Tutz, Reithinger:

Flexible semiparametric mixed models

Sonderforschungsbereich 386, Paper 448 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

Flexible Semiparametric Mixed Models

Gerhard Tutz & Florian Reithinger
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
{tutz,flo}@stat.uni-muenchen.de

27.07.2005

Abstract

In linear mixed models the influence of covariates is restricted to a strictly parametric form. With the rise of semi- and nonparametric regression also the mixed model has been expanded to allow for additive predictors. The common approach uses the representation of additive models as mixed models. An alternative approach that is proposed in the present paper is likelihood based boosting. 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. Likelihood based boosting is a general method which may be seen as an extension of L2 boost. In additive mixed models the advantage of boosting techniques in the form of componentwise boosting is that it is suitable for high dimensional settings where many influence variables are present. It allows to fit additive models for many covariates with implicit selection of relevant variables and automatic selection of smoothing parameters. Moreover, boosting techniques may be used to incorporate the subject-specific variation of smooth influence functions by specifying "random slopes" on smooth effects. This results in flexible semiparametric mixed models which are appropriate in cases where a simple random intercept is unable to capture the variation of effects across subjects.

Keywords: Mixed Model, boosting, random slopes, additive models, smoothing

1 Introduction

There is an extensive body of literature on the linear mixed model, early highlights being Henderson (1953), Laird & Ware (1982) and Harville (1977). Nice overviews including more recent work are found in Verbeke & Molenberghs (2001), McCulloch & Searle (2001). In common linear mixed models the influence of covariates is restricted to a strictly parametric form. While in regression models much work has been done to extend the strict parametric form to more flexible forms of semi- and nonparametric regression, much less has been done to develop flexible mixed model. For overviews on semiparametric regression models see Hastie & Tibshirani (1990), Green & Silverman (1994) and Schimek (2000).

A first step to more flexible mixed models is the generalization to additive mixed models where a random intercept is included. With response y_{it} for observation t on individual i and covariates $u_{i1}, \dots u_{im}$ the basic form is

$$y_{it} = \beta_0 + \alpha_{(1)}(u_{i1}) + \cdots + \alpha_{(m)}(u_{im}) + b_{i0} + \varepsilon_{it} \tag{1}$$

where $\alpha_{(1)}(.), \ldots, \alpha_{(m)}(.)$ are unspecified functions of covariates $u_{i1}, \ldots, u_{im}, b_{i0}$ is a subject-specific random intercept with $b_{i0} \sim N(0, \sigma_b^2)$ and ε_{it} is an additional noise variable. Estimation for this model may be based on the observation that regression models with smooth components may be fitted by mixed model methodology. Speed (1991) indicated that the fitted cubic smoothing splines is a best linear unbiased predictor. Subsequently the approach has been used in several papers to fit mixed models, see e.g. Verbyla, Cullis, Kenward & Welham (1999), Parise, Wand, Ruppert & Ryan (2001), Lin & Zhang (1999), Brumback & Rice (1998), Zhang, Lin, Raz & Sowers (1998),Wand (2003). Bayesian approaches have been considered e.g. by Fahrmeir & Lang (2001).

In model (1) it is assumed that the effects of covariates do not vary across individuals. This restriction is rather severe. For example in the analysis of growth curves it has to be assumed that not only the starting values differ for individuals but also that the speed of growth depends on individuals. In order to avoid overparameterization we propose to model the variation across individuals by random "slopes" of smooth functions. A simple model of this type is

$$y_{it} = \beta_o + \alpha(u_i) + b_{i0} + b_{i1}\alpha(u_i) + \varepsilon_{it}, \tag{2}$$

where for simplicity only one variable u_i is considered which has the smooth effect $\alpha(u_i)$ but which is modified by the random effect b_{i1}. By assuming $(b_{i0}, b_{it}) \sim N(0, \Sigma_b)$ the intercept as well as the slope are considered as subject-specific random effects. Model (2) raises problems in estimation since the random slopes b_{i1} are linked to the function $\alpha(u_i)$ which is not a known predictor as in the usual linear mixed model. In order to illustrate flexible models we will consider the following examples.

Application 1: AIDS Study

The data were collected within the Multicenter AIDS Cohort Study (MACS), which has followed nearly 5000 gay or bisexual men from Baltimore, Pittsburgh, Chicago and Los Angeles since 1984 (see Kaslow, Ostrow, Detels, Phair, Polk & Rinaldo (1987), Zeger & Diggle (1994)). The study includes 1809 men who were infected with HIV when the study began and another 371 men who were seronegative at entry and seroconverted during the follow-up. In the study 369 seroconverters with 2376 measurements in total were used, two subjects were dropped since covariate information was not available. The interesting response variable is the number or percent of CD4 cells by which progression of disease may be assessed. Covariates include years since seroconversion, packs of cigarettes a day, recreational drug use (yes/no), number of sexual partners, age and a mental illness score. Zeger & Diggle (1994) motivate extensively the interest in the typical time course of CD4 cell decay and the variability across subjects. Since the forms of the effects is not known, time since seroconversion, age and the mental illness score may be considered as unspecified additive effects. Figure 1 shows the smooth effect of time on CD4 cell decay for a random intercept model together with the data, Figure 2 shows the observations for three men with differing number of observed time points (dashed lines) and the fitted curves for individual time decay (for details see Section 3).

Application 2: Jimma Study

The Jimma Infant Survival Differential Longitudinal Study which is extensively described in Lesaffre, Asefa & Verbeke (1999) is a cohort study examining the live births which took place during a one year period from September 1992 until September 1993 in Ethiopia. The study involves about 8000 households with live births in that period. The children were followed up for one year to determine the risk factors for infant mortality. Following Lesaffre, Asefa & Verbeke (1999) we consider 495 singleton live births from the town of Jimma and look for the determinants of growth of the children in terms of body weight (in kg). Weight has been measured at delivery and repeatedly afterwards. In addition we consider the socio-economic and demographic covariates age of mother in years (AGEM), educational level of mother (0-5: illiterate, read and write, elementary school, junior high school, high school, college and above), place of delivery (DELIV,1-3: hospital, health center, home), number of antenatal visits (VISIT,

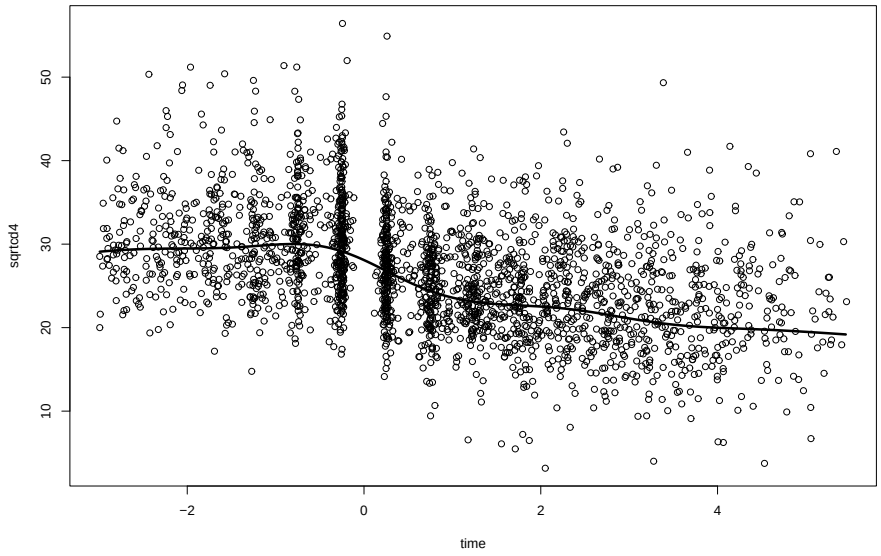


Figure 1: Smoothed time effect on the CD4 cell from Multicenter AIDS Cohort Study (MACS)

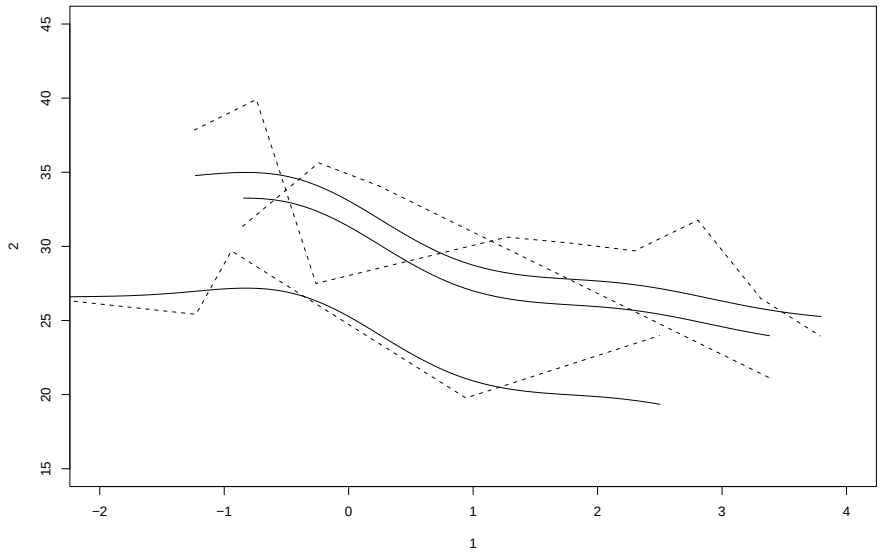


Figure 2: Smoothed time effect on the CD4 cell from Multicenter AIDS Cohort Study (MACS) and the decay of CD4 cells of 3 members of the study over time

0,⩾1), month of birth (TIME,1:Jan.-June, 0:July-Dec.), sex of child (1:male, 0:female). For more details and motivation of the study see Lesaffre, Asefa & Verbeke (1999). Figure 3 shows the overall evolution of weight and Figure 4 shows the growth curve of four children (observations and fitted curves) for an additive mixed model with random slopes on the additive age effect. It is seen that random slopes are definitely necessary for modelling since speed of growth varies strongly across children.

In the following estimation procedures are proposed for the additive model as well as for models with random slopes on smooth effects. In the case of the additive model they may be seen as an alternative to the approach based on Speed's (1991) observation. To our knowledge the fitting of random slopes of smooth effects has not been considered in the literature because of the underlying multiplicative structure of effects. The proposed estimation procedures for both semiparametric structures are based on boosting techniques which for regression models have been suggested e.g. by Bühlmann & Yu (2003). It is shown that in particular in high dimensional settings boosting approaches yield quite efficient estimation procedures for additive models and they may be modified to allow random slopes of curves.

