	87.83	0.40		0.22
	Fixed L1, adapt., GCV	81.22	0.07		0.11	83.03	0.37		0.14
	Fixed L1, adapt., CV, R	81.10	0.06		0.23	81.98	0.97		0.40
	Fixed L1, adapt., GCV, R	99.54	0.06		0.25	95.49	1.49		0.29
	Mixed	71.04	0.06		0.00	124.92	2.44		0.00
	Finite AIC	124.80	0.07		0.60	132.84	2.66		0.95
	Finite BIC	169.01	0.07		0.81	132.11	2.77		1.00
15	Fixed	100.60	0.08	1.00	0.00	106.59	0.21	1.00	0.00
	Fixed L1, CV	63.19	0.06	0.52	0.47	103.97	1.86	0.25	0.74
	Fixed L1, GCV	51.74	0.07	0.93	0.05	84.38	0.86	0.92	0.07
	Fixed L1, adapt., CV	69.42	0.07	0.79	0.17	81.99	0.35	0.74	0.21
	Fixed L1, adapt., GCV	70.21	0.07	0.83	0.13	77.95	0.31	0.83	0.14
	Fixed L1, adapt., CV, R	56.22	0.06	0.61	0.35	77.14	0.87	0.55	0.39
	Fixed L1, adapt., GCV, R	56.86	0.06	0.62	0.31	81.58	1.08	0.61	0.30
	Mixed	48.51	0.07	1.00	0.00	101.27	1.91	1.00	0.00
	Finite AIC	71.43	0.07	0.11	0.84	105.97	2.08	0.04	0.95
	Finite BIC	67.90	0.06	0.02	0.98	105.06	2.16	0.00	1.00
5	Fixed	100.60	0.08	1.00	0.00	106.70	0.21	1.00	0.00
	Fixed L1, CV	68.73	0.07	0.70	0.27	113.98	2.00	0.27	0.73
	Fixed L1, GCV	61.54	0.07	0.94	0.04	88.91	0.98	0.93	0.06
	Fixed L1, adapt., CV	75.39	0.07	0.80	0.13	83.09	0.35	0.79	0.15
	Fixed L1, adapt., GCV	76.42	0.07	0.84	0.11	82.85	0.31	0.82	0.12
	Fixed L1, adapt., CV, R	69.10	0.06	0.64	0.27	84.10	0.94	0.56	0.37
	Fixed L1, adapt., GCV, R	79.37	0.06	0.61	0.27	90.09	1.33	0.61	0.27
	Mixed	58.95	0.07	1.00	0.00	106.76	1.87	1.00	0.00
	Finite AIC	99.80	0.06	0.19	0.68	116.29	2.02	0.05	0.93
	Finite BIC	100.96	0.06	0.04	0.92	115.35	2.34	0.00	1.00

TABLE 5: *Results for the settings with Gaussian response,* $\beta_{i0} \sim N(1,4)$, $n_i = 10$.

