Generalized Linear Mixed Models Based on Boosting

Gerhard Tutz and Andreas Groll

Abstract A likelihood-based boosting approach for fitting generalized linear mixed models is presented. In contrast to common procedures it can be used in high-dimensional settings where a large number of potentially influential explanatory variables is available. Constructed as a componentwise boosting method it is able to perform variable selection with the complexity of the resulting estimator being determined by information criteria. The method is investigated in simulation studies and illustrated by using a real data set.

Keywords Generalized linear mixed model, Boosting, Linear models, Variable selection

Gerhard Tutz
Institut für Statistik, Ludwigstraße 33, Ludwig-Maximilians-Universität München, Germany, e-mail: gerhard.tutz@stat.uni-muenchen.de

Andreas Groll
Institut für Statistik, Ludwigstraße 33, Ludwig-Maximilians-Universität München, Germany, e-mail: andreas.groll@stat.uni-muenchen.de

1 Introduction

Generalized linear mixed models (GLMMs) as an extension of generalized linear models that incorporate random effects have been an area of intensive research. Various methods have been proposed ranging from numerical integration techniques (for example Booth & Hobert 1999) over "joint maximization methods" (Breslow & Clayton 1993, Schall 1991), in which parameters and random effects are estimated simultaneously, to fully Bayesian approaches (Fahrmeir & Lang 2001). Overviews on current methods are found in McCulloch & Searle (2001) and Fahrmeir & Tutz (2001). Due to the already heavy computational problems in GLMMs modelling usually is restricted to few predictor variables. When many predictors are given, the selection of predictors is often based on test statistics with the usual problems of forward-backward algorithms with stability of estimates.

In the present article boosting techniques for the selection of predictors are proposed. Boosting was developed within the machine learning community as a method to improve classification. A first breakthrough was the AdaBoost algorithm proposed by Freund & Schapire (1996). Breiman (1998) considered the AdaBoost algorithm as a gradient descent optimization technique and Friedman (2001) extended boosting methods to include regression problems. Bühlmann & Yu (2003) succeeded in proving an exponential dependence between the bias and the variance of the boosted model, which explains the resistance against overfitting. They showed how to fit smoothing splines by boosting base learners and introduced the idea of componentwise boosting, which may be exploited to select predictors. For a detailed overview of componentwise boosting, see Bühlmann & Yu (2003) and Bühlmann & Hothorn (2008).

The paper is structured as follows. In Section 2 we introduce the generalized linear mixed model. In Section 3 we present the boosting algorithm with its computational details and give further information about starting values, stopping criteria and selection. Then the performance of the boosting algorithm is investigated in two simulation studies, one for the random intercept Poisson model and one for the random intercept Bernoulli model. An application to the Multicenter AIDS Cohort Study (MACS, see Kaslow et al. 1987, Zeger & Diggle 1994) is considered in Section 4, which is based on the CD4 data and deals with gay or bisexual men infected with HIV.

2 Generalized Linear Mixed Models - GLMM

Let y_{it} denote observation t in cluster i, $i = 1,\dots,n$, $t = 1,\dots,T_i$, collected in $\mathbf{y}_i^T = (y_{i1},\dots,y_{iT_i})$. Let $\mathbf{x}_{it}^T = (x_{it1},\dots,x_{itp})$ be the covariate vector associated with fixed effects and $\mathbf{z}_{it}^T = (z_{it1},\dots,z_{its})$ the covariate vector associated with random effects. It is assumed that the observations y_{it} are conditionally independent with means $\mu_{it} = E(\mathbf{y}_{it}|\mathbf{b}_i,\mathbf{x}_{it},\mathbf{z}_{it})$ and variances $var(y_{it}|\mathbf{b}_i) = \phi\upsilon(\mu_{it})$, where $\upsilon(.)$ is a known variance function and ϕ is a scale parameter. The generalized linear mixed model that we consider in the following has the form

$$g(\mu_{it}) = \beta_0 + \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{z}_{it}^T\mathbf{b}_i = \beta_0 + \eta_{it}^{\text{par}} + \eta_{it}^{\text{rand}}, \tag{1}$$

where g is a monotonic and continuously differentiable link function, β_0 is the intercept, $\eta_{it}^{\text{par}} = \mathbf{x}_{it}^T\boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_1,\dots,\beta_p)$ and $\eta_{it}^{\text{rand}} = \mathbf{z}_{it}^T\mathbf{b}_i$ contains the cluster-specific random effects $\mathbf{b}_i \sim N(\mathbf{0},\mathbf{Q})$, with covariance matrix $\mathbf{Q}$.

An alternative form that we also use in the following is

$$\mu_{it} = h(\eta_{it}), \quad \eta_{it} = \beta_0 + \eta_{it}^{\text{par}} + \eta_{it}^{\text{rand}},$$

where $h = g^{-1}$ is the inverse link function.

A closed representation of model (1) is obtained by using matrix notation. Let $\mathbf{X}_i^T = (\mathbf{x}_{i1},\dots,\mathbf{x}_{iT_i})$ denote the design matrix of the i-th covariate and $\tilde{\boldsymbol{\beta}}^T = (\beta_0,\boldsymbol{\beta}^T)$ the linear parameter vector including intercept. Let $\tilde{\mathbf{X}}_i = [\mathbf{1},\mathbf{X}_i]$ be the corresponding design matrix, where $\mathbf{1}^T = (1,\dots,1)$ is a vector of ones having suitable length. By collecting observations within one cluster the model has the form

$$g(\boldsymbol{\mu}_i) = \tilde{\mathbf{X}}_i\tilde{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{b}_i,$$

where $\mathbf{Z}_i^T = (\mathbf{z}_{i1},\dots,\mathbf{z}_{iT_i})$. For all observations one obtains

$$g(\boldsymbol{\mu}) = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{b},$$

with $\tilde{\mathbf{X}}^T = [\tilde{\mathbf{X}}_1^T,\dots,\tilde{\mathbf{X}}_n^T]$ and block-diagonal matrix $\mathbf{Z} = Blockdiag(\mathbf{Z}_1,\dots,\mathbf{Z}_n)$. For the random effect $\mathbf{b}$ one has a normal distribution with covariance matrix $\mathbf{Q}_{\mathbf{b}} = Blockdiag(\mathbf{Q},\dots,\mathbf{Q})$.

Focusing on generalized linear mixed models we assume that the conditional density of y_{it}, given explanatory variables and the random effect $\mathbf{b}_i$, is of exponential family type

$$f(y_{it}|\mathbf{X}_i,\mathbf{b}_i) = \exp\left\{\frac{(y_{it}\theta_{it} - \kappa(\theta_{it}))}{\phi} + c(y_{it},\phi)\right\}, \tag{2}$$

where $\theta_{it} = \theta(\mu_{it})$ denotes the natural parameter, $\kappa(\theta_{it})$ is a specific function corresponding to the type of exponential family, $c(.)$ the log normalization constant and ϕ the dispersion parameter (compare Fahrmeir & Tutz 2001).