In Section 2 the additive mixed model is considered. It is shown how estimates of parameters and variances are obtained for given smoothing parameters and how smoothing parameters may be obtained as ML or REML estimates. In Section 3 the proposed estimators are outlined. By utilizing boosting techniques the selection of potentially high dimensional smoothing parameters is avoided and stable estimates are obtained. It is shown that in particular in high dimensional problems where many unspecified procedures of potential influence have to be considered the boosting approach outperforms alternative approaches. Moreover, the CD4 cells example is considered in section 3. In Section 4 the modelling of random slopes of smooth effects is considered and an estimation technique is given that is able to cope with the multiplicative structure of effects. In Section 5 extensions to varying-coefficients models and interactions is briefly sketched.

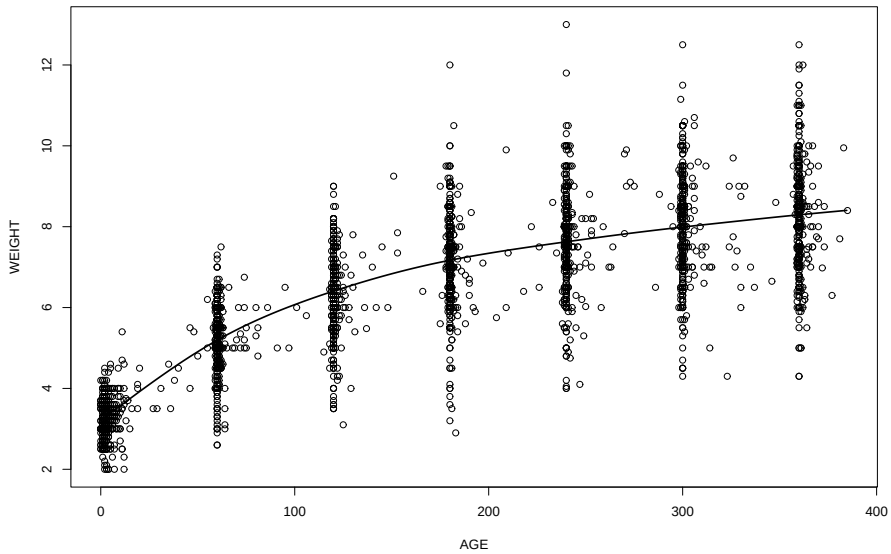


Figure 3: Evolution of average weight(kg) as function of age

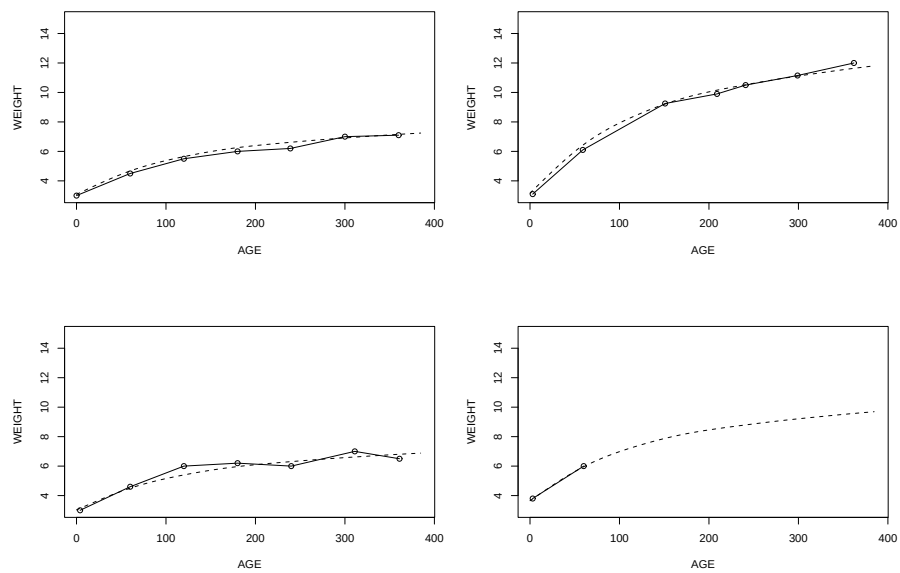


Figure 4: Individual infant curves (observed and predicted)

2 Additive Mixed Model

2.1 The Model

Let the data be given by $(y_{it}, x_{it}, u_{it}, w_{it}), \quad i = 1, \ldots, n, \quad t = 1, \ldots, T_i$, where y_{it} is the response for observation t within cluster i and $x_{it}^T = (x_{it1}, \ldots, x_{itp}), \quad u_{it}^T = (u_{it1}, \ldots, u_{itm}), \quad w_{it}^T = (w_{it1}, \ldots, w_{its})$ are vectors of covariates, which may vary across clusters and observations. The semiparametric mixed model that is considered in the following has the general form

$$\begin{aligned} y_{it} &= x_{it}^T\beta + \sum_{j=1}^{m} \alpha_{(j)}(u_{itj}) + w_{it}^T b_i + \epsilon_{it} \\ &= \mu_{it}^{par} + \mu_{it}^{add} + \mu_{it}^{rand} + \epsilon_{it} \end{aligned} \tag{3}$$

where

$\mu_{it}^{par} = x_{it}^T\beta$ is a linear parametric term,

$\mu_{it}^{add} = \sum_{j=1}^{m} \alpha_{(j)}(u_{it_j})$ is an additive term with unspecified influence functions $\alpha_{(1)}, \ldots, \alpha_{(m)}$,

$\mu_{it}^{rand} = w_{it}^T b_i$ contains the cluster-specific random effect $b_i, \quad b_i \sim N(0, Q(\rho))$, where $Q(\rho)$ is a parameterized covariance matrix and

ϵ_{it} is the noise variable, $\epsilon_{it} \sim N(0, \sigma^2 I), \quad \epsilon_{it}, b_i$ independent.

In spline methodology the unknown functions $\alpha_{(j)}$ are approximated by basis functions. A simple basis is known as the truncated power series basis of degree d, yielding

$$\alpha_{(j)}(u) = \gamma_0^{(j)} + \gamma_1^{(j)} u + \ldots \gamma_d^{(j)} u^d + \sum_{s=1}^{M} \alpha_s^{(j)} (u - k_s^{(j)})_+^d \ ,$$

where $k_1^{(j)} < \ldots < k_M^{(j)}$ are distinct knots. More generally one uses

$$\alpha_{(j)}(u) = \sum_{s=1}^{M} \alpha_s^{(j)} \Phi_s^{(j)}(u) = \alpha_j^T \Phi_j(u) \tag{4}$$

where $\Phi_s^{(j)}$ denotes the s-th basis function for variable $j, \alpha_j^T = (\alpha_1^{(j)}, \ldots, \alpha_M^{(j)})$ are unknown parameters and $\Phi_j(u)^T = (\Phi_1^{(j)}(u), \ldots, \Phi_M^{(j)}(u))$ represent the vector-valued evaluations of the basis functions.
For semi- and nonparametric regression models Eilers & Marx (1996), Marx & Eilers (1998) propose the numerically stable B-splines which have also been used by Wood (2004). For further investigation of basis functions see also Wand (2000), Ruppert & Carroll (1999).

By collecting observations within one cluster the model has the form

$$\begin{aligned} y_i &= Z_i\beta + \Phi_{i1}\alpha_1 + \ldots + \Phi_{im}\alpha_m + W_i b_i + \epsilon_i, \\ & \begin{bmatrix} \epsilon_i \\ b_i \end{bmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 I & \\ & Q(\rho) \end{pmatrix}\right), \end{aligned} \tag{5}$$

where $Z_i\beta$ contains the linear term, $\Phi_{ij}\alpha_j$ represents the one additive term and $W_i\beta$ the random term. Vectors and matrices are given by $y_i^T = (y_{i1}, \ldots, y_{iT_i})$, $Z_i^T = (x_{i1}, \ldots, x_{iT_i}), \Phi_{ij}^T = (\Phi_1^{(j)}(u_{i1j}), \ldots, \Phi_M^{(j)}(u_{iT_ij}))$, $W_i^T = (w_{i1}, \ldots, w_{iT_i})$, $\epsilon_i^T = (\epsilon_{i1}, \ldots, \epsilon_{iT_i})$. In the case of the truncated power series the "fixed" term $\gamma_0^{(j)} + \gamma_1^{(j)}u + \ldots + \gamma_d^{(j)}u^d$ is taken into the linear term $Z_i\beta$ without specifying Z_i and β explicitly.
In matrix form one obtains

$$y = Z\beta + \Phi_1\alpha_1 + \ldots + \Phi_m\alpha_m + Wb + \epsilon$$

where $y^T = (y_1^T, \ldots, y_n^T)$, $b^T = (b_1^T, \ldots, b_n^T)$, $\epsilon^T = (\epsilon_1^T, \ldots, \epsilon_n^T)$,
$Z^T = (Z_1^T, \ldots, Z_n^T)$, $\Phi_j^T = (\Phi_{1j}^T, \ldots, \Phi_{nj}^T)$, $W^T = (W_1^T, \ldots, W_n^T)$.
Parameters to be estimated are the fixed effects, collected in $\delta^T = (\beta^T, \alpha_1^T, \ldots, \alpha_m^T)$ and the variance specific parameters $\theta^T = (\sigma_\varepsilon, \rho^T)$ which determine the covariances $cov(\epsilon_{it}) = \sigma_\varepsilon^2 I_{T_i}$ and $cov(b_i) = Q(\rho)$. In addition one wants to estimate the random effects b_i. Since b_i is a random variable the latter is often called prediction rather than estimation.

2.2 Penalized Maximum Likelihood Approach

Starting from the marginal version of the model

$$y_i = Z_i\beta + \Phi_{i1}\alpha_1 + \ldots + \Phi_{im}\alpha_m + \epsilon_i^*, \tag{6}$$

$$\epsilon_i^* \sim N(0, V_i(\theta)), \quad V_i(\theta) = \sigma^2 I_{T_i} + W_i Q(\rho) W_i^T,$$

estimates for δ may be based on the *penalized log-likelihood*

$$l_p(\delta;\theta) = -\frac{1}{2}\sum_{i=1}^n \log(|V_i(\theta)|) - \sum_{i=1}^n \frac{1}{2}(y_i - Z_i\delta)^T (V_i(\theta))^{-1}(y - Z_i\delta) - \frac{1}{2}\delta^T K\delta, \tag{7}$$

where $\delta^T K\delta$ is a penalty term which penalized the coefficients $\alpha_1, \ldots, \alpha_n$. For the truncated power series an appropriate penalty is given by

$$K = Diag(0, \lambda_1 I, \ldots, \lambda_m I),$$

where I denotes the identity matrix and λ_j steers the smoothness of the function $\alpha_{(j)}$. For $\lambda_j \to \infty$ a polynomial of degree d is fitted. P-splines ((Eilers & Marx 1996)) use $K = D^T D$ where D is a matrix that builds the difference between adjacent parameters yielding the penalty $\delta^T K\delta = \Sigma_j \lambda_j \Sigma_s (\alpha_{s+1}^{(j)} - \alpha_s^{(j)})^2$ or higher differences.