$\beta_{i0} \sim \chi^2_3$		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	Fixed	100.59	0.08		0.00	107.58	0.21		0.00
	Fixed L1, CV	63.26	0.07		0.13	152.42	2.69		0.60
	Fixed L1, GCV	61.28	0.06		0.04	104.45	1.10		0.06
	Fixed L1, adapt., CV	70.25	0.07		0.17	84.49	0.40		0.15
	Fixed L1, adapt., GCV	71.23	0.07		0.13	84.18	0.31		0.12
	Fixed L1, adapt., CV, R	64.20	0.07		0.32	94.60	1.11		0.36
	Fixed L1, adapt., GCV, R	67.74	0.06		0.29	103.27	1.56		0.29
	Mixed	67.49	0.07		0.00	142.41	2.85		0.00
	Finite AIC	95.60	0.07		0.71	154.66	3.09		0.93
	Finite BIC	141.13	0.07		0.87	154.41	3.23		1.00
15	Fixed	100.59	0.08	1.00	0.00	107.47	0.21	1.00	0.00
	Fixed L1, CV	59.45	0.07	0.73	0.25	182.86	3.28	0.47	0.52
	Fixed L1, GCV	57.94	0.06	0.94	0.05	114.49	1.45	0.94	0.05
	Fixed L1, adapt., CV	70.89	0.07	0.81	0.15	86.28	0.31	0.79	0.15
	Fixed L1, adapt., GCV	72.59	0.06	0.84	0.13	86.71	0.29	0.82	0.12
	Fixed L1, adapt., CV, R	62.93	0.06	0.61	0.31	91.39	0.99	0.59	0.33
	Fixed L1, adapt., GCV, R	65.92	0.05	0.63	0.27	104.79	1.45	0.62	0.27
	Mixed	59.48	0.07	1.00	0.00	172.74	3.54	1.00	0.00
	Finite AIC	98.97	0.07	0.15	0.75	211.47	4.03	0.07	0.90
	Finite BIC	100.45	0.07	0.04	0.93	211.69	4.48	0.00	1.00
5	Fixed	100.59	0.08	1.00	0.00	106.72	0.21	1.00	0.00
	Fixed L1, CV	63.66	0.07	0.75	0.23	81.90	0.76	0.39	0.61
	Fixed L1, GCV	57.58	0.06	0.93	0.04	66.92	0.44	0.93	0.06
	Fixed L1, adapt., CV	70.31	0.07	0.79	0.15	74.80	0.20	0.78	0.18
	Fixed L1, adapt., GCV	73.07	0.06	0.83	0.11	77.55	0.16	0.83	0.12
	Fixed L1, adapt., CV, R	62.05	0.07	0.63	0.28	63.62	0.43	0.59	0.35
	Fixed L1, adapt., GCV, R	67.88	0.06	0.61	0.28	68.55	0.58	0.62	0.29
	Mixed	59.78	0.07	1.00	0.00	69.25	0.70	1.00	0.00
	Finite AIC	100.53	0.07	0.16	0.70	85.09	0.89	0.09	0.88
	Finite BIC	101.74	0.07	0.04	0.91	83.34	0.98	0.00	1.00

TABLE 6: *Results for the settings with Gaussian response,* $\beta_{i0} \sim \chi^2_3$, $n_i = 10$.

$\beta_{i0} \sim N(1,4)$		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	Fixed	211.68	0.14		0.00	227.37	0.38		0.00
	Fixed L1, CV	116.26	0.10		0.41	123.60	2.21		0.80
	Fixed L1, GCV	110.70	0.11		0.05	138.87	1.16		0.07
	Fixed L1, adapt., CV	140.18	0.11		0.18	135.57	0.50		0.23
	Fixed L1, adapt., GCV	150.01	0.11		0.13	149.85	0.40		0.14
	Fixed L1, adapt., CV, R	114.44	0.12		0.38	119.18	0.91		0.43
	Fixed L1, adapt., GCV, R	111.48	0.10		0.31	116.53	1.31		0.34
	Mixed	96.58	0.11		0.00	122.92	2.27		0.00
	Finite AIC	135.56	0.12		0.85	125.04	2.35		0.96
	Finite BIC	127.63	0.11		0.99	123.33	2.46		1.00
15	Fixed	211.68	0.14	1.00	0.00	227.55	0.38	1.00	0.00
	Fixed L1, CV	125.50	0.12	0.59	0.40	121.54	2.13	0.19	0.81
	Fixed L1, GCV	116.01	0.12	0.94	0.05	138.83	1.18	0.95	0.07
	Fixed L1, adapt., CV	145.39	0.12	0.81	0.16	139.17	0.52	0.74	0.24
	Fixed L1, adapt., GCV	149.66	0.12	0.85	0.13	149.10	0.43	0.85	0.14
	Fixed L1, adapt., CV, R	121.62	0.10	0.66	0.31	118.65	1.16	0.49	0.48
	Fixed L1, adapt., GCV, R	119.03	0.11	0.65	0.29	114.72	1.29	0.63	0.31
	Mixed	99.81	0.13	1.00	0.00	120.46	2.14	1.00	0.00
	Finite AIC	148.00	0.12	0.14	0.81	122.16	2.28	0.03	0.97
	Finite BIC	140.59	0.12	0.01	0.99	120.83	2.33	0.00	1.00
5	Fixed	211.68	0.14	1.00	0.00	226.37	0.38	1.00	0.00
	Fixed L1, CV	91.91	0.12	0.40	0.59	118.93	2.10	0.18	0.82
	Fixed L1, GCV	103.26	0.11	0.94	0.06	134.74	1.08	0.94	0.07
	Fixed L1, adapt., CV	134.73	0.12	0.78	0.19	141.33	0.45	0.76	0.21
	Fixed L1, adapt., GCV	144.08	0.12	0.83	0.14	146.58	0.40	0.84	0.14
	Fixed L1, adapt., , CVR	101.19	0.11	0.59	0.37	113.66	0.91	0.53	0.43
	Fixed L1, adapt., GCV, R	97.70	0.11	0.65	0.30	113.06	1.25	0.63	0.31
	Mixed	80.51	0.12	1.00	0.00	115.93	1.95	1.00	0.00
	Finite AIC	98.20	0.12	0.09	0.88	119.91	2.25	0.04	0.96
	Finite BIC	90.09	0.11	0.00	1.00	117.35	2.25	0.00	1.00