One popular method to maximize generalized linear mixed models is penalized quasi-likelihood (PQL), which has been suggested by Breslow & Clayton (1993), Lin & Breslow (1996) and Breslow & Lin (1995). Typically the covariance matrix $\mathbf{Q}(\boldsymbol{\rho})$ of the random effects $\mathbf{b}_i$ dependents on an unknown parameter vector $\boldsymbol{\rho}$. In penalization-based concepts the joint likelihood-function is specified by the parameter vector of the covariance structure $\boldsymbol{\rho}$ together with the dispersion parameter ϕ, which are collected in $\boldsymbol{\gamma}^T = (\phi,\boldsymbol{\rho}^T)$ and parameter vector $\boldsymbol{\delta}^T = (\beta_0,\boldsymbol{\beta}^T,\mathbf{b}^T)$, $\mathbf{b}^T = (\mathbf{b}_1^T,\dots,\mathbf{b}_n^T)$. The corresponding log-likelihood is

$$l(\boldsymbol{\delta},\boldsymbol{\gamma}) = \sum_{i=1}^{n} \log\left(\int f(\mathbf{y}_i|\boldsymbol{\delta},\boldsymbol{\gamma})p(\mathbf{b}_i,\boldsymbol{\gamma})d\mathbf{b}_i\right), \tag{3}$$

where $p(\mathbf{b}_i,\boldsymbol{\gamma})$ denotes the density of the random effects. Approximation of (3) along the lines of Breslow & Clayton (1993) yields the penalized likelihood

$$l^P(\boldsymbol{\delta},\boldsymbol{\gamma}) = \sum_{i=1}^{n} \log(f(\mathbf{y}_i|\boldsymbol{\delta},\boldsymbol{\gamma})) - \frac{1}{2}\mathbf{b}^T\mathbf{Q}(\boldsymbol{\rho})^{-1}\mathbf{b}, \tag{4}$$

where the penalty term $\mathbf{b}^T\mathbf{Q}(\boldsymbol{\rho})^{-1}\mathbf{b}$ is due to the approximation based on the Laplace method.

PQL usually works within the profile likelihood concept. It is distinguished between the estimation of $\boldsymbol{\delta}$ given the plugged-in estimate $\hat{\boldsymbol{\gamma}}$ resulting in the profile-likelihood $l^P(\boldsymbol{\delta},\hat{\boldsymbol{\gamma}})$ and the estimation of $\boldsymbol{\gamma}$. The PQL method is implemented in the macro `GLIMMIX` and proc `GLMMIX` in SAS (Wolfinger 1994), in the `glmmPQL` and `gamm` functions of the R-packages `MASS` (Venables & Ripley 2002) and `mgcv` (Wood 2006). Further notes were given by Wolfinger & O'Connell (1993), Littell et al. (1996) and Vonesh (1996).

3 Boosted Generalized Linear Mixed Models - bGLMM

Boosting originates in the machine learning community where it has been proposed as a technique to improve classification procedures by combining estimates with reweighted observations. Since it has been shown in Breiman (1999) and Friedman (2001) that reweighting corresponds to minimizing iteratively a loss function, boosting has been extended to regression problems in a L2-estimation framework by Bühlmann & Yu (2003). The boosting algorithm presented in this paper is based on the likelihood function and works by iterative fitting of residuals using "weak learners" and implies selection of components.

3.1 The Boosting Algorithm

The following algorithm uses componentwise boosting. Componentwise boosting means that only one component of the predictor, in our case one linear term, is fitted at a time. More precisely, a model containing the intercept and only one linear term $x_r\beta_r$ is fitted in one iteration step. We will use the notation $\mathbf{x}_{i.r}^T = (x_{i1r},\dots,x_{iT_ir})$ for the covariate vector of the r-th linear effect in cluster i and define $\mathbf{x}_{.r}^T = (\mathbf{x}_{1.r}^T,\dots,\mathbf{x}_{n.r}^T), r = 1,\dots,p$. Hence the corresponding r-th design matrix containing intercept and only r-th covariate vector is given by

$$\mathbf{X}_{i.r} = [\mathbf{1},\mathbf{x}_{i.r}] \quad \text{and} \quad \mathbf{X}_{.r} = [\mathbf{1},\mathbf{x}_{.r}],$$

for cluster i and the whole sample, respectively. For cluster i the predictor that contains only the r-th covariate has the form $\boldsymbol{\eta}_{ir} = \mathbf{X}_{i.r}\tilde{\boldsymbol{\beta}}_r + \mathbf{Z}_i\mathbf{b}_i$, with $\tilde{\boldsymbol{\beta}}_r^T = (\beta_0, \beta_r)$, and for the whole sample one obtains

$$\boldsymbol{\eta}_r = \mathbf{X}_{.r}\tilde{\boldsymbol{\beta}}_r + \mathbf{Z}\mathbf{b}.$$

In the following boosting algorithm the vectors $\tilde{\boldsymbol{\beta}}_r^T = (\beta_0, \beta_r)$ and $\boldsymbol{\delta}_r^T = (\beta_0, \beta_r, \mathbf{b}^T)$ contain only the r-th fixed effect.

Algorithm `bGLMM`

---

1. *Initialization*
   Compute starting values $\hat{\boldsymbol{\mu}}^{(0)}, \mathbf{Q}^{(0)}, \mathbf{b}^{(0)}$ (see Section 3.2.3) and set $\boldsymbol{\eta}^{(0)} = \mathbf{X}\hat{\hat{\boldsymbol{\beta}}}^{(0)} + \mathbf{Z}\mathbf{b}^{(0)}$.
2. *Iteration*
   For $l = 1, 2, \ldots, l_{max}$
   a. *Refitting of residuals*
      i. Computation of parameters
         For $r \in \{1, \ldots, p\}$ derive the penalized score function $\mathbf{s}_r^P(\boldsymbol{\delta}) = \partial l^P / \partial \boldsymbol{\delta}_r$ and the penalized pseudo Fisher matrix $\mathbf{F}_r^P(\boldsymbol{\delta})$ (see Section 3.2.1). Based on the general form of one step in Fisher scoring given by

$$\hat{\boldsymbol{\delta}}^{(l)} = \hat{\boldsymbol{\delta}}^{(l-1)} + (\mathbf{F}^P(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\mathbf{s}^P(\hat{\boldsymbol{\delta}}^{(l-1)}).$$

         an update of the r-th component is computed. Because the fit is within an iterative procedure it is sufficient to use just one single step. In order to obtain an additive correction of the already fitted terms (the offset), we use one step in Fisher scoring with starting value $\boldsymbol{\delta} = \mathbf{0}$. Therefore Fisher scoring for the r-th component takes the simpler form

$$\hat{\boldsymbol{\delta}}_r^{(l)} = (\mathbf{F}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\mathbf{s}_r(\hat{\boldsymbol{\delta}}^{(l-1)}) \tag{5}$$

         with variance-covariance components being replaced by their current estimates $\hat{\mathbf{Q}}^{(l-1)}$.
      ii. Selection step
         Select from $r \in \{1, \ldots, p\}$ the component j that leads to the smallest $AIC_r^{(l)}$ or $BIC_r^{(l)}$ as given in Section 3.2.3 and select the corresponding $(\hat{\boldsymbol{\delta}}_j^{(l)})^T = \left(\hat{\beta}_0^*, \hat{\beta}_j^*, (\hat{\mathbf{b}}^*)^T\right)$.
      iii. Update
         Set

$$\hat{\beta}_0^{(l)} = \hat{\beta}_0^{(l-1)} + \hat{\beta}_0^*, \qquad \hat{\mathbf{b}}^{(l)} = \hat{\mathbf{b}}^{(l-1)} + \hat{\mathbf{b}}^*$$

         and for $r = 1, \ldots, p$ set

$$\hat{\beta}_r^{(l)} = \begin{cases} \hat{\beta}_r^{(l-1)} & \text{if } r \neq j \\ \hat{\beta}_r^{(l-1)} + \hat{\beta}_r^* & \text{if } r = j, \end{cases}$$

$$(\hat{\boldsymbol{\delta}}^{(l)})^T = \left(\hat{\beta}_0^{(l)}, \hat{\beta}_1^{(l)}, \ldots, \hat{\beta}_p^{(l)}, (\hat{\mathbf{b}}^{(l)})^T\right).$$

With $\mathbf{A} := [\mathbf{X}, \mathbf{Z}]$ update

$$\hat{\boldsymbol{\eta}}^{(l)} = \mathbf{A}\hat{\boldsymbol{\delta}}^{(l)}$$

b. *Computation of variance-covariance components*
Estimates of $\hat{\mathbf{Q}}^{(l)}$ are obtained as approximate REML-type estimates or alternative methods (see Section 3.2.2)

3.2 Computational Details of bGLMM

In the following we give a more detailed description of the single steps of the `bGLMM` algorithm. First we describe the derivation of the score function and the Fisher matrix. Then two estimation techniques for the variance-covariance components are given. Finally, we give details of the computation of starting values and the selection procedure.