From the derivative of $l_p(\delta, \theta)$ one obtains the estimation equation $\partial l_p(\delta, \phi)/\partial\delta = 0$ which yields

$$\sum_{i=1}^n (Z_i^T (V_i(\theta))^{-1} y_i) = (\sum_{i=1}^n (Z_i^T (V_i(\theta))^{-1} Z_i + K)^{-1})\hat{\delta}$$

and therefore

$$\hat{\delta} = (\sum_{i=1}^{n}(Z_i^T(V_i(\theta))^{-1}Z_i + K))^{-1}\sum_{i=1}^{n} Z_i^T(V_i(\theta))^{-1}y_i$$

which depends on the variance parameters θ. It is well known that maximization of the log-likelihood with respect to θ yields biased estimates since maximum likelihood does not take into account that fixed parameters have been estimated (see Patterson & Thompson (1974)). The same holds for the penalized log-likelihood (7). Therefore for the estimation of variance parameters often restricted maximum likelihood estimates (REML) are preferred which are based on the log-likelihood

$$\begin{aligned} l_r(\delta,\theta) &= -\frac{1}{2}\sum_{i=1}^{n}\log(|V_i(\theta)|) - \frac{1}{2}\sum_{i=1}^{n}(y_i - Z_i\beta)^T V_i(\theta)^{-1}(y_i - Z_i\beta) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\log(|Z_i^T V_i(\theta) Z_i|), \end{aligned}$$

see Harville (1974), Harville (1977) and Verbeke & Molenberghs (2001).
The restricted log-likelihood differs from the log-likelihood by an additional component. One has

$$l_r(\delta,\theta) = l(\delta,\theta) - \frac{1}{2}\sum_{i=1}^{n}\log(|Z_i^T V_i(\theta) Z_i|).$$

It should be noted that for the estimation of θ the penalization term $\delta^T K\delta$ may be omitted since it has no effect. Details on REML is given in the Appendix.

BLUP Estimates

Usually one also wants estimates of the random effects. Best linear unbiased prediction (BLUP) is a framework to obtain estimates for β and $b_1,\ldots,b_n$ for given variance components. There are several ways to motivate BLUP (see Robinson (1991)). One way is to consider the joint density of y and b which is normal and maximize with respect to δ and b. By adding the penalty term $\delta^T K\delta$ one has to minimize

$$\sum_{i=1}^{n}\frac{1}{\sigma^2}(y_i - Z_{\Phi i}\delta - W_i b_i)^T(y_i - Z_{\Phi i}\delta - W_i b_i) + b_i^T Q(\rho)^{-1} b_i + \delta^T K\delta \tag{8}$$

where $Z_{\Phi i} = [Z_i, \Phi_{i1},\ldots,\Phi_{im}]$, $\bar{Q}(\rho) = Diag(Q(\rho)\ldots Q(\rho))$.
With $Z_\Phi^T = (Z_{\Phi 1}^T \ldots Z_{\Phi m}^T)$ the criterion (8) may be rewritten as

$$\frac{1}{\sigma^2}(y - Z_\Phi\delta - Wb)^T(y - Z_\Phi\delta - Wb) + b^T\bar{Q}(\rho)^{-1}b^T + \delta^T K\delta$$

which yields the "ridge regression" solution

$$\begin{bmatrix}\hat{\delta}\\ \hat{b}\end{bmatrix} = \left(C^T\tfrac{1}{\sigma_\varepsilon^2}IC + B\right)^{-1} C^T\frac{1}{\sigma_\varepsilon^2}Iy$$

with $C = (Z_\Phi, W)$ and

$$B = \begin{pmatrix} K & 0 \\ 0 & \bar{Q}(\rho)^{-1}\end{pmatrix}.$$

Some matrix derivation shows that $\hat{\delta}$ has the form

$$\hat{\delta} = (Z_\Phi^T V(\theta)^{-1} Z_\Phi + K)^{-1} Z^T V(\theta)^{-1} y,$$

where $V(\theta) = Diag(V_1(\theta) \dots V_n(\theta))$, and for the vector of random coefficients $b^T = (b_1^T, \dots, b_n^T)$ one obtains

$$\hat{b} = Q(\rho) W^T V(\theta)^{-1} (y - Z_\Phi \hat{\delta}).$$

In simpler form BLUP estimates are given by

$$\begin{aligned}
\hat{\delta} &= (\sum_{i=1}^{n} (Z_i^T (V_i(\theta))^{-1} Z_i + K))^{-1} \sum_{i=1}^{n} Z_i^T (V_i(\theta))^{-1} y_i, \\
\hat{b}_i &= Q W_i^T V_i(\theta)^{-1} (y_i - Z_{\Phi i} \hat{\delta}).
\end{aligned}$$

2.3 Mixed Model approach to Smoothing

For the computation of $\hat{\delta}$ it is necessary to specify the smoothing parameters $\lambda_1, \dots, \lambda_m$. With cross-validation techniques problems arise if the number of smooth covariates is high. An approach that works for moderate number of smooth covariates uses the ML or REML estimates of variance components. The basic concept ist to reformulate the estimation as a more general mixed model. Let us consider again the criterion for BLUP estimates (8) which has the form

$$\begin{aligned}
& \frac{1}{\sigma^2} (y - Z_\Phi \delta - Wb)^T (y - Z_\Phi \delta - Wb) + \alpha^T K_\alpha \alpha + b^T \bar{Q}(\rho)^{-1} b \\
= \quad & \frac{1}{\sigma^2} (y - Z\beta - \Phi\alpha - Wb)^T (y - Z\beta - \Phi\alpha - Wb) + (\alpha^T b^T) \begin{pmatrix} K_\alpha & 0 \\ 0 & \bar{Q}(\rho)^{-1} \end{pmatrix} \begin{pmatrix} \alpha \\ b \end{pmatrix}
\end{aligned} \tag{9}$$

where $\Phi = [\Phi_1 \dots \Phi_m]$ and K_α for the truncated power series has the form $K_\alpha = Diag(\lambda_1 I, \dots, \lambda_m I)$. Thus (9) corresponds to the BLUP criterion of the mixed model

$$y = Z\beta + \begin{bmatrix} \Phi & W \end{bmatrix} \begin{pmatrix} \alpha \\ b \end{pmatrix} + \epsilon$$

$$\text{with} \qquad \begin{pmatrix} \alpha \\ b \\ \epsilon \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_\alpha^{-1} & 0 & 0 \\ 0 & \bar{Q}(\rho) & 0 \\ 0 & 0 & \sigma_\varepsilon^2 I \end{pmatrix} \right).$$

Since $K_\alpha = Diag(\lambda_1 I, \dots, \lambda_m I)$ the smoothing parameters $\lambda_1, \dots, \lambda_m$ correspond to the variance of the random effects $\alpha_1, \dots, \alpha_m$ for which $cov(\alpha_j) = \lambda_j I$ is assumed. Thus for the purpose of estimation $\alpha_1, \dots, \alpha_m$ are treated as random effects. REML estimates yield $\hat{\lambda}_1, \dots, \hat{\lambda}_m$. For alternative basis function like B-splines some reformation is necessary to obtain the simple independence structure of the random effects (see Appendix).

3 Boosting Approach to Additive Mixed Models

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 that reweighting corresponds to minimizing iteratively a loss function (Breiman (1999), Friedman (2001)) boosting has been extended to regression problems in a L_2-estimation framework by Bühlmann & Yu (2003).In the following boosting is used to obtain estimates for the semiparametric mixed model. Instead of using REML estimates for the choice of smoothing parameters the estimates of the smooth components are obtained by using "weak learners" iteratively. The weak learner that is used is the estimate of δ based on a fixed, very large smoothing parameter λ which is used for all components. By iterative fitting of the residual and selection of components (see algorithm) the procedure adapts automatically to the possibly varying smoothness 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 smooth term $\Phi_{ij}\alpha_j$, is refitted at a time. That means that a model containing the linear term and only one smooth component is fitted in one iteration step. For simplicity we will use the notation

$$Z_{i(r)} = [Z_i \ \Phi_{ir}] \quad , \quad \delta_r^T = (\beta^T, \alpha_r^T)$$

for the design matrix with predictor $Z_{i(r)} = Z_i\beta + \Phi_{ir}\alpha_r$.
The corresponding penalty matrix is denoted by K_r, which for the truncated power series has the form

$$K_r = Diag(0, \ \lambda I).$$

BoostMixed

1. *Initialization*
   Compute starting values $\hat{\beta}^{(0)}, \hat{\alpha}_1^{(0)}, \ldots \hat{\alpha}_m^{(0)}$ and set $\eta_i^{(0)} = Z_i\hat{\beta}^{(0)} + \Phi_{i1}\hat{\alpha}_1^{(0)} + \ldots + \Phi_{im}\hat{\alpha}_m^{(0)}$.

2. *Iteration*
   For l=1,2,...