TABLE 7: *Results for the settings with Gaussian response,* $\beta_{i0} \sim N(1,4)$, $n_i = 5$.

$\beta_{i0} \sim \chi^2_3$		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	Fixed	211.68	0.14		0.00	226.58	0.38		0.00
	Fixed L1, CV	98.35	0.11		0.36	287.29	4.84		0.67
	Fixed L1, GCV	106.86	0.11		0.05	203.37	2.24		0.05
	Fixed L1, adapt., CV	136.76	0.09		0.18	160.75	0.72		0.18
	Fixed L1, adapt., GCV	149.79	0.09		0.13	160.97	0.63		0.12
	Fixed L1, adapt., CV, R	107.07	0.10		0.33	172.88	1.99		0.39
	Fixed L1, adapt., GCV, R	104.50	0.10		0.33	181.41	2.76		0.28
	Mixed	101.08	0.12		0.00	269.26	5.30		0.00
	Finite AIC	158.18	0.12		0.83	291.85	5.72		0.93
	Finite BIC	153.81	0.13		0.98	290.34	6.04		1.00
15	Fixed	211.67	0.14	1.00	0.00	225.78	0.38	1.00	0.00
	Fixed L1, CV	101.66	0.11	0.66	0.30	305.92	4.75	0.42	0.58
	Fixed L1, GCV	113.22	0.11	0.94	0.05	207.59	2.27	0.94	0.05
	Fixed L1, adapt., CV	140.25	0.10	0.78	0.18	171.49	0.74	0.76	0.20
	Fixed L1, adapt., GCV	150.37	0.10	0.85	0.13	169.86	0.70	0.84	0.12
	Fixed L1, adapt., CV, R	113.10	0.09	0.58	0.38	166.74	1.73	0.56	0.37
	Fixed L1, adapt., GCV, R	106.85	0.10	0.64	0.30	190.39	2.92	0.64	0.29
	Mixed	113.89	0.12	1.00	0.00	272.15	4.95	1.00	0.00
	Finite AIC	176.66	0.11	0.14	0.80	311.39	5.84	0.08	0.91
	Finite BIC	176.88	0.14	0.02	0.96	310.41	6.02	0.00	1.00
5	Fixed	211.68	0.14	1.00	0.00	225.95	0.38	1.00	0.00
	Fixed L1, CV	129.26	0.10	0.66	0.32	176.74	3.01	0.27	0.73
	Fixed L1, GCV	118.10	0.11	0.94	0.04	163.23	1.51	0.94	0.06
	Fixed L1, adapt., CV	146.52	0.12	0.81	0.15	154.50	0.69	0.76	0.19
	Fixed L1, adapt., GCV	152.55	0.12	0.85	0.12	157.10	0.49	0.83	0.12
	Fixed L1, adapt., CV, R	126.97	0.10	0.65	0.28	151.50	1.30	0.51	0.43
	Fixed L1, adapt., GCV, R	131.66	0.11	0.64	0.28	153.66	1.78	0.63	0.29
	Mixed	109.32	0.11	1.00	0.00	171.66	2.77	1.00	0.00
	Finite AIC	156.39	0.10	0.14	0.79	181.35	3.12	0.05	0.95
	Finite BIC	153.51	0.10	0.02	0.97	178.47	3.28	0.00	1.00

