Variable Selection for Generalized Additive Mixed Models by Likelihood-based Boosting

Andreas Groll[1] and Gerhard Tutz[2]

[1] Department of Statistics, University of Munich, Akademiestrasse 1, D-80799, Munich, Germany
(E-mail: andreas.groll@stat.uni-muenchen.de)
[2] Department of Statistics, University of Munich, Akademiestrasse 1, D-80799, Munich, Germany
(E-mail: tutz@stat.uni-muenchen.de)

Abstract. With the emergence of semi- and nonparametric regression the generalized linear mixed model has been expanded to account for additive predictors. In the present paper an approach for generalized additive mixed models is proposed which is based on likelihood-based boosting. In contrast to common procedures it can be used in high-dimensional settings where many covariates with unknown form of influence are available. It is constructed as a componentwise boosting method and hence is able to perform variable selection. The complexity of the resulting estimator is determined by information-based criteria.
Keywords: Generalized additive mixed model, Boosting, Smoothing, Variable selection, Penalized quasi-likelihood, Laplace approximation.

1 Introduction

General additive mixed models (GAMMs) are an extension of generalized additive models incorporating random effects. In the present article a boosting approach for the selection of additive predictors is proposed.

Boosting originates in the machine learning community and turned out to be a successful and practical strategy to improve classification procedures by combining estimates with reweighted observations. The idea of boosting has become especially important in the last decade as the issue of estimating high-dimensional models has become more urgent. Since Freund and Schapire (1996) have presented their famous AdaBoost many extensions have been developed (e.g. gradient boosting by Friedman et al., 2000, generalized linear and additive regression based on the L_2-loss by Bühlmann and Yu, 2003).

In the following the concept of likelihood-based boosting is extended to GAMMs which are sketched in Section 2. The fitting procedure is outlined in Section 3 and a simulation study is reported in Section 4.

2 Generalized Additive Mixed Models - GAMMs

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 $\mathbf{x}_{it}^T = (1, x_{it1}, \dots, x_{itp})$ be the covariate vector associated with fixed effects and $\mathbf{z}_{it}^T = (z_{it1}, \ldots, z_{itq})$ the covariate vector associated with random effects. Then the classical parametric random effects model assumes that the mean $\mu_{it} = E(y_{it}|\mathbf{b}_i, \mathbf{x}_{it}, \mathbf{z}_{it})$ is given by $\mu_{it} = \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{z}_{it}^T\mathbf{b}_i$, where $\mathbf{b}_i$ is a random effect. In addition it is assumed that the observations y_{it} are conditionally independent with variances $var(y_{it}|\mathbf{b}_i) = \phi\upsilon(\mu_{it})$, where $\upsilon(.)$ is a known variance function and ϕ is a scale parameter.

More generally, we include nonparametric effects and a general link. Let $\mathbf{u}_{it}^T = (u_{it1}, \ldots, u_{itm})$ denote the covariate vector associated with the nonparametric effects. The generalized semiparametric mixed model that is considered has the form

$$g(\mu_{it}) = \mathbf{x}_{it}^T\boldsymbol{\beta} + \sum_{j=1}^{m} \alpha_{(j)}(u_{itj}) + \mathbf{z}_{it}^T\mathbf{b}_i, \tag{1}$$

where g is a monotonic differentiable link function, $\mathbf{x}_{it}^T\boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$, including the intercept, $\sum_{j=1}^{m} \alpha_{(j)}(u_{itj})$ is an additive term with unspecified influence functions $\alpha_{(1)}, \ldots, \alpha_{(m)}$ and finally $\mathbf{z}_{it}^T\mathbf{b}_i$ contains the cluster-specific random effects $\mathbf{b}_i \sim N(0, \mathbf{Q})$, where $\mathbf{Q}$ is a $q \times q$ dimensional known or unknown covariance matrix. An alternative form that we also use in the following is $\mu_{it} = h(\eta_{it})$, where $h = g^{-1}$ is the inverse link function, called response function.