   (a) *Refitting of residuals*

      i. Computation of parameters
      For $r \in \{1, \ldots, m\}$ the model for residuals

      $$y_i - \eta_i^{(l-1)} \sim N(\eta_{i(r)}, V_i(\theta))$$

      with

      $$\eta_{i(r)} = Z_{i(r)}\delta_r = Z_i\beta + \Phi_{ir}\alpha_r$$

      is fitted, yielding

      $$\hat{\delta}_r = (\sum_{i=1}^n (Z_{i(r)}^T (V_i(\theta^{(l-1)}))^{-1} Z_{i(r)} + K_r))^{-1} \sum_{i=1}^n Z_{i(r)}^T (V_i(\theta^{(l-1)}))^{-1} (y_i - \eta_i^{(l-1)}).$$

      ii. Selection step
      Select from $r \in \{1, \ldots, m\}$ the component j that leads to the smallest $AIC_r^{(l)}$ or $BIC_r^{(l)}$ as given in Section 3.2.

      iii. Update
      Set
      $$\hat{\beta}^{(l)} = \hat{\beta}^{(l-1)} + \hat{\beta},$$
      and
      $$\hat{\alpha}_r^{(l)} = \begin{cases} \hat{\alpha}_r^{(l-1)} & \text{if } r \neq j \\ \hat{\alpha}_r^{(l-1)} + \hat{\alpha}_r & \text{if } r = j, \end{cases}$$
      $$\hat{\delta}^{(l)} = ((\hat{\beta}^{(l)})^T, \ (\hat{\alpha}_1^{(l)})^T, \ \ldots (\hat{\alpha}_m^{(l)})^T)^T.$$
      Update for $i = 1, \ldots, n$
      $$\eta_i^{(l)} = \eta_i^{(l-1)} + Z_{i(j)}\hat{\delta}_j.$$

   (b) *Computation of Variance Components*
   The computation is based on the penalized log-likelihood

   $$\begin{aligned} l_p(\theta|\eta^{(l)}; \delta_l) \quad &= -\tfrac{1}{2}\textstyle\sum_{i=1}^n \log(|V_i(\theta)|) + \sum_{i=1}^n (y_i - \eta^{(l)})^T V_i(\theta)^{-1}(y_i - \eta^{(l)}) \\ &-\tfrac{1}{2}(\hat{\delta}^{(l)})^T K \hat{\delta}^{(l)}. \end{aligned}$$

   Maximization yields $\hat{\theta}^{(l)}$.

We chose componentwise boosting techniques since they turn out to be very stable in the high dimensional case where many potential predictors are under consideration. In this case the procedure automatically selects the relevant variables and may be seen as a tool for variable selection with respect to unspecified smooth functions. In the case of few predictors one may also use boosting techniques without the selection step by refitting the residuals for the full model with design matrix $[Z_i \Phi_{i1} \dots \Phi_{im}]$.

3.2 Stopping criteria and Selection in BoostMixed

In boosting often cross-validation is used to find the appropriate complexity of the fitted model (e.g. Dettling & Bühlmann (2003), ...). In the present setting cross-validation turns out to be too time consuming to be recommended. An alternative is to use the effective degrees of freedom which are given by the trace of the hat matrix (compare Hastie & Tibshirani (1990)). In the following the hat matrix is derived.
For the derivation of the hat matrix the matrix representation of the mixed model is preferred (see (6))

$$y = Z\beta + \Phi_1\alpha_1 + \ldots + \Phi_m\alpha_m + \epsilon^*$$

$$\text{where} \quad \epsilon^* \sim N(0, V), \quad V(\theta) = Diag(V_1(\theta), \ldots, V_n(\theta)).$$

Since in one step only one component is refitted one has to consider the model for the residual refit of the rth component.

$$residual = Z_r\delta_r$$

$$\text{where} \quad Z_r^T = (Z_{1(r)}^T \dots Z_{n(r)}^T), \quad Z_{i(r)} = [Z_i \ \Phi_{ir}], \quad \delta_r^T = (\beta^T, \alpha_r^T).$$

The refit in the lth step is given by

$$\begin{aligned} \hat{\delta}_r &= \left(Z_r^T V(\theta^{(l-1)})^{-1} Z_r + \lambda K_r\right)^{-1} Z_r^T V^{-1}(\theta^{(l-1)})(y - \eta^{(l-1)}) \\ &= M_r^{(l)}(y - \eta^{(l-1)}) \end{aligned} \tag{10}$$

where

$$M_r^{(l)} = \left(Z_r^T V(\theta^{(l-1)})^{-1} Z_r + \lambda K_r\right)^{-1} Z_r^T V^{-1}(\theta^{(l-1)})$$

refers to the rth component in the lth refitting step. Then the corresponding fit has the form

$$\hat{\eta}_r^{(l)} = Z_r\hat{\delta}_r = Z_r M_r^{(l)}(y - \hat{\eta}^{(l-1)}) = H_r^{(l)}(y - \eta^{(l-1)})$$

where

$$H_r^{(l)} = Z_r M_r^{(l)}.$$

Let now j_l denote the index of the variable that is selected in the lth boosting step and $H^{(l)} = H_{je}^{(l)}$ denote the resulting "hat" matrix of the refit. One obtains with starting matrix $H^{(0)}$

$$\eta^{(1)} = H^{(0)}y + H^{(1)}(y - \hat{\eta}^{(0)}) = (H^{(0)} + H^{(1)}(I - M^{(0)}))y$$

and more general

$$\hat{\eta}^{(l)} = G^{(l)}y$$

where

$$G^{(l)} = \sum_{s=0}^{l} H^{(s)} \prod_{k=0}^{s-1} (I - H^{(k)})$$

is the global hat matrix after the lth step. It is sometimes useful to rewrite G as

$$G^{(l)} = I - \prod_{k=0}^{l} (I - H^{(k)})$$

(compare Bühlmann & Yu (2003)).
For the selection step one evaluates the hat matrices for candidates which for the rth component in the lth step have the form

$$G_r^{(l)} = G^{(l-1)} + H_r^{(l)} \prod_{k=0}^{l-1} (I - H^{(k)}).$$

Given the hat matrix $G_r^{(l)}$, complexity of the model may be determined by information criteria. When considering in the lth step the rth component one uses the criteria

$$\begin{aligned} AIC_r^{(l)} &= -2\left\{ -\frac{1}{2}\sum_{i=1}^{n} log(V(\hat{\theta}^{(l-1)})) + \sum_{i=1}^{n} (y_i - \hat{\eta}^{(l-1)})^T V_i(\hat{\theta}^{(l-1)})^{-1}(y_i - \hat{\eta}^{(l-1)}) \right\} \\ &\quad + 2\, trace\, (G_r^{(l)}), \\ BIC_r^{(l)} &= -2\left\{ -\frac{1}{2}\sum_{i=1}^{n} log(V(\hat{\theta}^{(l-1)})) + \sum_{i=1}^{n} (y_i - \hat{\eta}_i^{(l-1)})(V_k(\theta)^{(l-1)})^{-1}(y_i - \eta_i^{(l-1)}) \right\} \\ &\quad + 2\, trace\, G_r^{(l)}\, log(n) \end{aligned}$$

In the rth step one selects from $r \in \{1, \ldots, m\}$ the component that minimizes $AIC_r^{(l)}$(or $BIC_r^{(l)}$) and obtains $AIC^{(l)} = AIC_{j_l}^{(l)}$ if j_l is selected in the rth step. If $AIC_r^{(l)}$(or $BIC_r^{(l)}$) is larger than the previous criterion $AIC^{(l-1)}$ iteration stops. It should be noted that in contrast to common boosting procedures the selection step reflects the complexity of the refitted model. In common componentwise boosting procedures (e.g. Bühlmann & Yu (2003)) one selects the component that maximally improves the fit and then evaluates if the fit including complexity of the model deteriorates. The proposed procedure selects the component in a way that the new lack-of-fit, including the augmented complexity, is minimized. In our simulations the suggested approach showed superior performance.

In the following the initialization of the boosting algorithm is shortly sketched. The basic concept is to select few relevant variables in order to obtain stable estimates of variance components. Therefore for large λ (in our appllication $\lambda = 1000$), the total model is fitted using the full design matrix $\tilde{Z} = [Z, \Phi_1, \ldots, \Phi_m]$ and covariance matrix $V_i(\theta) = I$. Then in a stepwise way the variables are selected (usually up to 5) that yield the best fit. These yield the initial estimates $\hat{\beta}^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_m^{(0)}$ and the initial hat matrix $G^{(0)}$.

3.3 Standard Errors

Approximate standard errors for the parameter β and the functions $\alpha_{(j)}(u) = \Phi_{(j)}(u)^T \alpha_j$ may be derived by considering the iterative refitting scheme. For the estimated parameter in the lth step $\delta^{(l)}$ one obtains

$$\hat{\delta}^{(l)} = \hat{\delta}^{(l-1)} + M^{(l)}(y - \hat{\eta}^{(l-1)})$$

where $M^{(l)}$ is a matrix that selects the components β and α_{j_l} which are updated in the lth step. It is given by

$$(M^{(l)})^T = \left(\ (M_{j_l,1}^{(l)})^T, 0, \ldots, (M_{j_l,2}^{(l)})^T, \ldots, 0 \ \right),$$

where $M_{j_l,1}, M_{j_l,2}$ denote the partitoning of $M_{j_l}^{(l)}$ into components that refer to β and α_{j_l} respectively, i.e.

$$M_{j_l}^{(l)} = \begin{pmatrix} M_{j_l,1} \\ M_{j_l,2} \end{pmatrix}.$$

One obtains for the refitting of δ with starting matrix $M^{(0)}$

$$\begin{aligned} \hat{\delta}^{(1)} &= M^{(0)}y + M^{(1)}(y - \hat{\eta}^{(0)}) \\ &= M^{(0)}y + M^{(1)}(I - H^{(0)})y, \end{aligned}$$

and more general

$$\hat{\delta}^{(l)} = D^{(l)} y,$$

where

$$D^{(l)} = \sum_{s=0}^{l} M^{(s)} \prod_{k=0}^{s-1} (I - H^{(k)}).$$

With L denoting the last refit one obtains with $\hat{\delta} = \hat{\delta}^{(L)}, D = D^{(L)}$, for the covariance of $\hat{\delta}$

$$\begin{aligned} cov(\hat{\delta}) &= D\ cov(y) D^T \\ &= D\ V(\theta) D^T \end{aligned}$$

Approximate variances follow by using $\hat{\theta} = \hat{\theta}^{(L)}$ to approximate $V(\theta)$. Standard errors for β and $\alpha_{(j)}(u) = \Phi_{ij}(u)^T \alpha_j$ are then easily derived since $\hat{\delta}^T = (\hat{\beta}^T, \alpha_1^T \ldots, \hat{\alpha}_M^T)$.

In boosting the crucial tuning parameter is the number of iterations. The smoothing parameter that is used in the algorithm should be chosen large to obtain a weak learner. For large λ the number of iterations increases. In order to limit the number of iterations we modified the algorithm slightly. If more than 1000 iterations are needed until the stopping criterion is met, then the algorithm is restarted with $\lambda/2$; the halving procedure is repeated if $\lambda/2$ also needs more than 1000 iterations.