TABLE 8: *Results for the settings with Gaussian response,* $\beta_{i0} \sim \chi^2_3$, $n_i = 5$.

$\beta_{i0} \sim N(-.3, 4)$		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	Fixed L1, CV	17.04	0.01		0.04	31.89	0.28		0.14
	Fixed L1, GCV	16.98	0.01		0.04	32.26	0.27		0.03
	Fixed L1, adapt., CV	16.18	0.01		0.08	39.64	0.37		0.15
	Fixed L1, adapt., GCV	16.34	0.01		0.08	39.64	0.37		0.15
	Fixed L1, adapt., CV, R	23.10	0.00		0.17	55.15	0.70		0.47
	Fixed L1, adapt., GCV, R	24.23	0.01		0.13	49.32	0.60		0.22
	Mixed	16.61	0.01		0.00	51.31	0.74		0.00
	Finite AIC	32.54	0.01		0.39	63.88	0.80		0.71
	Finite BIC	34.22	0.01		0.47	76.08	1.09		0.88
15	Fixed L1, CV	26.64	0.01	0.75	0.08	45.66	0.34	0.87	0.04
	Fixed L1, GCV	23.15	0.01	0.77	0.07	44.85	0.34	0.89	0.02
	Fixed L1, adapt., CV	25.85	0.01	0.65	0.11	71.92	0.61	0.67	0.13
	Fixed L1, adapt., GCV	25.62	0.01	0.65	0.11	71.92	0.61	0.67	0.13
	Fixed L1, adapt., CV, R	37.98	0.01	0.49	0.22	98.54	0.96	0.44	0.32
	Fixed L1, adapt., GCV, R	35.74	0.01	0.60	0.14	92.54	0.91	0.58	0.18
	Mixed	23.86	0.01	1.00	0.00	76.12	0.84	1.00	0.00
	Finite AIC	45.20	0.01	0.19	0.39	84.88	0.83	0.20	0.49
	Finite BIC	53.95	0.01	0.13	0.48	85.77	0.95	0.12	0.66
5	Fixed L1, CV	17.80	0.01	0.82	0.03	51.64	0.27	0.88	0.02
	Fixed L1, GCV	17.51	0.01	0.83	0.02	51.64	0.27	0.88	0.02
	Fixed L1, adapt., CV	17.30	0.01	0.70	0.05	78.55	0.61	0.68	0.12
	Fixed L1, adapt., GCV	17.30	0.01	0.71	0.05	78.55	0.61	0.68	0.12
	Fixed L1, adapt., CV, R	22.64	0.01	0.57	0.10	108.19	1.06	0.45	0.25
	Fixed L1, adapt., GCV, R	23.80	0.01	0.65	0.06	95.01	0.84	0.61	0.14
	Mixed	17.25	0.01	1.00	0.00	67.50	0.60	1.00	0.00
	Finite AIC	32.43	0.01	0.21	0.27	93.57	0.71	0.18	0.37
	Finite BIC	37.21	0.01	0.12	0.36	93.57	0.80	0.14	0.45

TABLE 9: *Results for the settings with binomial response,* $\beta_{i0} \sim N(-.3, 4)$, $n_i = 10$.