Versions of the additive model (1) have been considered e.g. by Lin and Zhang (1999), who use natural cubic smoothing splines for the estimation of the unknown functions $\alpha_{(j)}$. In the following regression splines are used, which have been widely used for the estimation of additive structures in recent years, see e.g. Marx and Eilers (1998) and Wood (2006).

In regression spline methodology the unknown functions $\alpha_{(j)}(.)$ are approximated by basis functions. A simple basis is known as the B-spline basis of degree d, yielding $\alpha_{(j)}(u) = \sum_{i=1}^{k} \alpha_i^{(j)}\ B_i^{(j)}(u; d)$, where $B_i^{(j)}(u; d)$ denotes the i-th basis function for variable j. Let $\boldsymbol{\alpha}_j^T = (\alpha_1^{(j)}, \ldots, \alpha_k^{(j)})$ denote the unknown parameter vector of the j-th smooth function and let $\mathbf{B}_j^T(u) = (B_1^{(j)}(u; d), \ldots, B_k^{(j)}(u; d))$ represent the vector-valued evaluations of the k basis functions. Then the parameterized model for (1) has the form

$$g(\mu_{it}) = \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{B}_1^T(u_{it1})\boldsymbol{\alpha}_1 + \cdots + \mathbf{B}_m^T(u_{itm})\boldsymbol{\alpha}_m + \mathbf{z}_{it}^T\mathbf{b}.$$

By collecting observations within one cluster we obtain the design matrix $\mathbf{X}_i^T = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i})$ for the i-th covariate, and analogously we set $\mathbf{Z}_i^T = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT_i})$, so that the model has the simpler form

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{B}_{i1}\boldsymbol{\alpha}_1 + \cdots + \mathbf{B}_{im}\boldsymbol{\alpha}_m + \mathbf{Z}_i\mathbf{b}_i,$$

where $\mathbf{B}_{ij}^T = [\mathbf{B}_j(u_{i1j}), \ldots, \mathbf{B}_j(u_{iT_ij})]$ denotes the transposed B-spline design matrix of the i-th cluster and variable j. Furthermore, let $\mathbf{X}^T =$

$[\mathbf{X}_1^T, \dots, \mathbf{X}_n^T]$, let $\mathbf{Z} = diag(\mathbf{Z}_1, \dots, \mathbf{Z}_n)$ be a block-diagonal matrix and let $\mathbf{b}^T = (\mathbf{b}_1^T, \dots, \mathbf{b}_n^T)$ be the vector collecting all random effects yielding the model in matrix form

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}_1\boldsymbol{\alpha}_1 + \ldots + \mathbf{B}_m\boldsymbol{\alpha}_m + \mathbf{Z}\mathbf{b} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b}, \qquad (2)$$

where $\boldsymbol{\alpha}^T = (\boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_m^T)$, $\mathbf{B} = (\mathbf{B}_1, \dots, \mathbf{B}_m)$.

The Penalized Likelihood Approach: Focusing on generalized mixed models we assume that the conditional density $f(y_{it}|\mathbf{x}_{it}, \mathbf{u}_{it}, \mathbf{b}_i)$ is of exponential family type. A popular method to maximize generalized mixed models is penalized quasi-likelihood (PQL), which has been suggested by Breslow and Clayton (1993), Lin and Breslow (1996) and Breslow and Lin (1995). In the following we shortly sketch the PQL approach for the semi-parametric model. In order to avoid too severe restrictions to the form of the functions $\alpha_{(j)}(\cdot)$, we use many basis functions in our approach, say about 20 for each function $\alpha_{(j)}(\cdot)$, and add a penalty term to the log-likelihood. With $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \mathbf{b}^T)$ one obtains the penalized log-likelihood

$$l^{\text{pen}}(\boldsymbol{\delta}, \mathbf{Q}) = \sum_{i=1}^{n} \log\left(\int f(\mathbf{y}_i|\boldsymbol{\delta}, \mathbf{Q}) p(\mathbf{b}_i, \mathbf{Q}) d\mathbf{b}_i\right) - \lambda \frac{1}{2} \sum_{j=1}^{m} \boldsymbol{\alpha}_j^T \mathbf{K}_j \boldsymbol{\alpha}_j, \qquad (3)$$

where $\mathbf{K}_j$ penalizes the parameters $\boldsymbol{\alpha}_j$ and smoothing parameter λ which controls the influence of the penalty term. When using P-splines one penalizes the difference between adjacent categories in the form $\boldsymbol{\alpha}_j^T \mathbf{K}_j \boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j^T (\boldsymbol{\Delta}^d)^T \boldsymbol{\Delta}^d \boldsymbol{\alpha}_j$, where $\boldsymbol{\Delta}^d$ denotes the difference operator matrix of degree d.