3.4 Simulation Study

We present part of a simulation study in which the performance of BoostMixed models is compared to alternative approaches. The underlying model is the random intercept model

$$y_{it} = b_i + \sum_{j=1}^{40} c * f_{(j)}(u_{it}) + \epsilon_{it}, i = 1, \ldots, 80, t = 1, \ldots, 5$$

with the smooth components given by

$$\begin{aligned} f_{(1)}(u) &= sin(u) \quad && u \in [-3, 3], \\ f_{(2)}(u) &= cos(u) \quad && u \in [-2, 8], \\ f_{(3)}(u) &= u^2 \quad && u \in [-3, 3], \\ f(u) &= 0 \quad && u \in [-3, 3], j = 4, \ldots, 40. \end{aligned}$$

The vectors $u_{it}^T = (u_{it1}, \ldots, u_{it40})$ have been drawn independently with components following a uniform distribution within the specified interval. For the covariates constant correlation is assumed, i.e. $corr(y_{itr}, y_{its}) = \rho$. The constant c determines the signal strength of the covariates. The random effect and the noise variable have been specified by $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 2$ and $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 2$. In the part of the study which is presented the number of observations has been chosen by $n = 80, T = 5$.

The fit of the model is based on B-splines of degree 3 with 15 equidistant knots. The performance of estimators is evaluated separately for the structural components and the variance. By averaging across 100 datasets we consider mean squared errors for $\eta, \sigma_b^2, \sigma_\varepsilon^2$ given by

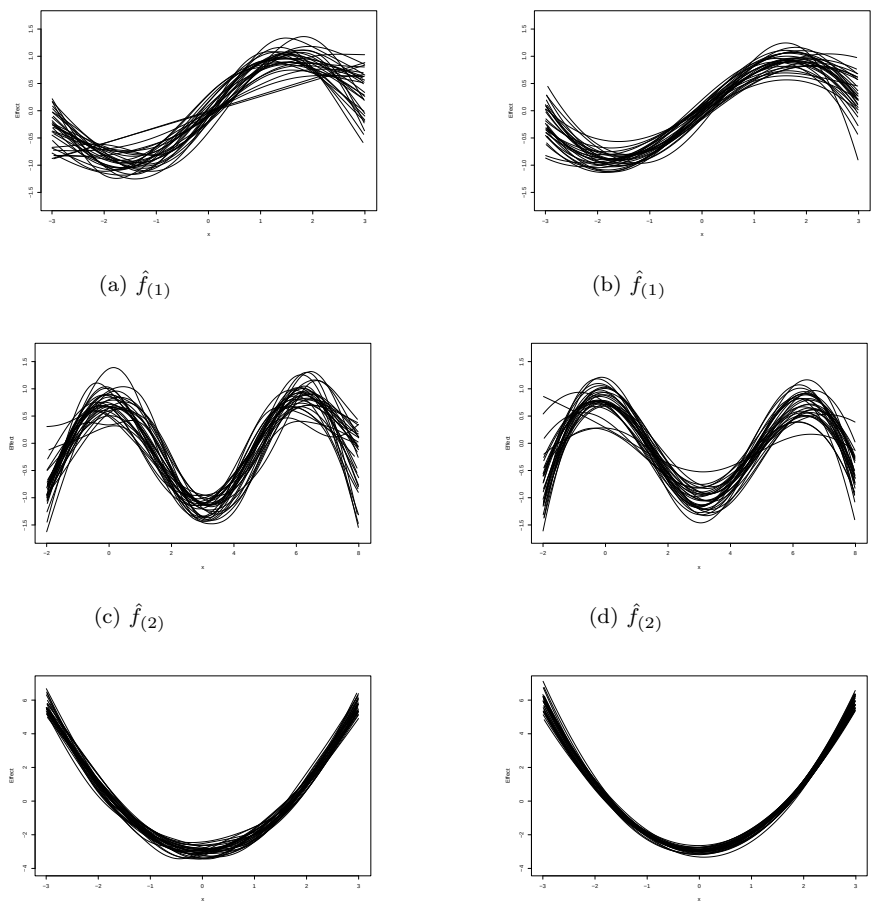
$$\begin{aligned} \text{mse}_\eta &= \textstyle\sum_{i=1}^n \sum_{t=1}^T (\eta_{it} - \hat{\eta}_{it})^2, \hat{\eta}_{it} = x_{it}^T \hat{\beta}, && \text{mse}_\beta = ||\beta - \hat{\beta}||^2, \\ \text{mse}_{\sigma_b^2} &= ||\sigma_b^2 - \hat{\sigma}_b^2||^2, && \text{mse}_{\sigma_\epsilon^2} = ||\sigma_\epsilon^2 - \hat{\sigma}_\epsilon^2||^2. \end{aligned}$$

as well as the mean squared error for the smooth components

$$\text{mse}_f = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\sum_{j=1}^{p}(f_{(j)}(u_{itj}) - \hat{f}_{(j)}(u_{itj}))^2,$$

which corresponds to the estimation of parameters in linear mixed models.

For illustration in Figure 5 the Mixed Model approach to smooth components (MM) is compared with BoostMixed for 30 datasets. It is seen that both methods detect the underlying smooth functions fairly well. However, it is seen that the mixed model approach has higher variability. For example for some datasets the component $f_{(1)}$ has been strongly oversmoothed yielding straight lines (rather than the *sin* function).



(a) $\hat{f}_{(1)}$ (b) $\hat{f}_{(1)}$

(c) $\hat{f}_{(2)}$ (d) $\hat{f}_{(2)}$

(e) $\hat{f}_{(3)}$ (f) $\hat{f}_{(3)}$

Figure 5: Thirty functions computed with mixed model methods(left panels) and boosting (right panels) $(c = 1, p = 3)$

In Tables 1 and 2 the resulting mean squared errors are given for the low correlation case ($\rho = 0.1$) and the medium correlation case ($\rho = 0.5$). It is seen that for all components mean squared errors are smaller when BoostMixed is used. The difference is rather large for high dimensional predictors which include noisy covariates. BoostMixed then has the advantage that it automatically selects the right predictors. The number of selected predictors as given in Table 1 and 2 has mean values between three and four, thus showing that selection was successful.

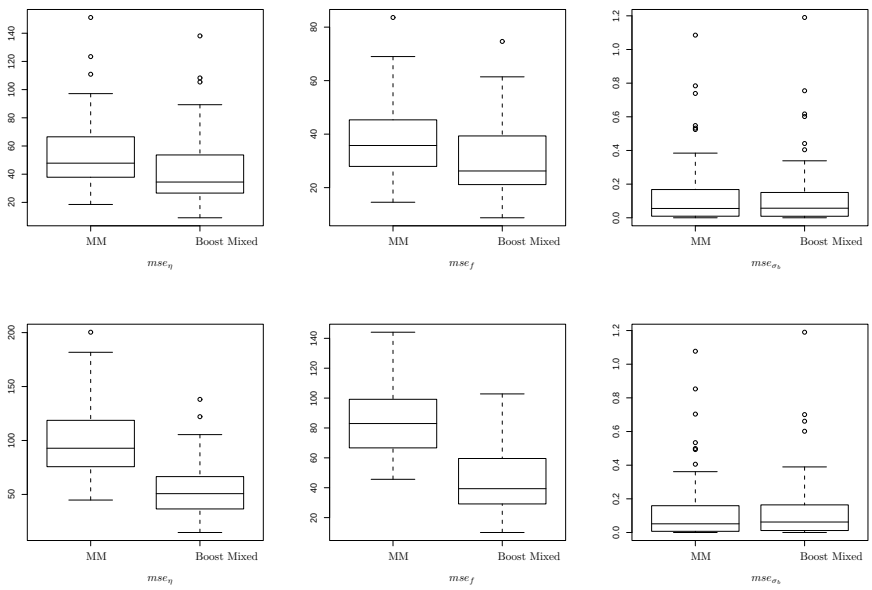


Figure 6: Boxplot for mse_η, mse_f and mse_{σ_b}additive simulation study with $c = 1$ and $p = 3$ (top) and $c = 1$ and $p = 15$ (bottom)

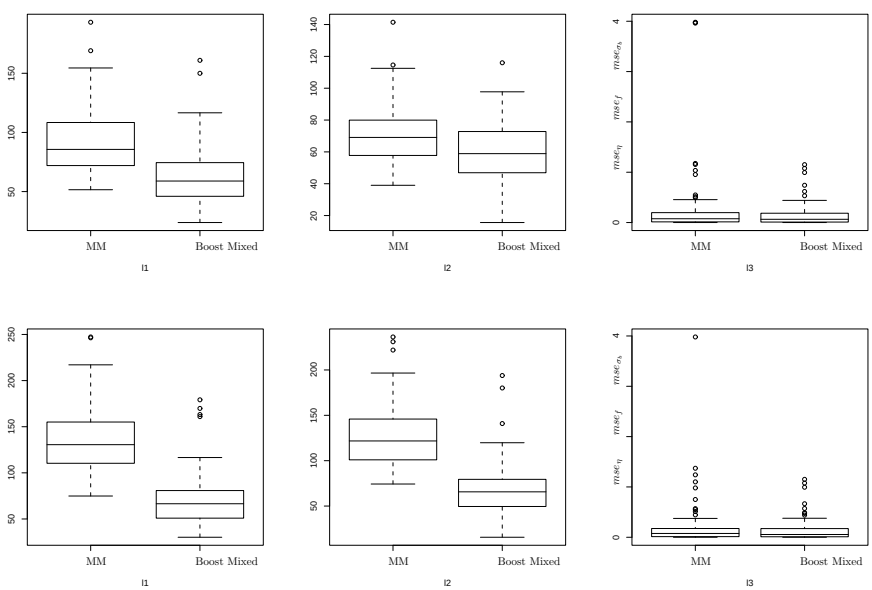


Figure 7: Boxplot for mse_η, mse_f and mse_{σ_b} additive simulation study with $c = 10$ and $p = 3$ (top) and $c = 1$ and $p = 15$ (bottom)