3.2.1 Score Function and Fisher Matrix

In this section we specify more precisely the single components which are derived in step 2 (a) of the `bGLMM` algorithm. For $r \in \{1, \ldots, p\}$ the penalized score function $\mathbf{s}_r^P(\boldsymbol{\delta}) = \partial l^P / \partial \boldsymbol{\delta}_r$, obtained by differentiating the log-likelihood from equation (4), has vector components

$$\begin{aligned} \mathbf{s}_{\tilde{\boldsymbol{\beta}}r}^P &= \sum_{i=1}^n \mathbf{X}_{ir}^T \mathbf{D}_i(\boldsymbol{\delta}) \boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\delta})), \\ \mathbf{s}_{ir}^P &= \mathbf{Z}_i^T \mathbf{D}_i(\boldsymbol{\delta}) \boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\delta})) - \mathbf{Q}^{-1}\mathbf{b}_i, \quad i = 1, \ldots, n, \end{aligned}$$

with $\mathbf{D}_i(\boldsymbol{\delta}) = \partial h(\boldsymbol{\eta}_i)/\partial \boldsymbol{\eta}, \boldsymbol{\Sigma}_i(\boldsymbol{\delta}) = cov(\mathbf{y}_i | \tilde{\boldsymbol{\beta}}, \mathbf{b}_i)$, and $\boldsymbol{\mu}_i(\boldsymbol{\delta}) = h(\boldsymbol{\eta}_i)$. The vector $\mathbf{s}_{\tilde{\boldsymbol{\beta}}r}^P$ has dimension $p+1$, while the vectors $\mathbf{s}_{ir}^P$ are of dimension s. Note that $\mathbf{s}_r^P(\boldsymbol{\delta})$ could be seen as a penalized score function because of the term $\mathbf{Q}^{-1}\mathbf{b}_i$.
The penalized pseudo-Fisher matrix $\mathbf{F}_r^P(\boldsymbol{\delta})$, $r \in \{1, \ldots, p\}$, which is partitioned into

$$\mathbf{F}_r^P(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}r} & \mathbf{F}_{\tilde{\boldsymbol{\beta}}1r} & \mathbf{F}_{\tilde{\boldsymbol{\beta}}2r} & \cdots & \mathbf{F}_{\tilde{\boldsymbol{\beta}}nr} \\ \mathbf{F}_{1\tilde{\boldsymbol{\beta}}r} & \mathbf{F}_{11r} & & & 0 \\ \mathbf{F}_{2\tilde{\boldsymbol{\beta}}r} & & \mathbf{F}_{22r} & & \\ \vdots & & & \ddots & \\ \mathbf{F}_{n\tilde{\boldsymbol{\beta}}r} & 0 & & & \mathbf{F}_{nnr} \end{bmatrix},$$

has single components

$$\begin{aligned} \mathbf{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}r} &= -E\left(\frac{\partial^2 l^P(\boldsymbol{\delta})}{\partial \tilde{\boldsymbol{\beta}}_r \partial \tilde{\boldsymbol{\beta}}_r^T}\right) = \sum_{i=1}^n \mathbf{X}_{ir}^T \mathbf{D}_i(\boldsymbol{\delta}) \boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1} \mathbf{D}_i(\boldsymbol{\delta}) \mathbf{X}_{ir}, \\ \mathbf{F}_{\tilde{\boldsymbol{\beta}}ir} = \mathbf{F}_{i\tilde{\boldsymbol{\beta}}r}^T &= -E\left(\frac{\partial^2 l^P(\boldsymbol{\delta})}{\partial \tilde{\boldsymbol{\beta}}_r \partial \mathbf{b}_i^T}\right) = \mathbf{X}_{ir}^T \mathbf{D}_i(\boldsymbol{\delta}) \boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1} \mathbf{D}_i(\boldsymbol{\delta}) \mathbf{Z}_i, \\ \mathbf{F}_{iir} &= -E\left(\frac{\partial^2 l^P(\boldsymbol{\delta})}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}\right) = \mathbf{Z}_i^T \mathbf{D}_i(\boldsymbol{\delta}) \boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1} \mathbf{D}_i(\boldsymbol{\delta}) \mathbf{Z}_i + \mathbf{Q}^{-1}. \end{aligned}$$

3.2.2 Variance-Covariance Components

For the estimation of variances (Breslow & Clayton 1993) maximize the profile likelihood that is associated with the normal theory model. By replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ one maximizes

$$\begin{aligned} l(\mathbf{Q_b}) = &-\frac{1}{2}\log(|\mathbf{V}(\hat{\boldsymbol{\delta}})|) - \frac{1}{2}\log(|\mathbf{X}^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}})\mathbf{X}|) \\ &-\frac{1}{2}(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}) - \mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1}(\hat{\boldsymbol{\delta}})(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}) - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{6}$$

with respect to $\mathbf{Q_b}$, with the pseudo-observations $\tilde{\boldsymbol{\eta}}(\boldsymbol{\delta}) = \mathbf{A}\boldsymbol{\delta} + \mathbf{D}^{-1}(\boldsymbol{\delta})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\delta}))$ and with matrices $\mathbf{V}(\boldsymbol{\delta}) = \mathbf{W}^{-1}(\boldsymbol{\delta}) + \mathbf{Z}\mathbf{Q_b}\mathbf{Z}^T$, $\mathbf{Q_b} = Blockdiag(\mathbf{Q},\ldots,\mathbf{Q})$ and $\mathbf{W}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta})\mathbf{D}(\boldsymbol{\delta})^T$. Having calculated $\hat{\boldsymbol{\delta}}^{(l)}$ in the l-th boosting iteration, we obtain the estimator $\hat{\mathbf{Q}}_{\mathbf{b}}^{(l)}$, which is an approximate REML-type estimate for $\mathbf{Q_b}$.

An alternative estimate, which can be derived as an approximate EM algorithm, uses the posterior mode estimates and posterior curvatures. One derives $(\mathbf{F}^P(\hat{\boldsymbol{\delta}}^{(l)}))^{-1}$, the inverse of the penalized pseudo Fisher matrix of the full model using the posterior mode estimates $\hat{\boldsymbol{\delta}}^{(l)}$ to obtain the posterior curvatures $\hat{\mathbf{V}}_{ii}^{(l)}$. Now compute $\hat{\mathbf{Q}}^{(l)}$ by

$$\hat{\mathbf{Q}}^{(l)} = \frac{1}{n}\sum_{i=1}^n (\hat{\mathbf{V}}_{ii}^{(l)} + \hat{\mathbf{b}}_i^{(l)}(\hat{\mathbf{b}}_i^{(l)})^T). \tag{7}$$

In general, the $\mathbf{V}_{ii}$ are derived via the formula

$$\mathbf{V}_{ii} = \mathbf{F}_{ii}^{-1} + \mathbf{F}_{ii}^{-1}\mathbf{F}_{i\tilde{\boldsymbol{\beta}}}(\mathbf{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} - \sum_{i=1}^{n}\mathbf{F}_{\tilde{\boldsymbol{\beta}}i}\mathbf{F}_{ii}^{-1}\mathbf{F}_{i\tilde{\boldsymbol{\beta}}})^{-1}\mathbf{F}_{\tilde{\boldsymbol{\beta}}i}\mathbf{F}_{ii}^{-1},$$

where $\mathbf{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}}, \mathbf{F}_{i\tilde{\boldsymbol{\beta}}}, \mathbf{F}_{ii}$ are the elements of the penalized pseudo Fisher matrix $\mathbf{F}^{P}(\boldsymbol{\delta})$ of the full model, for details see for example Fahrmeir & Tutz (2001).