$\beta_{i0} \sim \chi^2_3$		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	Fixed L1, CV	40.34	0.01		0.07	81.59	0.29		0.03
	Fixed L1, GCV	37.48	0.01		0.06	81.41	0.29		0.03
	Fixed L1, adapt., CV	34.92	0.01		0.11	104.34	0.46		0.14
	Fixed L1, adapt., GCV	34.92	0.01		0.11	104.34	0.46		0.14
	Fixed L1, adapt., CV, R	42.24	0.01		0.23	132.70	0.90		0.33
	Fixed L1, adapt., GCV, R	42.71	0.01		0.14	122.15	0.67		0.20
	Mixed	42.28	0.01		0.00	108.48	0.64		0.00
	Finite AIC	51.39	0.01		0.46	107.62	0.58		0.50
	Finite BIC	58.52	0.01		0.56	108.26	0.65		0.59
15	Fixed L1, CV	16.84	0.01	0.83	0.06	60.17	0.28	0.87	0.03
	Fixed L1, GCV	17.04	0.01	0.85	0.05	60.73	0.30	0.87	0.02
	Fixed L1, adapt., CV	14.90	0.01	0.72	0.10	88.25	0.56	0.67	0.13
	Fixed L1, adapt., GCV	14.87	0.01	0.72	0.10	88.25	0.56	0.67	0.13
	Fixed L1, adapt., CV, R	18.79	0.01	0.56	0.20	116.55	1.02	0.46	0.32
	Fixed L1, adapt., GCV, R	19.70	0.01	0.66	0.13	106.71	0.83	0.60	0.17
	Mixed	17.81	0.01	1.00	0.00	91.84	0.78	1.00	0.00
	Finite AIC	29.05	0.01	0.19	0.42	99.60	0.87	0.19	0.48
	Finite BIC	35.04	0.01	0.14	0.50	98.93	1.00	0.13	0.61
5	Fixed L1, CV	15.96	0.01	0.82	0.05	28.61	0.19	0.89	0.02
	Fixed L1, GCV	15.83	0.01	0.85	0.04	29.45	0.24	0.89	0.02
	Fixed L1, adapt., CV	16.00	0.01	0.74	0.07	39.12	0.35	0.71	0.11
	Fixed L1, adapt., GCV	15.91	0.01	0.74	0.07	39.12	0.35	0.71	0.11
	Fixed L1, adapt., CV, R	21.02	0.01	0.53	0.20	60.72	0.69	0.51	0.26
	Fixed L1, adapt., GCV, R	20.47	0.01	0.67	0.10	55.76	0.60	0.64	0.16
	Mixed	14.98	0.01	1.00	0.00	46.50	0.52	1.00	0.00
	Finite AIC	25.30	0.01	0.19	0.38	59.27	0.52	0.19	0.45
	Finite BIC	26.94	0.01	0.14	0.46	58.48	0.59	0.13	0.58

TABLE 10: *Results for the settings with binomial response,* $\beta_{i0} \sim \chi^2_3$, $n_i = 10$.