		MM						BoostMixed						
c	p	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Selected	Steps	Time
0.5	3	48.919	37.610	0.119	0.028	14.7	0.14	44.178	38.435	0.114	0.026	2.9	21	0.13
	6	59.117	48.360	0.119	0.029	17.9	0.52	51.380	47.964	0.112	0.028	3.2	21	0.27
	15	92.049	85.762	0.127	0.031	26.2	9.01	60.406	58.639	0.111	0.028	3.6	20	0.72
	25							70.528	70.860	0.108	0.030	3.8	19	0.93
1	3	54.240	37.535	0.124	0.024	11.0	0.10	41.470	30.457	0.119	0.026	3.0	61	0.35
	6	63.671	48.900	0.118	0.024	15.4	0.45	45.094	34.757	0.119	0.027	3.2	61	0.64
	15	97.211	85.477	0.120	0.028	21.4	7.36	53.980	45.249	0.121	0.030	3.7	62	1.62
	25							62.094	55.092	0.118	0.032	4.0	81	2.59
5	3	74.485	60.585	0.186	0.032	12.9	0.12	51.907	46.045	0.181	0.030	3.0	456	1.66
	6	85.335	72.724	0.185	0.031	14.3	0.42	52.756	47.277	0.181	0.031	3.0	457	3.25
	15	119.919	114.034	0.188	0.036	20.2	6.97	57.385	53.415	0.177	0.035	3.2	464	8.94
	25							61.299	58.286	0.176	0.038	3.4	464	13.43
10	3	91.144	71.836	0.264	0.026	13.8	0.13	62.942	60.312	0.140	0.029	3.0	1834	5.51
	6	101.424	83.810	0.186	0.026	15.1	0.44	64.687	62.413	0.131	0.029	3.0	1833	12.38
	15	135.990	126.305	0.197	0.034	19.5	6.76	70.245	69.627	0.137	0.034	3.3	1814	22.03
	25							77.409	78.312	0.138	0.036	3.6	1812	32.19

Table 1: Comparison between additive mixed model fit and BoostMixed ($\rho = 0.1$).

		MM						BoostMixed						
c	p	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Steps	Time	mse_η	mse_f	mse_{σ_b}	$\text{mse}_{\sigma_\epsilon}$	Selected	Steps	Time
0.5	3	50.605	35.003	0.133	0.023	16.1	0.1	44.400	34.979	0.134	0.025	2.9	27.9	0.1
	6	61.019	48.787	0.134	0.024	18.6	0.5	51.515	44.691	0.134	0.026	3.2	27.2	0.3
	15	94.837	93.356	0.134	0.031	29.6	10.0	64.801	61.745	0.132	0.028	3.7	25.1	0.8
	25							73.055	72.109	0.131	0.031	4.0	23.5	1.1
1	3	53.324	41.481	0.147	0.034	11.3	0.1	39.049	32.489	0.144	0.034	3.0	55.2	0.3
	6	64.692	55.124	0.147	0.037	16.7	0.4	42.398	36.626	0.145	0.036	3.1	55.5	0.5
	15	96.471	98.067	0.150	0.039	22.1	7.5	50.575	47.069	0.143	0.037	3.6	55.8	1.4
	25							56.293	54.101	0.146	0.037	3.9	56.3	1.9
5	3	76.088	63.533	0.155	0.024	12.8	0.1	52.205	48.831	0.154	0.025	3.0	385.0	1.4
	6	86.457	77.309	0.155	0.025	13.9	0.4	53.503	50.494	0.155	0.026	3.0	385.3	2.6
	15	118.606	119.500	0.159	0.029	17.7	6.1	56.342	54.229	0.154	0.027	3.2	382.5	6.5
	25							60.006	58.927	0.152	0.028	3.3	376.4	10.6
10	3	96.354	77.674	0.188	0.028	13.7	0.1	67.639	63.196	0.185	0.028	3.0	1568.7	4.7
	6	108.771	93.913	0.264	0.029	15.2	0.4	69.957	66.484	0.184	0.029	3.0	1561.5	10.2
	15	143.923	141.908	0.297	0.035	18.5	6.3	75.311	73.641	0.180	0.032	3.3	1553.5	19.6
	25							83.490	84.629	0.179	0.034	3.6	1522.0	27.9

Table 2: Comparison between additive mixed model fit and BoostMixed ($\rho = 0.5$).

But it should be noted that also in the case where only the variables are included which carry information, the mean squared errors are still smaller when BoostMixed is used. For higher number of predictors (p>20) the Mixed Model fit did not work, therefore no values are shown in Table 1 and 2. The strongest reduction in terms of mean squared error is found for the estimation of mse_η the effect becomes stronger with increasing signal c and parameters p, see for example $mse_\eta = 41.470$ for BoostMixed and $mse_\eta = 54.240$ for the additive model with $c = 1$, $p = 3$. In Figure 6 and 7 the mean squared errors are given for the pure information case (p=3) and the case that includes several noise variables (p=15).

3.5 Application to CD4 data

For the AIDS Cohort Study MACS we considered the semi-parametric mixed model from Section 1

$$y_{it} = \mu_{it}^{par} + \mu_{it}^{add} + b_{it} + \epsilon_{it},$$

where y_{it} denotes the square root CD4 number of cells for subject i on measurement t (taken at irregular time intervals). The parametric and nonparametric term are given by

$$\begin{aligned} \mu_i^{\text{par}} &= \beta_0 + drugs_i\beta_D + partners_i\beta_P, \\ \mu_{it}^{\text{add}} &= \alpha_T(time) + \alpha_A(age_i) + \alpha_C(cesd). \end{aligned}$$

where *cesd* is a mental illness score. The square root transformation has been used since the original CD4 cell number varies over a wide range. This is a kind of stabilization transformation for variances. The estimated effect of time was modelled smoothly with the resulting curve given in Figure 1. This smooth curve can be compared to the results of Zeger & Diggle (1994) who applied generalized estimation equations. In Figure 8 the smooth effects of age, the mental illness score and time are given. It is seen that there is a slight increase in CD4 cells for increasing age and a decease with higher values of the mental illness score. Table 3 shows the estimates for the parameters. Comparison between BoostMixed and the mixed model approach shows that the estimates are well comparable.

	BoostMixed		Mixed Model	
Intercept	24.6121	(0.294)	24.8233	(0.286)
Drugs	0.5211	(0.279)	0.5473	(0.292)
partners	0.0633	(0.049)	0.0595	(0.034)
σ_ϵ	4.2531	-	4.26138	-
σ_b	4.3870	-	4.43180	-

Table 3: Estimates for the AIDS Cohort Study MACS with BoostMixed and mixed model approach (standard deviations given in brackets)

4 Random slopes on smooth effects

4.1 Estimation by boosting techniques

The semiparametric additive model (3) allows for additive effects of covariates, including multivariate random effects. For example random slopes for linear terms are already included. With $w_{it} = x_{it}$ model (3) becomes

$$y_{it} = \sum_{j=1}^{m} \alpha_{(j)}(u_{itj}) + x_{it}^T\beta + x_{it}^T b_i + \varepsilon_{it}$$

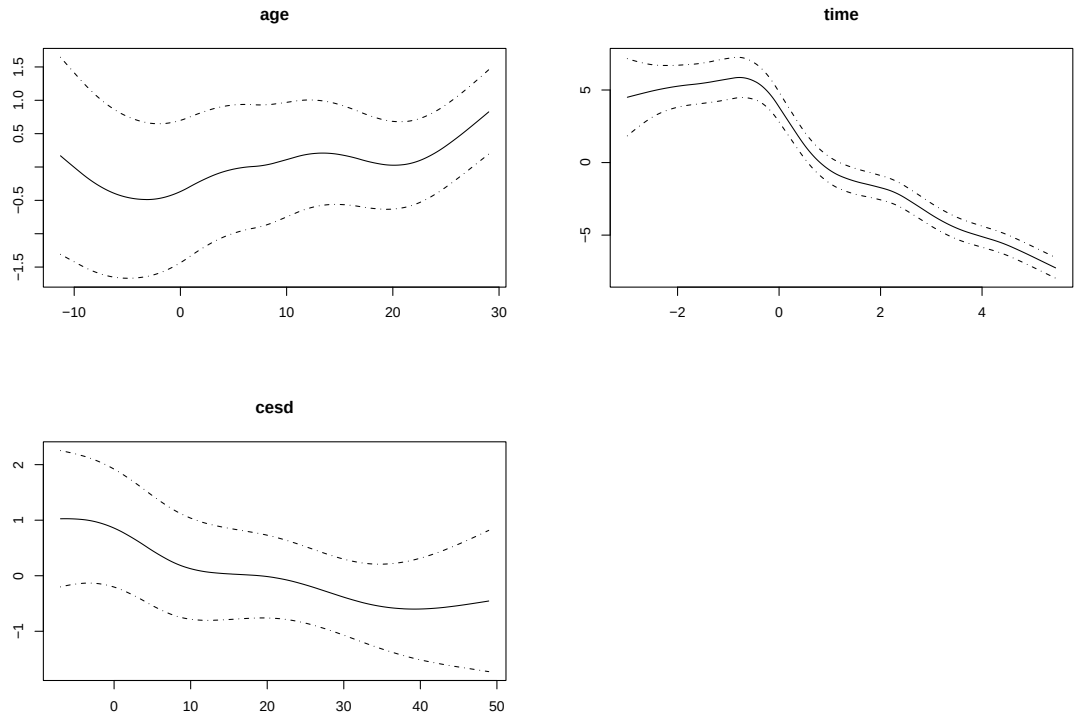


Figure 8: Estimated effect of age, the illness score cesd and time based on BoostMixed

and b_i represents random slopes on the variables x_{it}. Quite a different challenge is the incorporation of random effects in additive terms. For simplicity of presentation we restrict consideration to one smooth effect. Let the smooth random intercept model

$$y_{it} = \beta_0 + \alpha(u_i) + b_{i0} + \varepsilon_{it},\ b_{i0} \sim N(0, \sigma^2),$$

be extended to

$$y_{it} = \beta_0 + \alpha(u_i) + \alpha(u_i) b_{i1} + b_{i0} + \varepsilon_{it}, \tag{11}$$

with $\big(b_{i0}, b_{i1}\big) \sim N(0, Q(\rho))$.
As usual the smooth component has to be centered for reasons of identifiability of effects, in our applications $\sum_i \alpha(u_i) = 0$ has been used. That means the "random slope" b_{i1} in model (11) is a parameter that, quite similar to random slopes in linear mixed models, lets the strength of the variable vary across subjects. The dependence on variable u_i becomes

$$\alpha(u_i) + \alpha(u_i) b_{i1} = \alpha(u_i)(1 + b_{i1})$$

showing that $\alpha(u_i)$ represents the basic effect of variable u_i but this effect can be stronger for individuals if $b_{i1} > 0$ and weaker if $b_{i1} < 0$. Thus b_{i1} strengthens or attenuates the effect of the variable u_i. If the variance of b_{i1} is very large it may even occur that $b_{i1} < 1$ meaning that the effect of u_i is "inverted" for some individuals. If $\alpha(u_i)$ is linear with $\alpha(u_i) = \beta u_i$, the influence term is given by $\alpha(u_i)(1 + b_{i1}) = u_i(\beta + \tilde{b}_{i1})$ where $\tilde{b}_{i1} = \beta b_{i1}$ represents the usual term in linear mixed models with random slopes. Thus comparison with the linear mixed model should be based on the rescaled random effect $\tilde{\beta}_{i1}$ with $E(\tilde{\beta}_{i1}) = 0, \mathrm{Var}(\tilde{\beta}_{i1}) = \beta^2 \mathrm{Var}(\beta_{i1})$.