3.2.3 Starting Values, Stopping Criteria and Selection in bGLMM

We compute the starting values $\hat{\boldsymbol{\mu}}^{(0)}, \mathbf{Q}^{(0)}$ from step 1. of the `bGLMM` algorithm by fitting the simple global intercept model with random effects given by

$$g(\mu_{it}) = \beta_0 + \mathbf{z}_{it}^T\mathbf{b}_i. \tag{8}$$

This can be done very easily, e.g. by using the R-function `glmmPQL` (Wood 2006) from the `MASS` library (Venables & Ripley 2002).

To find the appropriate complexity of our model we use the effective degrees of freedom, which corresponds to the trace of the hat matrix (Hastie & Tibshirani 1990). In the following we derive the hat matrix corresponding to the l-th boosting step for the r-th component (compare Tutz & Binder 2007, Leitenstorfer 2008). Let $\mathbf{A}_r := [\mathbf{X}_r, \mathbf{Z}]$ and $\mathbf{K} = Blockdiag(0, 0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$ be a block diagonal penalty matrix with a diagonal of two zeros corresponding to intercept and r-th fixed effect and n times the matrix $\mathbf{Q}^{-1}$. Then the Fisher matrix $\mathbf{F}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)})$ and the score vector $\mathbf{s}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)})$ are given in closed form as

$$\mathbf{F}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)}) = \mathbf{A}_r\mathbf{W}_l\mathbf{A}_r + \mathbf{K}$$

and

$$\mathbf{s}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)}) = \mathbf{A}_r^T\mathbf{W}_l\mathbf{D}_l^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}) - \mathbf{K}\hat{\boldsymbol{\delta}}_r^{(l-1)}$$

where $\mathbf{W}_l = \mathbf{W}(\hat{\boldsymbol{\delta}}^{(l-1)}), \mathbf{D}_l = \mathbf{D}(\hat{\boldsymbol{\delta}}^{(l-1)}), \boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}(\hat{\boldsymbol{\delta}}^{(l-1)})$ and $\hat{\boldsymbol{\mu}}^{(l-1)} = h(\hat{\boldsymbol{\eta}}^{(l-1)}) = h(\mathbf{A}\hat{\boldsymbol{\delta}}^{(l-1)})$. For $r = 1, \ldots, m$ the refit in the l-th iteration step by Fisher scoring (5) is given by

$$\begin{aligned}\hat{\boldsymbol{\delta}}_r^{(l)} &= (\mathbf{F}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\mathbf{s}_r(\hat{\boldsymbol{\delta}}^{(l-1)}) \\ &= (\mathbf{A}_r\mathbf{W}_l\mathbf{A}_r + \mathbf{K})^{-1}\mathbf{A}_r^T\mathbf{W}_l\mathbf{D}_l^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}).\end{aligned}$$

We define the predictor corresponding to the r-th refit in the l-th iteration step as

$$\begin{aligned}\hat{\boldsymbol{\eta}}_r^{(l)} &:= \hat{\boldsymbol{\eta}}^{(l-1)} + \mathbf{A}_r\hat{\boldsymbol{\delta}}_r^{(l)}, \\ \hat{\boldsymbol{\eta}}_r^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)} &= \mathbf{A}_r\hat{\boldsymbol{\delta}}_r^{(l)} \\ &= \mathbf{A}_r(\mathbf{A}_r\mathbf{W}_l\mathbf{A}_r + \mathbf{K})^{-1}\mathbf{A}_r^T\mathbf{W}_l\mathbf{D}_l^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}).\end{aligned}$$

Taylor approximation of first order $h(\hat{\boldsymbol{\eta}}) = h(\boldsymbol{\eta}) + \frac{\partial h(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ yields

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_r^{(l)} &\approx \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{D}_l(\hat{\boldsymbol{\eta}}_r^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)}),\\
\hat{\boldsymbol{\eta}}_r^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)} &\approx \mathbf{D}_l^{-1}(\hat{\boldsymbol{\mu}}_r^{(l)} - \hat{\boldsymbol{\mu}}^{(l-1)}),
\end{aligned}$$

and therefore

$$\mathbf{D}_l^{-1}(\hat{\boldsymbol{\mu}}_r^{(l)} - \hat{\boldsymbol{\mu}}^{(l-1)}) \approx \mathbf{A}_r\,(\mathbf{A}_r\mathbf{W}_l\mathbf{A}_r + \mathbf{K})^{-1}\,\mathbf{A}_r^T\mathbf{W}_l\mathbf{D}_l^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}).$$

Multiplication with $\mathbf{W}_l^{1/2}$ and using $\mathbf{W}^{1/2}\mathbf{D}^{-1} = \boldsymbol{\Sigma}^{-1/2}$ yields

$$\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}}_r^{(l)} - \hat{\boldsymbol{\mu}}^{(l-1)}) \approx \tilde{\mathbf{H}}_r^{(l)}\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}),$$

where $\tilde{\mathbf{H}}_r^{(l)} := \mathbf{W}_l^{1/2}\mathbf{A}_r\,(\mathbf{A}_r\mathbf{W}_l\mathbf{A}_r + \mathbf{K})^{-1}\,\mathbf{A}_r^T\mathbf{W}_l^{1/2}$ denotes the usual generalized ridge regression hat-matrix. Defining $\mathbf{M}_r^{(l)} := \boldsymbol{\Sigma}_l^{1/2}\tilde{\mathbf{H}}_r^{(l)}\boldsymbol{\Sigma}_l^{-1/2}$ yields the approximation

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_r^{(l)} &\approx \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{M}_r^{(l)}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)})\\
&= \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{M}_r^{(l)}[(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-2)}) - (\hat{\boldsymbol{\mu}}^{(l-1)} - \hat{\boldsymbol{\mu}}^{(l-2)})]\\
&\approx \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{M}_r^{(l)}[(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-2)}) - \mathbf{M}_r^{(l-1)}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-2)})].
\end{aligned}$$

The hat matrix corresponding to the global intercept model from equation (8) is

$$\mathbf{M}^{(0)} = \mathbf{A}_1(\mathbf{A}_1^T\mathbf{W}_1\mathbf{A}_1 + \mathbf{K}_1)\mathbf{A}_1^T\mathbf{W}_1,$$

with matrices $\mathbf{A}_1 := [\mathbf{1}, \mathbf{Z}]$ and $\mathbf{K}_1 := Blockdiag(0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$. As the approximation $\hat{\boldsymbol{\mu}}^{(0)} \approx \mathbf{M}^{(0)}\mathbf{y}$ holds, one obtains

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_r^{(1)} &\approx \hat{\boldsymbol{\mu}}^{(0)} + \mathbf{M}_r^{(1)}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)})\\
&\approx \mathbf{M}^{(0)}\mathbf{y} + \mathbf{M}_r^{(1)}(\mathbf{I} - \mathbf{M}^{(0)})\mathbf{y}.
\end{aligned}$$