$\beta_{i0} \sim N(-.3, 4)$		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	Fixed L1, CV	43.14	0.03		0.11	75.43	1.21		0.50
	Fixed L1, GCV	39.94	0.03		0.09	52.66	0.78		0.05
	Fixed L1, adapt., CV	42.54	0.02		0.16	68.20	1.21		0.36
	Fixed L1, adapt., GCV	42.54	0.02		0.16	68.22	1.22		0.37
	Fixed L1, adapt., CV, R	49.83	0.01		0.19	82.89	1.39		0.68
	Fixed L1, adapt., GCV, R	60.92	0.02		0.17	73.12	1.23		0.40
	Mixed	44.03	0.03		0.00	75.26	1.48		0.00
	Finite AIC	63.98	0.02		0.51	84.96	1.46		0.89
	Finite BIC	63.02	0.03		0.52	84.71	1.52		0.98
15	Fixed L1, CV	22.83	0.01	0.80	0.10	97.05	1.07	0.57	0.37
	Fixed L1, GCV	23.14	0.01	0.82	0.08	77.70	0.81	0.88	0.04
	Fixed L1, adapt., CV	21.96	0.01	0.70	0.15	100.36	1.13	0.52	0.33
	Fixed L1, adapt., GCV	21.96	0.01	0.70	0.15	99.77	1.14	0.52	0.32
	Fixed L1, adapt., CV, R	23.48	0.01	0.62	0.21	122.43	1.40	0.24	0.67
	Fixed L1, adapt., GCV, R	27.93	0.01	0.67	0.17	108.69	1.21	0.49	0.36
	Mixed	23.41	0.01	1.00	0.00	106.91	1.51	1.00	0.00
	Finite AIC	34.08	0.02	0.18	0.50	126.51	1.50	0.11	0.83
	Finite BIC	34.59	0.02	0.16	0.54	126.96	1.55	0.02	0.97
5	Fixed L1, CV	36.35	0.02	0.71	0.07	53.08	0.80	0.64	0.30
	Fixed L1, GCV	32.50	0.03	0.73	0.07	42.45	0.62	0.89	0.04
	Fixed L1, adapt., CV	35.19	0.02	0.59	0.12	53.95	0.79	0.58	0.30
	Fixed L1, adapt., GCV	35.19	0.02	0.59	0.12	53.27	0.79	0.58	0.30
	Fixed L1, adapt., CV, R	45.18	0.02	0.51	0.17	67.98	1.02	0.29	0.64
	Fixed L1, adapt., GCV, R	56.21	0.02	0.57	0.12	58.37	0.93	0.55	0.33
	Mixed	36.61	0.03	1.00	0.00	58.37	1.10	1.00	0.00
	Finite AIC	64.74	0.03	0.19	0.43	74.96	1.14	0.14	0.78
	Finite BIC	60.74	0.03	0.17	0.48	74.84	1.20	0.02	0.96

TABLE 11: *Results for the settings with binomial response,* $\beta_{i0} \sim N(-.3, 4)$, $n_i = 5$.

$\beta_{i0} \sim \chi^2_3$		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	Fixed L1, CV	35.13	0.02		0.11	144.52	1.20		0.17
	Fixed L1, GCV	32.05	0.03		0.10	137.26	0.99		0.06
	Fixed L1, adapt., CV	33.44	0.02		0.17	166.16	1.42		0.28
	Fixed L1, adapt., GCV	33.44	0.02		0.17	166.16	1.42		0.28
	Fixed L1, adapt., CV, R	41.92	0.02		0.27	191.15	1.90		0.48
	Fixed L1, adapt., GCV, R	47.38	0.02		0.19	184.65	1.59		0.31
	Mixed	36.84	0.03		0.00	172.23	1.93		0.00
	Finite AIC	54.99	0.03		0.53	218.36	2.01		0.73
	Finite BIC	54.71	0.03		0.55	218.69	2.11		0.85
15	Fixed L1, CV	86.90	0.03	0.71	0.11	200.60	1.35	0.74	0.13
	Fixed L1, GCV	79.55	0.03	0.74	0.10	196.30	1.28	0.81	0.06
	Fixed L1, adapt., CV	81.39	0.03	0.60	0.16	230.92	1.61	0.48	0.27
	Fixed L1, adapt., GCV	81.39	0.03	0.60	0.16	231.56	1.72	0.47	0.28
	Fixed L1, adapt., CV, R	97.63	0.02	0.48	0.26	273.13	2.19	0.25	0.57
	Fixed L1, adapt., GCV, R	103.46	0.03	0.58	0.18	259.38	1.92	0.45	0.29
	Mixed	88.96	0.03	1.00	0.00	238.12	2.52	1.00	0.00
	Finite AIC	114.89	0.04	0.20	0.50	262.16	2.34	0.17	0.63
	Finite BIC	114.76	0.04	0.18	0.53	311.54	2.41	0.08	0.76
5	Fixed L1, CV	23.64	0.02	0.71	0.13	93.98	1.09	0.71	0.17
	Fixed L1, GCV	23.12	0.02	0.74	0.11	88.04	0.95	0.80	0.09
	Fixed L1, adapt., CV	18.99	0.02	0.60	0.18	109.30	1.25	0.48	0.32
	Fixed L1, adapt., GCV	18.99	0.02	0.60	0.18	109.30	1.23	0.49	0.31
	Fixed L1, adapt., CV, R	23.27	0.01	0.47	0.30	137.24	1.59	0.28	0.58
	Fixed L1, adapt., GCV, R	23.99	0.01	0.57	0.20	125.81	1.39	0.46	0.35
	Mixed	26.68	0.02	1.00	0.00	126.30	1.88	1.00	0.00
	Finite AIC	34.78	0.02	0.14	0.53	170.65	1.86	0.16	0.71
	Finite BIC	33.58	0.02	0.11	0.55	172.05	1.94	0.05	0.89