By approximating the likelihood in (3) along the lines of Breslow and Clayton (1993) one obtains the double penalized log-likelihood:

$$l^{\text{pen}}(\boldsymbol{\delta}, \mathbf{Q}) = \sum_{i=1}^{n} \log(f(\mathbf{y}_i|\boldsymbol{\delta}, \mathbf{Q})) - \frac{1}{2}\mathbf{b}^T \mathbf{Q}_b^{-1} \mathbf{b} - \lambda \frac{1}{2} \sum_{j=1}^{m} \boldsymbol{\alpha}_j^T \mathbf{K}_j \boldsymbol{\alpha}_j, \qquad (4)$$

with block-diagonal matrix $\mathbf{Q}_b = diag(\mathbf{Q}, \dots, \mathbf{Q})$. The first penalty term $\mathbf{b}^T \mathbf{Q}_b^{-1} \mathbf{b}$ is due to the approximation based on the Laplace method, the second penalty term $\lambda \sum_{j=1}^{m} \boldsymbol{\alpha}_j^T \mathbf{K}_j \boldsymbol{\alpha}_j$ determines the smoothness of the functions $\alpha_{(j)}(.)$ depending on the chosen smoothing parameter λ.

PQL usually works within the profile likelihood concept. For GAMMs it is implemented in the `gamm` function (R-package `mgcv`, Wood, 2006).

3 Boosted Generalized Additive Mixed Models

In Tutz and Groll (2010) and Tutz and Groll (2011a) boosting approaches for generalized linear mixed models were introduced. The boosting algorithm that is presented in the following extends these approaches to the framework of *additive mixed models*.

It is used that spline coefficients $\boldsymbol{\alpha}_j$ and B-spline design matrices $\mathbf{B}_j$ from equation (2) can be decomposed into an unpenalized and a penalized part (see Fahrmeir et al., 2004) with new corresponding diagonal penalty matrices $\mathbf{K} := \mathbf{K}_j = diag(0, \ldots, 0, 1, \ldots, 1)$. Furthermore, we drop the first column of each $\mathbf{B}_j$ as the parametric term of the model already contains the intercept and hence the smooth functions must be centered around zero in order to avoid identification problems.

The following algorithm uses componentwise boosting and is based on the EM-type algorithm that can be found in Fahrmeir and Tutz (2001). Only one component $\boldsymbol{\alpha}_j$ of the additive predictor is fitted at a time. That means that a model containing the linear term and only one smooth component is fitted within one iteration step.

The predictor containing all covariates associated with fixed effects and only the covariate vector of the r-th smooth effect yields

$$\boldsymbol{\eta}_{i \cdot r} = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{B}_{ir} \boldsymbol{\alpha}_r + \mathbf{Z}_i \mathbf{b}_i,$$

for cluster i. Altogether the predictor considering only the r-th smooth effect has the form

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

We further want to introduce the vector $\boldsymbol{\delta}_r^T := (\boldsymbol{\beta}^T, \boldsymbol{\alpha}_r^T, \mathbf{b}^T)$, containing only the spline coefficients of the r-th smooth component.