The main problem in model (11) is the estimation of the random effect. If $\alpha(u)$ is expanded in basis functions by $\alpha(u) = \sum_s \alpha_s \Phi_s(u)$ one obtains

$$\alpha(u_i) b_i = \sum_s \alpha_s b_i \Phi_s(u)$$

which is a multiplicative model since α_s *and* b_i are unknown and cannot be observed. However, boosting methodology may be used to obtain estimates for the model. The basic concept in boosting is that in one step the refitting of $\alpha(u_i)$ is done by using a weak learner which in our case corresponds to large λ in the penalization term.

Thus in one step the change from iteration $\alpha^{(l)}$ to $\alpha^{(l+1)}$ is small. Consider model in vector form with predictor

$$\eta_i = \mathbf{1}\beta_0 + \Phi_i\alpha + (\mathbf{1}\,\Phi_i\alpha)\begin{pmatrix} b_i \\ b_{i1} \end{pmatrix}$$

where $1^T = (1, \ldots, 1)$ is a vector of 1s, Φ_i is the corresponding matrix containing evaluations of basis functions and $\alpha^T = (\alpha_1, \ldots \alpha_n)$ denotes the corresponding weights. Then the refitting of residuals in the iteration step is modified in the following way.

Let $\eta_i^{(l-1)}$ denote the estimate from the previous step. Then the refitting of residuals (without selection) is done by fitting the model

$$y_i - \eta_i^{(l-1)} \sim N(\eta_i,\, V_i(\theta))$$

with

$$\eta_i = \mathbf{1}\beta_0 + \Phi_i\alpha + (1,\, \Phi_i\hat{\alpha}^{(l-1)})\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \tag{12}$$

where β_0, α are the parameters to be estimated and the estimate from the previous step $\hat{\alpha}^{(l-1)}$ is considered as known parameter. With resulting estimates $\hat{\beta}_0$, $\hat{\alpha}$ the corresponding update step takes the form

$$\hat{\alpha}^{(l)} = \hat{\alpha}^{(l-1)} + \hat{\alpha} \quad , \quad \hat{\beta}_0^{(l)} = \hat{\beta}_0^{(l-1)} + \hat{\beta}_0.$$

The basic idea behind the refitting is that forward iterative fitting procedures like boosting are weak learners. Thus the previous estimate is considered as known in the last term of (12). Only the additive term $\Phi_i\alpha$ is refitted within one iteration step. Of course after the refit the variance components corresponding to (b_{i0}, b_{i1}) have to be estimated.

4.2 Application to Jimma Study

For the Jimma data from Section 1 we focus on the effect of age (in days) on the weight of children. Since growth measurements usually do not evolve linearly in time the use of a linear mixed model involves to find an appropriate scale of age. Lesaffre, Asefa & Verbeke (1999) found that weight is approximately linearly related with the square root of age. An even better approximation, they actually used in their analysis is the transformation $(age - log(age + 1) - 0.02 \times age)^{1/2}$. Since in growth curve analysis random slopes are needed , they had to find the scale before using mixed model methodology. The big advantage of the approach proposed here is that the scale of age has not to be found separately but is determined by the (flexible) mixed model itself. The model we consider includes random slopes on the age effects, smooth effect of age of mother and several parametric terms for the categorical variables. It has predictor

$$\eta_{it} = \beta_0 + \alpha_A(Age_i) + b_{i0} + b_{i1}\alpha_A(Age_i) + \alpha_{AM}(Age\,of\,Mother_i) + \;\; parametric\;term.$$

Figure 3 shows the overall dependence (of children). Figure 9 shows the (centered) dependence on age and age of mother. It is seen that the effect of age of mothers is hardly linear (as assumed in the linear mixed models). Body weight of children seems to increase with age of mother up to about 30 years, then the effect remains rather stable. Table 4 gives the estimates of the parametric terms. For comparison the estimates for the linear mixed model with random slopes on the transformed age and linear effect of age of mother are given in Table 4 . As transformed age we use $(age - log(age + 1) - 0.02 \times age)^{1/2}$ as suggested by Lesaffre, Asefa & Verbeke (1999). It is seen that the effects of the categoriacal covariates are quite comparable. The differing

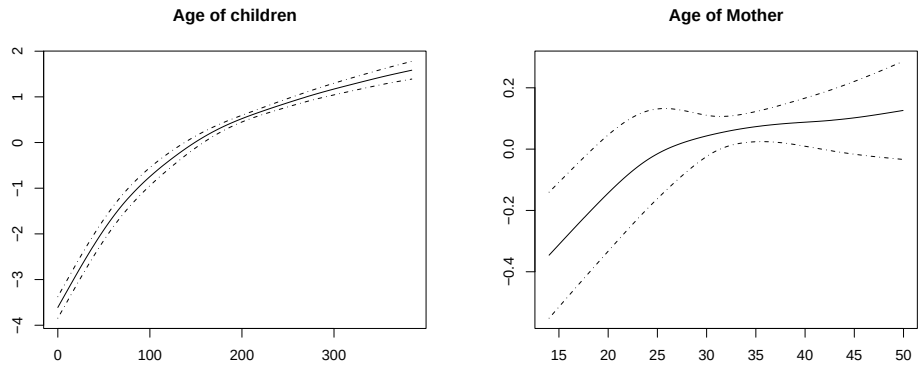


Figure 9: Effects of age of children (in days) and age of the mother (in years) in the Jimma study

intercepts are due to centering of variables. For age of mother the linear model shows a distinct increase (0.014 with standard deviation 0.004).

Table 5 shows the estimated variance of (b_{i0}, b_{i1}) for the flexible model and the linear mixed model with transformed age.

	BoostMixed		Mixed Model	
INTER	6.819	0.174	2.664	0.176
SEX	0.304	0.049	0.296	0.081
EDUC0	-0.051	0.066	-0.085	0.118
EDUC1	-0.021	0.151	-0.044	0.236
EDUC2	0.041	0.051	0.009	0.093
EDUC3	0.036	0.029	-0.005	0.060
EDUC4	-0.005	0.019	-0.042	0.042
VISIT	-0.078	0.072	-0.078	0.117
TIME	-0.177	0.065	-0.169	0.107
DELIV1	-0.027	0.007	-0.019	0.010
DELIV2	-0.148	0.031	-0.141	0.052
AGE			0.886	0.004
AGEM			0.014	0.004

Table 4: Effects of categorical covariates in Jimma study

BoostMixed		Mixed Model	
0.825962	0.196618	0.171369	-0.017506
0.196618	0.057253	-0.017506	0.045134

Table 5: Covariance matrix for random intercept and slope for Jimma data

5 Some Extensions

By allowing the variables $u_1, \ldots, u_m$ to have additive form model (3) represents a partially additive mixed model. More flexible predictors have been proposed in regression models, in particular varying coefficients models (Hastie & Tibshirani (1993)) and interactions between covariates. In the following the extension to more flexible forms of predictors in mixed models is considered briefly. For simplicity we consider only one additional term and variables that do not

vary across replications. The effect of variable z_i varies with variable u_i within a mixed model framework if the (additional) nonparametric term is given by

$$\mu_{it}^{nonp.} = \alpha(u_i) + z_i\gamma(u_i),$$

where $\alpha(u_i)$ and $\gamma(u_i)$ are unspecified functions of the continuous variable u_i Hastie & Tibshirani (1993) call u_i an effect modifier since the effect of z_i depends on the value of u_i. Often z_i is a factor represented by a 0-1 dummy variable. The (simplified) model has the form

$$\begin{aligned} y_{it} &= \mu_{it}^{nonp.} + \mu_{it}^{rand} + \varepsilon_{it} \\ &= \alpha(u_i) + z_i\gamma(u_i) + w_{it}^T b_i + \varepsilon_{it} \end{aligned}$$

which in general is enlarged by further linear and additive terms. For the vector representation one obtains

$$y_i = \Phi_i\alpha + \Phi_i(z_i)\gamma + W_i b_i + \varepsilon_i$$

where Φ_i represents the basis function for the additive term $\alpha(u_i)$ (see Section 2.1) and $\Phi_i(z)$ is a matrix composed from observations z and basis functions for $\gamma(u_i)$. Let $\gamma(u)$ be represented by

$$\gamma(u) = \sum_{s=1}^{n} \gamma_s \Phi_s^{(z)}(u),$$

then one obtains

$$z_i\gamma(u_i) = \sum_{s=1}^{n} \gamma_s z_i \Phi_s^{(z)}(u_i)$$

and the matrix $\Phi_i(z) = (\Phi_{rs})$ has elements $\Phi_{rs} = z_r\Phi_5^{(z)}(u_r)$. The corresponding vector γ is given by $\gamma^T = (\gamma_1, \ldots \gamma_M)$. Thus the model has the form (4) and may be estimated by boosting. After the additive terms have been fitted, the varying coefficients term $\Phi_i(z_i)\gamma$ is included by fitting in Step 2 of the algorithm the model for residuals

$$y_i - \eta_i^{(l-1)} \sim N(\Phi_i(z_i)\gamma,\, V_i(\theta))$$

yielding an estimate for γ. If one wants to consider more candidates for varying coefficients a selection step should be included.

6 Concluding Remarks

Alternative estimates have been proposed that yield stable estimates of additive mixed models also in the high dimensional case. If additive structures with a random intercept are not sufficient to capture the variation across subjects it is recommended to include an additional random slope which strengthens or attenuates the effect of a covariate. The model with random slopes is simply structured and adds only two additional parameters, the variance of the slope and the covariance between slope and intercept. It is therefore very parsimonious and allows simple interpretation. By using few additional parameters it has a distinct advantage over methods that allows subjects to have their own function, yielding as many functions as subjects (see Verbyla, Cullis, Kenward & Welham (1999)).