In the following, to indicate that the hat matrices of the former steps have been fixed, let $j_k \in \{1, \ldots, p\}$ denote the index of the component selected in boosting step k. Then we can abbreviate $\mathbf{M}_{j_k} := \mathbf{M}_{j_k}^{(k)}$ for the matrix corresponding to the component that has been selected in the k-th iteration. Further, in a recursive manner, we get

$$\hat{\boldsymbol{\mu}}_r^{(l)} \approx \mathbf{H}_r^{(l)}\mathbf{y},$$

where

$$\begin{aligned}
\mathbf{H}_r^{(l)} &= \mathbf{I} - (\mathbf{I} - \mathbf{M}_r^{(l)})(\mathbf{I} - \mathbf{M}_{j_{l-1}})(\mathbf{I} - \mathbf{M}_{j_{l-2}}) \cdot \ldots \cdot (\mathbf{I} - \mathbf{M}^{(0)}) \\
&= \mathbf{M}_r^{(l)} \prod_{i=0}^{l-1} (\mathbf{I} - \mathbf{M}_{j_i}) + \sum_{k=0}^{l-1} \mathbf{M}_{j_k} \prod_{i=0}^{k-1} (\mathbf{I} - \mathbf{M}_{j_i}) \\
&= \sum_{k=0}^{l} \mathbf{M}_{j_k} \prod_{i=0}^{k-1} (\mathbf{I} - \mathbf{M}_{j_i}),
\end{aligned}$$

is the hat matrix corresponding to the l-th boosting step considering the r-th component, whereas $\mathbf{M}_{j_l} := \mathbf{M}_r^{(l)}$ is not fixed yet.

In general, given hat matrix $\mathbf{H}$, the complexity of the model may be determined by the information criteria. We will use

$$AIC = -2\, l(\boldsymbol{\mu}) + 2\, \mathrm{trace}\,(\mathbf{H}), \tag{9}$$

$$BIC = -2\, l(\boldsymbol{\mu}) + 2\, \mathrm{trace}\,(\mathbf{H})\, log(n), \tag{10}$$

where

$$l(\boldsymbol{\mu}) = \sum_{i=1}^{n} l_i(\boldsymbol{\mu}_i) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\boldsymbol{\mu}_i) \tag{11}$$

denotes the general log-likelihood and $l_i(\boldsymbol{\mu}_i)$ the log-likelihood contribution of $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$. In general, the log-likelihood (4) can also be written with $\boldsymbol{\mu}$ instead of $\boldsymbol{\delta}$ in the argument, considering the definition of the natural parameter $\theta = \theta(\boldsymbol{\mu})$ in (2) and using $\boldsymbol{\mu} = h(\boldsymbol{\eta}) = h(\boldsymbol{\eta}(\boldsymbol{\delta}))$. In (9) and (10) the nonpenalized log-likelihood is used.

For exponential family distributions $\log f(\mathbf{y}_i|\boldsymbol{\mu}_i)$ has a well-known form. For example in the case of binary responses, one obtains

$$\log f(\mathbf{y}_i|\boldsymbol{\mu}_i) = \sum_{t=1}^{T_i} y_{it} \log \mu_{it} + (1 - y_{it}) \log (1 - \mu_{it}),$$

whereas in the case of Poisson responses, one has

$$\log f(\mathbf{y}_i|\boldsymbol{\mu}_i) = \sum_{t=1}^{T_i} y_{it} \log \mu_{it} - \mu_{it}.$$

Based on (11), the information criteria (9) and (10) used in the l-th boosting step, considering the r-th component, have the form

$$\begin{aligned}
AIC_r^{(l)} &= -2\, l(\hat{\boldsymbol{\mu}}_r^{(l)}) + 2\, \mathrm{trace}\,(\mathbf{H}_r^{(l)}), \\
BIC_r^{(l)} &= -2\, l(\hat{\boldsymbol{\mu}}_r^{(l)}) + 2\, \mathrm{trace}\,(\mathbf{H}_r^{(l)})\, log(n),
\end{aligned}$$

with

$$l(\hat{\boldsymbol{\mu}}_r^{(l)}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_{ir}^{(l)}). \tag{12}$$

In the l-th step one selects from $r \in \{1,\dots,p\}$ the component j_l that minimizes $AIC_r^{(l)}$ or $BIC_r^{(l)}$ and obtains $AIC^{(l)} := AIC_{j_l}^{(l)}$. We choose a number l_{max} of maximal boosting steps, e.g. $l_{max} = 1000$, and stop the algorithm at iteration l_{max}. Then we select from $\mathscr{L} := \{1,2,\dots,l_{max}\}$ the component l_{opt}, where $AIC^{(l)}$ or $BIC^{(l)}$ is smallest, that is

$$\begin{aligned} l_{opt} &= \arg \min_{l \in \mathscr{L}} AIC^{(l)}, \\ l_{opt} &= \arg \min_{l \in \mathscr{L}} BIC^{(l)}. \end{aligned}$$

Finally, we obtain the parameter estimates $\hat{\boldsymbol{\delta}}^{(l_{opt})}, \hat{\mathbf{Q}}^{(l_{opt})}$ and the corresponding fit $\hat{\boldsymbol{\mu}}^{(l_{opt})}$.

It should be noted that similar to Tutz & Reithinger (2006) our selection step reflects the complexity of the refitted model, which is in contrast to established componentwise boosting procedures. For example Bühlmann & Yu (2003), select the component that maximally improves the fit and then evaluate if the fit including model complexity deteriorates. The procedure proposed here selects the component such that the new lack-of-fit, including the augmented complexity, is minimized.

3.3 Simulation Study

In the following simulation studies the performance of the `bGLMM` algorithm is compared to alternative approaches.

Poisson Link The underlying model is the random intercept Poisson model

$$\begin{aligned} \eta_{it} &= \sum_{j=1}^{p} x_{itj}\beta_j + b_i, \quad i = 1,\dots,40, \quad t = 1,\dots,10, \\ E[y_{it}] &= \exp(\eta_{it}) := \lambda_{it}, \qquad y_{it} \sim \text{Pois}(\lambda_{it}), \end{aligned}$$

with linear effects given by $\beta_1 = -4, \beta_2 = -6, \beta_3 = 10$ and $\beta_j = 0, j = 4,\dots,50$. We choose the different settings $p = 3,5,10,20,50$. For $j = 1,\dots,50$ the vectors $\mathbf{x}_{it}^T = (x_{it1},\dots,x_{it50})$ follow a uniform distribution within the interval $[-0.3, 0.3]$. The number of observations is determined by $n = 40, T_i := T = 10, i = 1,\dots,n$. The random effect and the noise variable have been specified by $b_i \sim N(0,\sigma_b^2)$ with $\sigma_b^2 = 0.6$.

The performance of estimators is evaluated separately for the structural components and the variance. We compare the results of our `bGLMM` algorithm with the results obtained by the R-function `glmmPQL` recommended in Wood (2006). The `glmmPQL` routine is supplied with the `MASS` library (Venables & Ripley 2002). It operates by iteratively calling the R-function `lme` from the `nlme` library and returns the fitted `lme` model object for the working model at convergence. For more details about the `lme` function, see Pinheiro & Bates (2000).

By averaging across 50 training data sets we consider mean squared errors for $\boldsymbol{\beta}$ and σ_b given by

$$\text{mse}_{\boldsymbol{\beta}} := ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2, \qquad \text{mse}_{\sigma_b} := ||\sigma_b - \hat{\sigma}_b||^2.$$

To avoid that single outliers distort the analysis, we present the medians of both quantities in Table 1. The corresponding boxplots are shown in Figure 1. Additionally, we present boxplots of the σ_b-difference

$$\triangle_{\sigma_b} := \sigma_b - \hat{\sigma}_b$$

in Figure 2, to investigate the bias of estimates the true value $\sigma_b = \sqrt{0.6}$.