Algorithm `bGAMM`

---

1. *Initialization*
   Compute starting values $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\mathbf{Q}}^{(0)}$ and set $\hat{\boldsymbol{\eta}}^{(0)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(0)} + \mathbf{B}\hat{\boldsymbol{\alpha}}^{(0)} + \mathbf{Z}\hat{\mathbf{b}}^{(0)}$, e.g. by fitting the generalized linear mixed model $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$ with the R-function `glmmPQL` (Wood, 2006) from the `MASS` library (Venables and Ripley, 2002) and setting $\hat{\boldsymbol{\alpha}}^{(0)} = \mathbf{0}$.
2. *Iteration*
   For $l = 1, 2, \ldots$
   (a) *Refitting of residuals*
   (i.) *Computation of parameters*
   For $r \in \{1, \ldots, m\}$ the model

$$g(\boldsymbol{\mu}) = \hat{\boldsymbol{\eta}}^{(l-1)} + \mathbf{X}\boldsymbol{\beta} + \mathbf{B}_r \boldsymbol{\alpha}_r + \mathbf{Z}\mathbf{b}$$

   is fitted, where $\hat{\boldsymbol{\eta}}^{(l-1)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(l-1)} + \mathbf{B}\hat{\boldsymbol{\alpha}}^{(l-1)} + \mathbf{Z}\hat{\mathbf{b}}^{(l-1)}$ is considered a known off-set. Estimation refers to $\boldsymbol{\delta}_r$. In order to obtain an additive correction of the already fitted terms, we use one step in Fisher scoring with starting value $\boldsymbol{\delta}_r = \mathbf{0}$. Therefore Fisher scoring for the r-th component takes the simple form

$$\hat{\boldsymbol{\delta}}_r^{(l)} = (\mathbf{F}_r^{\text{pen}\,(l-1)})^{-1} \mathbf{s}_r^{(l-1)}$$

with penalized pseudo Fisher matrix $\mathbf{F}_r^{\text{pen }(l-1)} = \partial^2 l^{\text{pen}} / \partial\boldsymbol{\delta}_r \partial\boldsymbol{\delta}_r^T$ and using the unpenalized version of the penalized score function $\mathbf{s}_r^{\text{pen }(l-1)} = \partial l^{\text{pen}} / \partial\boldsymbol{\delta}_r$ with l^{pen} from (4). The variance-covariance components are replaced by their current estimates $\hat{\mathbf{Q}}^{(l-1)}$.

(ii.) *Selection step*
Select from $r \in \{1, \ldots, m\}$ the component j that leads to the smallest AIC or BIC and select $(\hat{\boldsymbol{\delta}}_j^{(l)})^T = \left((\hat{\boldsymbol{\beta}}^*)^T, (\hat{\boldsymbol{\alpha}}_j^*)^T, (\hat{\mathbf{b}}^*)^T\right)$.

(iii.) *Update*
Set

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

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

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

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

With $\mathbf{A} := [\mathbf{X}, \mathbf{B}, \mathbf{Z}]$ 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 EM- or REML-type estimates or alternative methods (see e.g. Tutz and Groll, 2010 and Tutz and Groll, 2011b for details).

4 Simulation study

In the following we present a simulation study to check the performance of the `bGAMM` algorithm and compare our algorithm to alternative approaches. The underlying model is the additive random intercept logit model with predictor

$$\eta_{it} = \sum_{j=1}^{p} f_j(u_{itj}) + b_i, \quad i = 1, \ldots, 40, \quad t = 1, \ldots, 10$$

which includes smooth effects given by $f_1(u) = 6\sin(u)$, $f_2(u) = 6\cos(u)$, $f_3(u) = u^2$, $f_4(u) = 0.4u^3$, $f_5(u) = -u^2$, $f_j(u) = 0$, for $j = 6, \ldots, 50$, with $u \in [-\pi, \pi]$ except for f_2, where $u \in [-\pi, 2\pi]$. We choose the different settings $p = 5, 10, 15, 20, 50$. For $j = 1, \ldots, 50$ the vectors $\mathbf{u}_{it}^T = (u_{it1}, \ldots, u_{it50})$ have been drawn independently with components following a uniform distribution within the specified interval. The number of observations is fixed as $n = 40, T_i := T = 10, \forall i = 1, \ldots, n$. The random effects are specified by $b_i \sim N(0, \sigma_b^2)$ with three different scenarios $\sigma_b = 0.4, 0.8, 1.6$. The identification of the optimal smoothing parameter λ has been carried out by BIC.