Acknowledgement

We gratefully acknowledge support from Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 386: Diskrete Strukturen). We thank Emannuel Lesaffre for letting us use the Jimma data.

A Appendix

A.1 ML for Variance Components

The estimation of the variance components is based on the profile log-likelihood that is obtained by plugging in the estimates $\hat{\delta}$ from a fixed boosting step in the penalized log-likelihood.

$$\begin{aligned} l(\hat{\delta};\theta) &= -\tfrac{1}{2}\textstyle\sum_{i=1}^n \log(|V_i(\theta)|) + \sum_{i=1}^n (y_i - \hat{\eta}^T V_i(\theta)^{-1}(y_i - \hat{\eta}) \\ &\quad -\tfrac{1}{2}\delta^T K \delta. \end{aligned}$$

Differentiation with respect to $\theta^T = (\sigma_\varepsilon, \varrho^T) = (\theta_1, \ldots, \theta_d)$ yields

$$s(\hat{\delta},\theta) = \tfrac{\partial l(\hat{\delta},\theta)}{\theta} = (s(\hat{\delta},\theta)_i)_{i=1,\ldots,d}$$
and
$$F(\hat{\delta},\theta) = -E(\tfrac{\partial^2 l(\hat{\delta},\theta)}{\partial\theta\partial\theta^T}) = (F(\hat{\delta},\theta)_{i,j})_{i,j=1,\ldots,d}$$

with

$$\begin{aligned} s(\hat{\delta},\theta)_i = \tfrac{\partial l(\hat{\delta},\theta)}{\theta_i} = \quad & -\tfrac{1}{2}\textstyle\sum_{k=1}^n \operatorname{trace}\left((V_k(\theta))^{-1}\,\tfrac{\partial V_k(\theta)}{\theta_i}\right) \\ & +\tfrac{1}{2}\textstyle\sum_{k=1}^n (y_k - \eta^{(l)})^T V_k(\theta)^{-1}\tfrac{\partial V_k(\theta)}{\theta_i} V_k(\theta)^{-1}(y_k - \eta^{(l)}) \end{aligned}$$

and

$$F(\hat{\delta},\theta)_{i,j} = \frac{1}{2}\sum_{k=1}^n \operatorname{trace}\left((V_k(\theta))^{-1}\,\frac{\partial V_k(\theta)}{\partial\theta_i}\,(V_k(\theta))^{-1}\,\frac{\partial V_k(\theta)}{\partial\theta_j}\right)$$

where

$$\frac{\partial V_k(\theta)}{\partial\theta_i} = \begin{cases} 2\sigma I_{T_k} & \text{if } i = 1 \\ W_k \frac{\partial Q(\varrho)}{\partial\theta_j} W_k^T & \text{if } j = i, i \neq 1. \end{cases}$$

It should be noted that maximization of $l(\hat{\delta},\theta)$ ignores the penalty term for δ.

For example, in the case of independence

$$Q(\varrho) = \varrho^2 * I$$

the elementwise derivative is

$$\frac{\partial Q(\varrho)}{\partial\varrho} = 2\varrho * I.$$

The estimator $\hat{\theta}$ can now be obtained by running a common Fisher scoring algorithm with

$$\hat{\theta}^{(s+1)} = \hat{\theta}^{(s)} + F(\hat{\delta},\theta^{(s)},)^{-1} s(\hat{\delta},\hat{\theta}^{(s)})$$

where s denotes the iteration index of the Fisher scoring algorithm. If Fisher scoring has converged the resulting $\hat{\theta}$ represents the estimates of variances for the considered boosting step.

A.2 Replacing the Truncated Power Series by B-Splines

In the following the use of B-Splines is sketched. For simplicity only one smooth component is considered with $\Phi_1(u), \ldots, \Phi_M(u)$ denoting the B-Splines for equidistant knots $k_1, \ldots, k_M$ and $\eta_i = Z_i\beta + \Phi_i\alpha$ denoting the predictor.

Let us first consider the difference matrix D^d corresponding to B-Spline penalization (see Eilers & Marx (1996)). With D being the $(M-1) \times M$ contrast matrix

$$D = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$$

one obtains higher order differences by the recursion $D^d = DD^{d-1}$ which is a $(M-d) \times M$ matrix. The penalty term is based on $\tilde{K} = (D^d)^T D^d$. New matrices $\tilde{X}_{(d)}$, depending on the order of the penalized differences are defined by

$$\tilde{X}_{(1)} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tilde{X}_{(2)} = \begin{pmatrix} 1 & k_1 \\ \vdots & \vdots \\ 1 & k_M \end{pmatrix}, \tilde{X}_{(3)} = \begin{pmatrix} 1 & k_1 & k_1^2 \\ \vdots & \vdots & \vdots \\ 1 & k_M & k_M^2 \end{pmatrix}.$$

For differences of order d one consider the $(M-d) \times M$ matrix $\tilde{W}_{(d)}^T = (D^d(D^d)^T)^{-1} D^d$. In the following we drop the notation of d and set $D := D^d$,$\tilde{W} := \tilde{W}_{(d)}$ and $\tilde{X} := \tilde{X}_{(d)}$. So $\tilde{W}$ and $\tilde{X}$ have the properties $D\tilde{X} = 0$,$\tilde{W}^T\tilde{X} = (DD^T)^{-1}D\tilde{X} = 0$,$\tilde{X}^T K\tilde{X} = 0 = \tilde{X}^T D^T D\tilde{X} = (D\tilde{X})^T(D\tilde{X})$. Most important is the equation

$$\tilde{W}^T K\tilde{W} = (DD^T)^{-1}DD^T DD^T(DD^T)^{-1} = I_{(M-d)}.$$

Since $\tilde{W}^T\tilde{X} = 0$, α can be decomposed into $\alpha = \tilde{X}\varphi + \tilde{W}\tilde{\alpha}$.

The predictor can now be rewritten in the form

$$\begin{aligned} \eta_i &= \left[Z_i, \Phi_i\right] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + W_i b_i = \left[Z_i, \Phi_i\right] \begin{bmatrix} \beta \\ \tilde{X}\varphi + \tilde{W}\tilde{\alpha} \end{bmatrix} + W_i b_i \\ &= \left[Z_i, \Phi_i\tilde{X}, \Phi_i\tilde{W}\right] \begin{bmatrix} \beta \\ \varphi \\ \tilde{\alpha} \end{bmatrix} + W_i b_i \\ &= \left[Z_i, \Phi_i\tilde{X}\right] \begin{bmatrix} \beta \\ \varphi \end{bmatrix} + \left[\Phi_i\tilde{W}, W_i\right] \begin{bmatrix} \tilde{\alpha} \\ b_i \end{bmatrix}. \end{aligned}$$

The penalized log-likelihood of the linear mixed model simplifies to

$$\begin{aligned} l_p(\delta) &= \textstyle\sum_{i=1}^n \log(f(y_i|\delta; b_i)p(b_i)) - \lambda\delta^T Diag(0_{(p\times p)}, \lambda K)\delta \\ &= \textstyle\sum_{i=1}^n \log(f(y_i|\delta; b_i)p(b_i)) - \lambda((\tilde{X}\varphi + \tilde{W}\tilde{\alpha})^T K(\tilde{X}\varphi + \tilde{W}\tilde{\alpha}) \\ &= \textstyle\sum_{i=1}^n \log(f(y_i|\delta; b_i)p(b_i)) - \frac{1}{2}\tilde{\alpha}^T 2 * \lambda I_{(M-d)}\tilde{\alpha}. \end{aligned}$$

This corresponds to the BLUP criterion of the mixed model

$$y_i = \tilde{Z}_i\tilde{\beta} + \begin{bmatrix} \Phi_i\tilde{W} & W \end{bmatrix} \begin{pmatrix} \tilde{\alpha} \\ b_i \end{pmatrix} + \epsilon_i$$

$$\text{with} \qquad \begin{pmatrix} \tilde{\alpha} \\ b_i \\ \epsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{2\lambda}I & 0 & 0 \\ 0 & Q(\rho) & 0 \\ 0 & 0 & \sigma_\epsilon^2 I \end{pmatrix} \right)$$

and $\tilde{\beta}^T = (\beta^T, \varphi^T)$, $\tilde{Z}_i = [Z_i, \Phi_i\tilde{X}]$. Thus, from decomposition $\alpha = \tilde{X}\varphi + \tilde{W}\tilde{\alpha}$ one obtains a mixed model with uncorrelated parameters $\tilde{\alpha}$.

References

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

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**, 961–976.

Bühlmann, P. and Yu, B. (2003). Boosting with l2 loss: Regression and classification. *Journal of the American Statistical Association* **98**, 324–339.

Dettling, M. and Bühlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19**, 1061–1069.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied Statistics* (to appear).

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

Green, D. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.

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

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society* **B 55**, 757–796.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226–252.

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

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

Lesaffre, E., Asefa, M., and Verbeke, G. (1999). Assessing the goodness-of-fit of the laird and ware model - an example: The jimma infant survival differential longitudinal study. *Statistics in Medicine* **18**, 835–854.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society* **B61**, 381–400.

Marx, D. B. and Eilers, P. (1998). Direct generalized additive modelling with penalized likelihood. *Comp. Stat. & Data Analysis* **28**, 193–209.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, linear and mixed models*. New York: Wiley.

Parise, H., Wand, M. P., Ruppert, D., and Ryan, L. (2001). Incorporation of historical controls using semiparametric mixed models. *Applied Statistics* **50**, 31–42.

Patterson, H. and Thompson, R. (1974). *Maximum Likelihood Estimation of Components of Variance*. Proceedings of the 8th International Biometric Conference.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science* **6**, 15–51.

Ruppert, D. and Carroll, R. J. (1999). Spatially-adaptive penalties for spline fitting. *Australian Journal of Statistics* **42**, 205–223.

Schimek, M. (2000). *Smoothing and Regression. Approaches, Computation and Application*. New York: Wiley.

Speed, T. (1991). That BLUP is a good thing: The estimation of random effects: Comment. *Statistical Science* **6**, 42–44.

Verbeke, G. and Molenberghs, G. (2001). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The anlysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics* **48**, 269–311.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics* **15**, 443–462.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additve models. *Journal of American Statistical Association* **99**, 673–686.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics* **50**, 689–699.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semi-parametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93**, 710–719.