Additional information on the performance of the algorithm was collected in *falseneg*, the mean over all 50 simulations of the number of variables $\beta_j, j = 1,2,3$, that were not selected and in *falsepos*, the mean over all 50 simulations of the number of variables $\beta_j, j = 4,\ldots,50$, that were selected. Notice at this point, that the `glmmPQL` function is not able to perform variable selection and therefore always estimates all p paramteres β_j.

The results for varying number p of covariates $x_{it1},\ldots,x_{itp}$ are summarized in Table 1. For the computation of the random effects variance-covariance components $\mathbf{Q}$ we used the two estimation techniques given in Section 3.2.2. The results using the EM-type estimates $\hat{\mathbf{Q}}$ from (7) are found in the `bGLMM` (EM) column of Table 1, results for the REML-type estimates $\hat{\mathbf{Q}}$, obtained by maximization of the profile likelihood in (6), are given in the third column. The corresponding results can be found in the `bGLMM` (REML) column of Table 1. It is seen that boosting estimates distinctly outperform the simple PQL algorithm when redundant variables are present. REML typre estimates turned out to be more stable than the EM-type estimates.

Table 1 Generalized linear mixed model (`glmmPQL`) and boosting (`bGLMM`) on Poisson data

	`glmmPQL`		`bGLMM` (EM)				`bGLMM` (REML)			
p	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	falsepos	falseneg	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	falsepos	falseneg
3	0.088	0.004	0.104	0.006	0	0	0.100	0.004	0	0
5	0.124	0.004	0.108	0.006	0.10	0	0.101	0.004	0.02	0
10	0.218	0.004	0.110	0.006	0.34	0	0.101	0.004	0.04	0
20	0.537	0.004	0.118	0.006	0.66	0	0.108	0.004	0.10	0
50	2.013	0.005	0.143	0.008	1.68	0	0.124	0.007	0.30	0

To illustrate how the `bGLMM` algorithm works we show in Figure 3 the paths of the three coefficients β_1, β_2 and β_3 for all simulations. It is seen that the algorithm always starts with updating coefficient β_3, which has the most influence on $\boldsymbol{\eta}$, as it has the biggest absolute value. Next, the coefficient β_2 is updated, while β_1 is the last coefficient which is refitted.

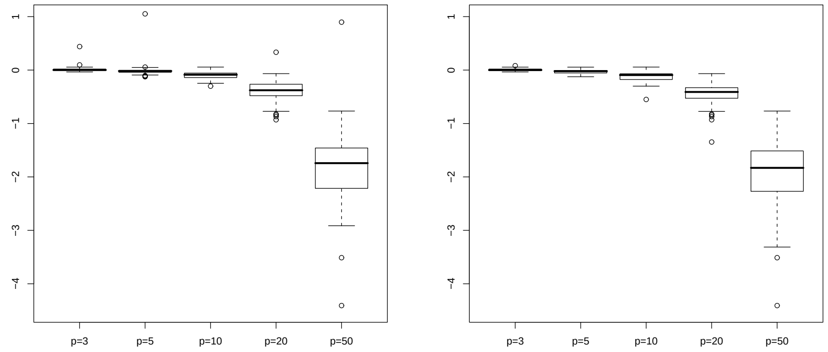

Fig. 1 Boxplots of $(\mathrm{mse}_{\boldsymbol{\beta}}^{\mathrm{bGLMM}} - \mathrm{mse}_{\boldsymbol{\beta}}^{\mathrm{glmmPQL}})$ for the EM model (left, without few extreme outliers) and the REML model (right)

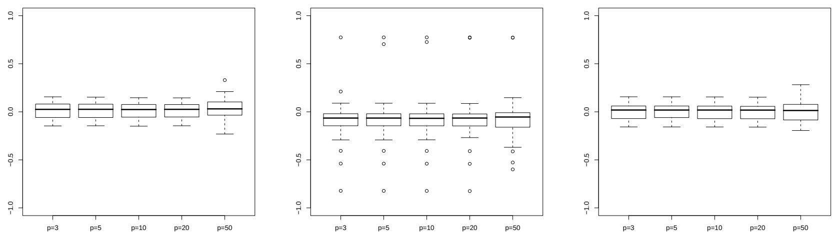

Fig. 2 Boxplots of $\triangle_{\sigma_b}$ for the `glmmPQL` model (left), for the `bGLMM` EM model (middle) and for the `bGLMM` REML model (right)

Bernoulli Link The underlying model is the random intercept Bernoulli model

$$\eta_{it} = \sum_{j=1}^{p} x_{itj}\beta_j + b_i, \quad i = 1,\ldots,40, \quad t = 1,\ldots,10$$

$$E[y_{it}] = \frac{\exp(\eta_{it})}{1+\exp(\eta_{it})} := \pi_{it} \qquad y_{it} \sim \mathrm{B}(1,\pi_{it})$$

with linear effects given by $\beta_1 = -5, \beta_2 = -10, \beta_3 = 15$ and $\beta_j = 0, j = 4,\ldots,50$. Again we choose the different settings $p = 3,5,10,20,50$. For $j = 1,\ldots,50$ the vectors $\mathbf{x}_{it}^T = (x_{it1},\ldots,x_{it50})$ have been drawn independently with components following a uniform distribution within the interval $[-0.1, 0.1]$. The number of observations remains $n = 40, T_i := T = 10, \forall i = 1,\ldots,n$. The random effect and the noise variable have been specified by $b_i \sim N(0,\sigma_b^2)$ with $\sigma_b^2 = 0.6$.

Again, we evaluate the performance of estimators separately for structural components and variance and compare the results of our `bGLMM` algorithm with the re-