Performance of estimators is evaluated separately for the structural components and variance σ_b. By averaging across 100 data sets we consider

$$\text{mse}_f := \sum_{t=1}^{N} \sum_{j=1}^{p} (f_j(v_{tj}) - \hat{f}_j(v_{tj}))^2, \qquad \text{mse}_{\sigma_b} := ||\sigma_b - \hat{\sigma}_b||^2,$$

where $v_{tj}, t = 1, \ldots, N$ denote fine and evenly spaced grids on the different predictor spaces for $j = 1, \ldots, p$. Additional information on the stability of the algorithms was collected in *notconv* (n.c.), the number of datasets, where numerical problems occurred during estimation. Moreover, *falseneg* (f.n.) reflects the mean number of functions $f_j, j = 1, 2, 3, 4, 5$, that were not selected, *falsepos* (f.p.) the mean number of functions $f_j, j = 6, \ldots, p$ wrongly selected, respectively. As the `gamm` function is not able to perform variable selection it always estimates all functions $f_j, j = 1, \ldots, p$.

The results of all quantities for different scenarios of σ_b and for varying number of noise variables can be found in Table 1 and Figures 1 and 2. We compare our `bGAMM` algorithm with the **R** function `gamm` recommended in Wood (2006), which is providing a penalized quasi-likelihood approach for the generalized additive mixed model.

		`gamm`			`bGAMM` (EM)				`bGAMM` (REML)			
σ_b	p	mse_f	mse_{σ_b}	n.c.	mse_f	mse_{σ_b}	f.p.	f.n.	mse_f	mse_{σ_b}	f.p.	f.n.
0.4	5	54809.28	0.188	64	33563.44	1.382	0	0	41671.53	0.280	0	0.05
0.4	10	54826.50	0.112	85	33563.44	1.382	0	0	41671.53	0.280	0	0.05
0.4	15	51605.63	0.151	93	33563.44	1.382	0	0	41671.53	0.280	0	0.05
0.4	20	54706.54	0.149	96	33530.58	1.395	0	0	41624.79	0.282	0	0.05
0.4	50	-	-	100	33648.53	1.359	0	0	41606.17	0.282	0	0.05
0.8	5	52641.67	0.470	55	34581.50	1.584	0	0	42755.58	0.545	0	0.08
0.8	10	53384.37	0.462	88	34581.50	1.584	0	0	42755.58	0.545	0	0.08
0.8	15	53842.01	0.272	95	34581.50	1.584	0	0	42755.58	0.545	0	0.08
0.8	20	55771.45	0.320	96	34581.50	1.584	0	0	42755.58	0.545	0	0.08
0.8	50	-	-	100	34581.50	1.584	0	0	42755.58	0.545	0	0.08
1.6	5	53909.80	1.683	58	32844.44	1.689	0	0	40306.13	0.927	0	0.36
1.6	10	54376.56	2.160	86	32844.44	1.646	0	0	40306.13	0.927	0	0.36
1.6	15	53100.51	2.110	93	32844.44	1.410	0	0	40306.13	0.927	0	0.36
1.6	20	-	-	100	32844.44	1.891	0	0	40306.13	0.927	0	0.36
1.6	50	-	-	100	32884.22	1.897	0	0	40449.15	0.935	0	0.36

Table 1. Generalized additive mixed model with `gamm`* and boosting (`bGAMM`) on Bernoulli data (* only those cases, where `gamm` did converge)

5 Concluding Remarks

An algorithm for the fitting of GAMMs with high-dimensional predictor structure was proposed and examined. It works much more stable than existing approaches and allows a selection of influential variables from a set of variables including irrelevant ones. Also the form of the single functions can be estimated more adequately.

An alternative boosting scheme is available in the `mboost` package (see Hothorn et al., 2010 and Bühlmann and Hothorn, 2007). It provides a variety of gradient boosting families to specify loss functions and the corresponding risk functions to be optimized. The `gamboost` function also allows to model heterogeneity in repeated measurements, but, in contrast to the presented approach, fits a fixed parameter model.