TABLE 12: *Results for the settings with binomial response,* $\beta_{i0} \sim \chi^2_3$, $n_i = 5$.

References

Agresti, A., B. Caffo, and P. Ohman-Strickland (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis 47*, 639–653.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*, 117–128.

Bates, D., M. Maechler, and B. Bolker (2013). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2.

Bondell, H. D. and B. J. Reich (2009). Simultaneous factor selection and collapsing levels in anova. *Biometrics 65*, 169–177.

Curtin, T., S. Ingels, S. Wu, and R. Heuer (2002). *National Education Longitudinal Study of 1988: Base-Year to Fourth Follow-up Data File User's Manual (NCES 2002-323)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). New York: Oxford University Press.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Follmann, D. A. and D. Lambert (1989). Generalizing logistic regression by nonparametric mixing. *J. Amer. Statist. Assoc. 84*(405), 295–300.

Fruehwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.

Gertheiss, J. and G. Tutz (2010). Sparse modeling of categorial explanatory variables. *Ann. Appl. Stat. 4*(4), 2150–2180.

Goldstein, H. (2011). *Multilevel statistical models*. Wiley.

Grilli, L. and C. Rampichini (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 7*(4), 121–133.

Grün, B. and F. Leisch (2008a). FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Soft. 28*(4), 1–35.

Grün, B. and F. Leisch (2008b). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *J. Classification 25*(2), 225–247.

Heagerty, P. J. and B. F. Kurland (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika 88*(4), 973–984.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc. 102*, 1025–1038.

Litière, S., A. Alonso, and G. Molenberghs (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics 63*, 1038–1044.

McCulloch, C. E. and J. M. Neuhaus (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statist. Sci. 26*(3), 388–402.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica 46*(1), 69–85.

Neuhaus, J. M. and C. E. McCulloch (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *R. Stat. Soc. Ser. B Stat. Methodol. 68*(5), 859–872.

Oelker, M.-R. (2013). *gvcm.cat: Regularized categorial effects/categorial effect modifiers in GLMs*. R package version 1.7.

Oelker, M.-R., J. Gertheiss, and G. Tutz (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modeling 14*(2), 157–177.

Oelker, M.-R. and G. Tutz (2013). A general family of penalties for combining differing types of penalties in generalized structured models. Technical Report 139, Ludwig-Maximilians-Universität München, Department of Statistics. http://epub.ub.uni-muenchen.de/17664/.

O'Sullivan, F., B. S. Yandell, and W. J. Raynor (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc. 81*, 96–103.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Städler, N., P. Bühlmann, and S. van de Geer (2010). l_1-penalization for mixture regression models. *Test 19*(2), 209–256.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Kneight (2005). Sparsity and smoothness via the fused lasso. *R. Stat. Soc. Ser. B Stat. Methodol. 67*, 91–108.

Townsend, Z., J. Buckley, M. Harada, and M. A. Scott (2013). The choice between fixed and random effects. In B. D. M. Marc A. Scott, Jeffrey S. Simonoff (Ed.), *The SAGE Handbook of Multilevel Modeling*. SAGE.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press.

Tutz, G. and G. Schauberger (2014). Extended ordered paired comparison models with application to football data from german bundesliga. *Advances in Statistical Analysis (to appear)*.

Verbeke, G. and G. Molenberghs (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis 52*, 5277–5286.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *J. Amer. Statist. Assoc. 68*, 49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*(476), 1418–1429.