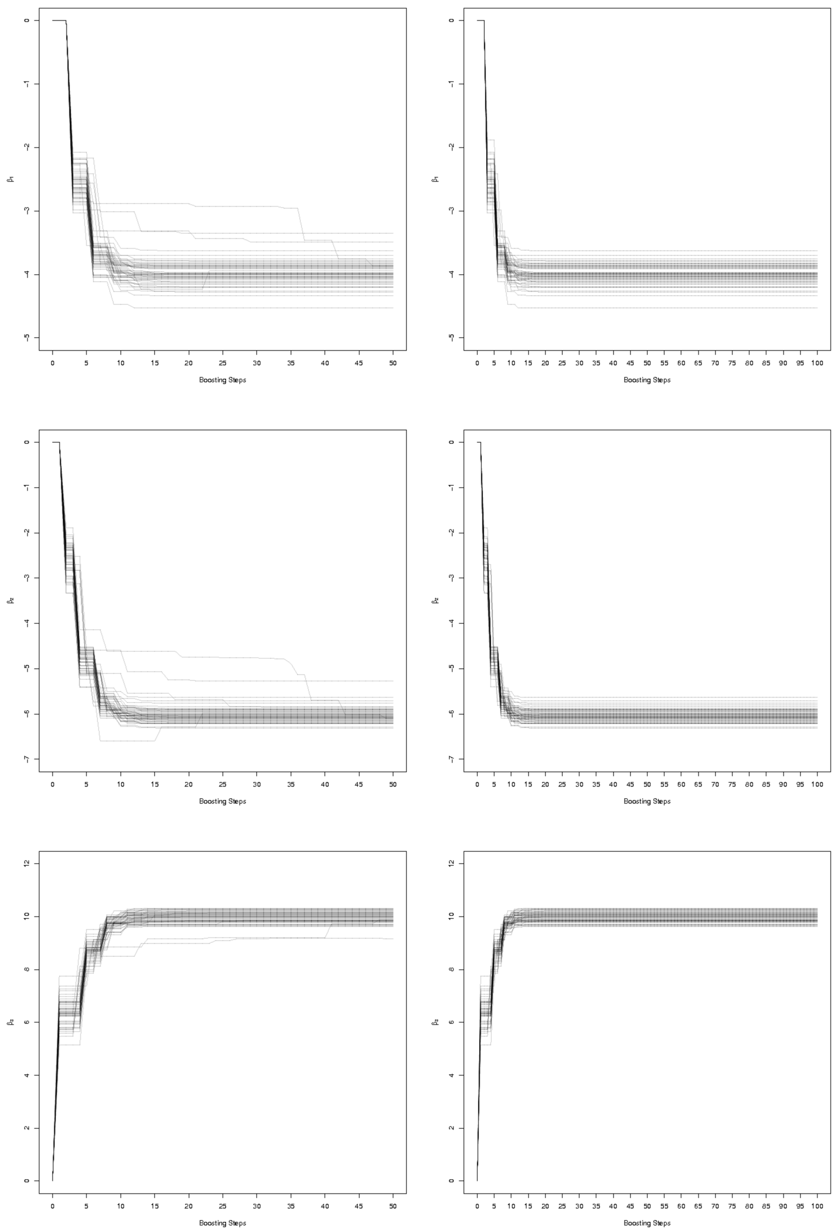

Fig. 3 Coefficient paths of β_1, β_2 and β_3 calculated by `bGLMM` algorithm for the generalized linear mixed Poisson EM (left) and REML (right) model in the $p = 20$ case

sults achieved via the `glmmPQL` function (Wood 2006). Therefore we use the same goodness-of-fit criteria as in the Poisson case.

Table 2 Generalized linear mixed model (`glmmPQL`) and boosting (`bGLMM`) on Bernoulli data

	`glmmPQL`		`bGLMM` (EM)				`bGLMM` (REML)			
p	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	falsepos	falseneg	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	falsepos	falseneg
3	9.70	0.016	36.66	0.552	0	0.84	36.92	0.043	0	0.86
5	19.80	0.014	36.66	0.553	0.02	0.82	36.93	0.044	0.02	0.86
10	44.92	0.012	39.93	0.554	0.10	0.82	37.92	0.036	0.10	0.86
20	90.82	0.015	34.29	0.553	0.12	0.66	35.08	0.029	0.14	0.72
50	294.01	0.030	48.39	0.525	0.46	0.66	45.08	0.016	0.46	0.60

The results for varying number p of covariates $x_{it1}, \ldots, x_{itp}$ and for the two different estimation methods for the random effects variance-covariance components $\mathbf{Q}$ are summarized in Table 2. Table 2 as well as Figures 4 to 5 show that in the Bernoulli case boosting is less convincing than in the Poisson case, in particular in terms of mse_{σ_b}. But the general trend, that, in case of many covariates, the $\boldsymbol{\beta}$-fit that is achieved using the `bGLMM` algorithm outperforms the fit obtained by the `glmmPQL` function, can still be observed. When variable selection is needed boosting estimates of $\boldsymbol{\beta}$ are distinctly better than estimates obtained by the `glmmPQL` function.

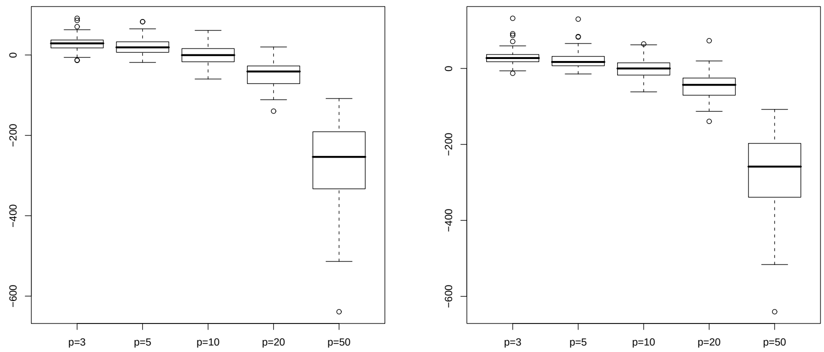

Fig. 4 Boxplots of $(\text{mse}_{\boldsymbol{\beta}}^{\text{bGLMM}} - \text{mse}_{\boldsymbol{\beta}}^{\text{glmmPQL}})$ for the EM model (left) and the REML model (right)

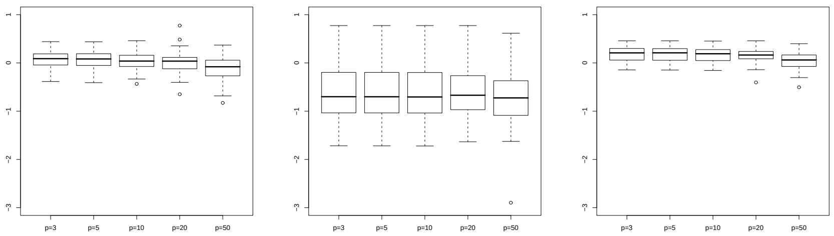

Fig. 5 Boxplots of $\triangle_{\sigma_b}$ for the `glmmPQL` model (left), for the `bGLMM` EM model (middle) and for the `bGLMM` REML model (right)

4 Application to CD4 Data

The data were collected within the Multicenter AIDS Cohort Study (MACS). In the study about 5000 infected gay or bisexual men from Baltimore, Pittsburgh, Chicago and Los Angeles have been observed since 1984 (see Kaslow et al. 1987, Zeger & Diggle 1994). The human immune deficiency virus (HIV) causes AIDS by attacking an immune cell called the CD4+ cell which coordinates the body's immunoresponse to infectious viruses and hence reduces a person's resistance against infection. According to Diggle et al. (2002) an uninfected individual has around 110 cells per milliliter of blood and since the number of CD4+ cells decreases with time from infection, one can use an infected person's CD4+ cell number to check disease progression. Within the MACS, $n = 369$ seroconverters with a total of $\sum_{i=1}^{n} T_i = 2376$ measurements were included with the number of CD4+ cells being the interesting response variable. Covariates include years since seroconversion ranging from 3 years before to 6 years after seroconversion, packs of cigarettes a day, recreational drug use (yes/no), number of sexual partners, age and a mental illness score (cesd). For observation t of individual i, the model that is considered in the following has the form

$$\begin{aligned} g(\mu_{it}) &= \beta_0 + \eta_{it}^{\text{par}} + \eta_{it}^{\text{rand}} \\ &= \beta_0 + \text{time}_{it}\beta_1 + \text{time}_{it}^2\beta_2 + \text{drugs}_{it}\beta_3 + \text{partners}_{it}\beta_4 \\ &\quad + \text{cigarettes}_{it}\beta_5 + \text{cesd}_{it}\beta_6 + \text{age}_{it}\beta_7 + b_i, \end{aligned}$$

with $b_i \sim N(0, \sigma_b^2)$. Our main objective is the typical time course of CD4+ decay and the variability across subjects. As the time effect may be nonlinear, we additionally consider the covariate "squared time". We fit an overdispersed Poisson model with natural link. The overdispersion parameter Φ is estimated by use of Pearson residuals

$$\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{(v(\hat{\mu}_{it}))^{\frac{1}{2}}}$$

by

$$\hat{\Phi} = \frac{1}{N - \text{trace}(\mathbf{H})} \sum_{i=1}^{n} \sum_{1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i.$$

For the estimation procedure we have standardized all covariates. The results for the `bGLMM` algorithm and for the `glmmPQL` function are given in Table 3. It is seen that the two boosting algorithms yield nearly the same estimates. The incorporated selection procedure suggests that drug use, pack of cigarettes a day and age are not needed in the predictor.