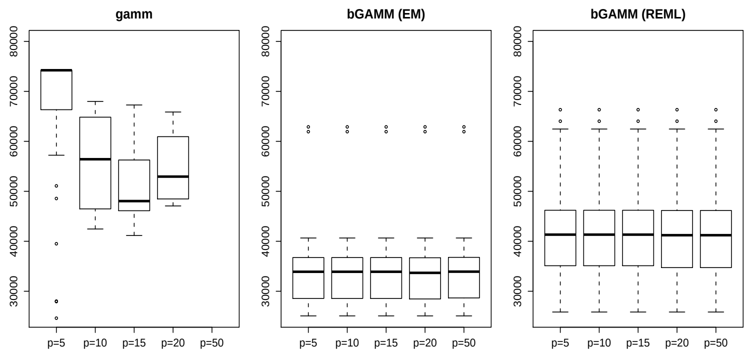

Fig. 1. Boxplots of mse_f for `gamm`* (left), `bGAMM` EM(middle) and `bGAMM` REML (right) for $p = 5, 10, 15, 20, 50$ (* only those cases, where `gamm` did converge)

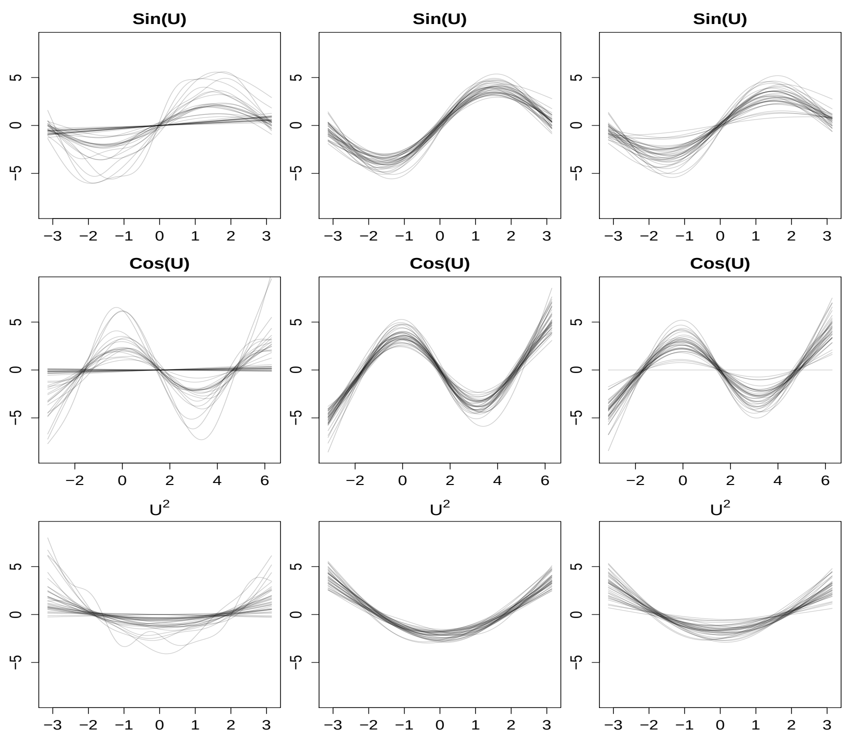

Fig. 2. First three smooth functions computed with the `gamm` model (left), the `bGAMM` EM model (middle) and the `bGAMM` REML model (right) for $p = 5, \sigma_b = 0.4$ for the cases where the `gamm` function did converge.

Bibliography

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association 88*, 9–25.

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

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22*, 477–522.

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

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica 14*, 731–761.

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

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

Friedman, J. H., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics 28*, 337–407.

Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner (2010). mboost: Model-based boosting. R package version 2.0-6 http://CRAN.R-project.org/package=mboost/.

Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*, 1007–1016.

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

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

Tutz, G. and A. Groll (2010). Generalized Linear Mixed Models Based on Boosting. In T. Kneib and G. Tutz (Eds.), *Statistical Modelling and Regression Structures - Festschrift in the Honour of Ludwig Fahrmeir*. Physica.

Tutz, G. and A. Groll (2011a). Binary and Ordinal Random Effects Models Including Variable Selection. Technical report, LMU Munich.

Tutz, G. and A. Groll (2011b). Binary and Ordinal Random Effects Models Including Variable Selection. *Comput. Stat. & Data Analysis*. Submitted.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

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