The maximal number of boosting steps has been chosen as $l_{max} = 100$ and the algorithm selected $l_{opt} = 19$ as optimal number of boosting steps. Coefficients build ups-for coefficients are found in Figure 6, with the vertical line indicating the optimal stopping point l_{opt}. It is seen that coefficient estimates are very stable after about 10 boosting steps.

Table 3 Estimates for the AIDS Cohort Study MACS with `glmmPQL` function (standard deviations in brackets) and `bGLMM` algorithm

	`glmmPQL`	`bGLMM` (EM)	`bGLMM` (REML)
Intercept	6.5547 (0.018)	6.5362	6.5362
Time	-0.2210 (0.011)	-0.2191	-0.2191
Time^2	-0.0156 (0.010)	-0.0197	-0.0197
Drugs	0.0126 (0.010)	0	0
Partners	0.0385 (0.010)	0.0568	0.0568
Packs of Cigarettes	0.057 (0.013)	0	0
Mental illness score (cesd)	-0.0304 (0.010)	-0.0388	-0.0388
Age	0.0020 (0.018)	0	0
σ_b^2	0.3025	0.3549	0.3539
Φ	66.8224	76.0228	76.0228

5 Concluding Remarks

Algorithms are derived that allow to estimate generalized mixed models with high-dimensional predictor structure. The incorporated selection procedure reduces the predictor space when redundant variables are present. Although penalized quasi-likelihood estimators work also in cases up to 50 predictors, performance deteriorates when many spurious variables are present. In these cases boosting approaches show better performance even in the binary response case. For low-dimensional settings boosting for binary responses still needs to be improved.

The approach proposed here can be extended to incorporate nonparametric effects. Let $\mathbf{u}_{it}^T = (u_{it1}, \ldots, u_{itm})^T$ be the covariate vector consisting of m different covariates associated with these nonparametric effects. The generalized semiparametric mixed model has the form

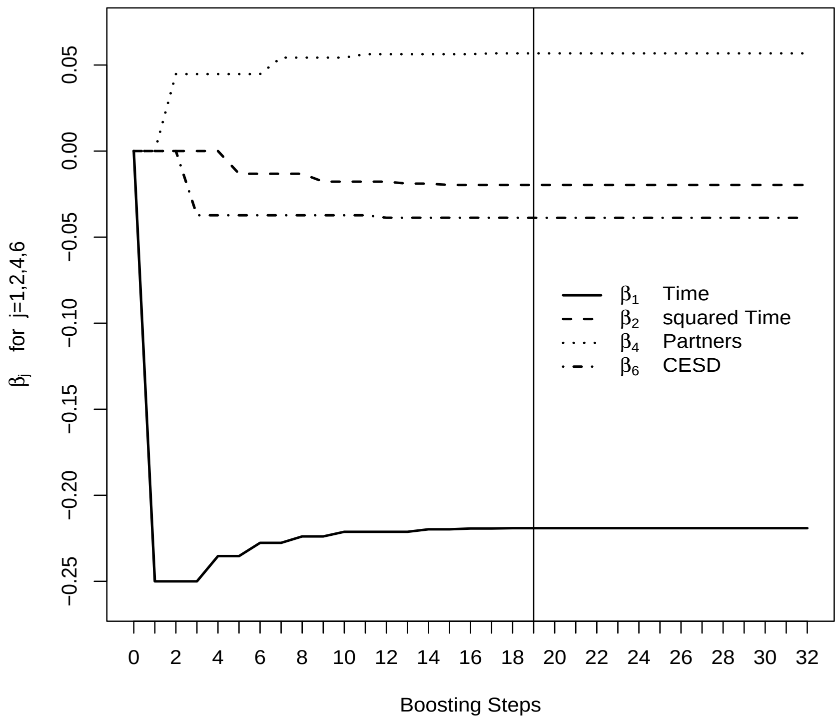

Fig. 6 Coefficient paths of $\beta_i \neq 0$ calculated by the generalized linear mixed Poisson REML model

$$\begin{aligned} g(\mu_{it}) &= x_{it}^T \boldsymbol{\beta} + \sum_{j=1}^{m} \alpha_{(j)}(u_{itj}) + w_{it}^T \mathbf{b}_i \\ &= \eta_{it}^{\text{par}} + \eta_{it}^{\text{add}} + \eta_{it}^{\text{rand}}, \end{aligned}$$

where g is a monotonic differentiable link function, $\eta_{it}^{\text{par}} = x_{it}^T \boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_p)$, now including the intercept, $\eta_{it}^{\text{add}} = \sum_{j=1}^{m} \alpha_{(j)}(u_{itj})$ is an additive term with unspecified influence functions $\alpha_{(1)}, \dots, \alpha_{(m)}$ and finally $\eta_{it}^{\text{rand}} = w_i^T \mathbf{b}_i$ contains the cluster-specific random effects $\mathbf{b}_i \sim N(0, \mathbf{Q})$, where $\mathbf{Q}$ is a known or unknown covariance matrix. By expanding nonparametric effects in basis functions and using a weak learner that refers to the updating of all coefficients corresponding to one nonparametric effect the model may be estimated with an incorporated selection procedure.

References

Booth, J. G. & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *J. R. Statist. Soc* **B 61**: 265–285.

Breiman, L. (1998). Arcing classifiers, *Annals of Statistics* **26**: 801–849.

Breiman, L. (1999). Prediction games and arcing algorithms, *Neural Computation* **11**: 1493–1517.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed model, **88**: 9–25.

Breslow, N. E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**: 81–91.

Bühlmann, P. & Hothorn, T. (2008). Boosting algorithms: regularization, prediction and model fitting, *Statistical Science* . accepted.

Bühlmann, P. & Yu, B. (2003). Boosting with the L2 loss: Regression and classification, *Journal of the American Statistical Association* **98**: 324–339.

Diggle, P. J., Heagerly, P., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*, Oxford University Press, Oxford.

Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors, *Applied Statistics* **50**: 201–220.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer-Verlag, New York.

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm, *Proceedings of the Thirteenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp. 148–156.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics* **29**: 337–407.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall, London.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. & Rinaldo, C. R. (1987). The multicenter aids cohort study: rationale, organization and selected characteristic of the participants, *American Journal of Epidemiology* **126**: 310–318.

Leitenstorfer, F. (2008). *Boosting in Nonparametric Regression: Constrained and Unconstrained Modeling Approaches*, Verlag Dr. Hut, München.

Lin, X. & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion, **91**: 1007–1016.

Littell, R., Milliken, G., Stroup, W. & Wolfinger, R. (1996). *SAS System for Mixed Models*, SAS Institue Inc., Cary, NC.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear and Mixed Models*, Wiley, New York.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*, Springer, New York.

Schall, R. (1991). Estimation in generalised linear models with random effects, *Biometrika* **78**: 719–727.

Tutz, G. & Binder, H. (2007). Generalized additive modelling with implicit variable selection by likelihood based boosting, *Biometrics* .

Tutz, G. & Reithinger, F. (2006). A boosting approach to flexible semiparametric mixed models, *Statistics in medicine* **26**: 2872–2900.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, fourth edn, Springer, New York.

Vonesh, E. F. (1996). A note on the use of laplace's approximatio for nonlinear mixed-effects models, *Biometrika* **83**: 447–452.

Wolfinger, R. W. (1994). Laplace's approximation for nonlinear mixed models, *Biometrika* **80**: 791–795.

Wolfinger, R. W. & O'Connell, M. (1993). Generalized linear mixed models; a pseudolikelihood approach, *Journal Statist. Comput. Simulation* **48**: 233–243.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall, London.

Zeger, S. L. & Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters, *Biometrics* **50**: 689–699.