Generalized Additive Models with Flexible Response Functions*

Elmar Spiegel[1*], Thomas Kneib[1], Fabian Sobotka[2]
[1]University of Goettingen
[2]Carl von Ossietzky University Oldenburg

Abstract

Common generalized linear models (GLM) depend on several assumptions: (i) the specified linear predictor, (ii) the chosen response distribution that determines the likelihood and (iii) the response function that maps the linear predictor to the conditional expectation of the response. Generalized additive models (GAM) provide a convenient way to overcome the restriction to purely linear predictors. Therefore the covariates may be included as flexible nonlinear or spatial functions to avoid potential bias arising from misspecification. Single index models, on the other hand, utilize flexible specifications of the response function and therefore avoid the deteriorating impact of a misspecified response function. However, such single index models are usually restricted to a linear predictor and aim to compensate for potential nonlinear structures only via the estimated response function. We will show that this is insufficient in many cases and present a solution by combining a flexible approach for response function estimation using monotonic P-splines with additive predictors as in GAMs. Our approach is based on maximum likelihood estimation and also allows us to provide confidence intervals of the estimated effects. To compare our approach with existing ones, we conduct extensive simulation studies and apply our approach on two empirical examples, namely the mortality rate in São Paulo due to respiratory diseases based on the Poisson distribution and credit scoring of a German bank with binary responses.
Keywords: flexible response function, generalized additive model, monotonic P-spline, single index model

1 Introduction

In standard generalized linear models (GLM) (McCullagh and Nelder, 1989), we assume that the conditional distribution of the observed responses y_i given covariates $\boldsymbol{x}_i$ for observations $i = 1, \ldots, n$ belongs to the simple exponential family and that the conditional expectation of the response can be related to the linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ via

$$\mu_i = \mathbb{E}[Y_i|\boldsymbol{x}_i] = h(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

where $\boldsymbol{\beta}$ comprises the regression coefficients and $h(\cdot)$ is a monotonically increasing, pre-specified response function. As a consequence, the expected value $\mu_i = \mathbb{E}[Y_i|\boldsymbol{x}_i]$ depends not only on the predictor $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, but also on the choice of the response function h and misspecification may result in biased and misleading estimates. As a motivating example, consider binary responses where the GLM can be derived from a latent model specification for the unobserved, continuous response $y_i^* = \boldsymbol{x}_i^\top \boldsymbol{\beta} - \varepsilon_i$, where ε_i follows a given distribution with cumulative distribution function (CDF) F ($\varepsilon_i \sim F$). The latent response is then related to the observed, binary responses via the thresholding mechanism

$$y_i = \begin{cases} 1, & \text{if } y_i^* > 0 \\ 0, & \text{if } y_i^* \leq 0 \end{cases}$$

and one can show that $\mathbb{E}[Y_i|\boldsymbol{x}_i] = F(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ such that the CDF of the latent error term determines the response function (Cameron and Trivedi, 2005).

*This is an author's version of the accepted, peer-reviewed manuscript. The original is available at `https://doi.org/10.1007/s11222-017-9799-6`. For supplementary material please contact `espiege@uni-goettingen.de`.

To achieve additional flexibility, many previous extensions of GLMs dealt with replacing the linear predictor with an additive version that accommodates for example nonlinear effects of continuous covariates, spatial effects, random effects or complex types of interactions (see Wood, 2017; Fahrmeir et al., 2013, for overviews). However, as demonstrated for the specific case of binomial GLMs, the quality of estimates and predictions depends not only on the predictor specification but also on the chosen response function. This problem is well known in the literature (see for example Czado and Santner, 1992). The most well known model class to deal with it are *single index models* as introduced by Ichimura (1993), where kernel density estimates are used to determine the response function. A similar design was used by Klein and Spady (1993) and Weisberg and Welsh (1994) were among the first to name this the missing link problem. Others like Carroll et al. (1997) and Wang et al. (2010) developed single index models further and applied them to the partial linear single index framework. Alternatively, Koenker and Yoon (2009) proposed to use more flexible, but still parametric response functions like the Pregibon response function. Friedman and Stuetzle (1981) also describe a method to estimate a nonlinear relationship between the response and the predictor. The kernel methods based on Ichimura (1993) all have the disadvantage to regularly estimate too flexible response functions which are often quite wiggly and do not ensure monotonicity of the response function. To stabilize the estimation, approaches based on penalized splines have been introduced by Yu and Ruppert (2002), Muggeo and Ferrara (2008) and Yu et al. (2017). They penalize the flexibility of the response function estimate such that the result is a smooth curve. The penalized estimation procedure of the response function has also been transferred to the boosting framework (Bühlmann and Hothorn, 2007) by Leitenstorfer and Tutz (2011) and Tutz and Petry (2012). However, these methods are only defined for linear predictors, while nonlinear effects may still occur even after considering a flexible specification for the response function. In our simulation study (Section 4) we show that classical single index models with linear predictors are not able to capture nonlinear covariate effects.
Novel approaches in this direction comprise Marx (2015) and Tutz and Petry (2016). While Marx (2015) defines his algorithm only for continuous responses and puts more emphasis on defining varying coefficient surfaces, Tutz and Petry (2016) build an additive version of Tutz and Petry (2012) with a particular focus on the variable selection enabled in the boosting framework. In contrast, we will combine the framework of single index models based on P-splines as described in Muggeo and Ferrara (2008) with the generalized additive model (GAM) (Hastie and Tibshirani, 1986) based on maximum likelihood (ML) methods. This has several advantages:

- We can readily combine the flexibility of GAMs in terms of predictor specifications with the data-driven determination of the shape of the response function.
- Ensuring monotonicity of the response function is easily integrated in the estimation framework, which also ensures that we obtain interpretable covariate effects.
- Embedding estimation in the ML framework allows us to also derive corresponding inferential statements, for example concerning the standard errors of estimated effects.
- The approach can be implemented based on existing software and in particular the R-package `mgcv`.

The rest of the paper is structured as follows: First, we give a short overview of GAMs in Section 2. In Section 3, we summarize the approach of Muggeo and Ferrara (2008) and introduce our new method including semiparametric predictors. Afterwards we report on our simulation study in Section 4 to compare our new method with previous suggestions. Furthermore, we introduce data on mortality rates in São Paulo due to respiratory diseases and credit scoring of a German bank, as applications in Section 5. The paper concludes with a discussion in Section 6.

2 Additive Models

2.1 Generalized Additive Models

Standard GLM depend on the assumption that the data follow a distribution which is member of the exponential family. Thus the corresponding density may be written as

$$f(y_i, \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right),$$

where θ_i are the unknown parameters and a, b, c are fixed functions depending on the specific distribution. Furthermore the moments of the distribution are given as

$$\mathbb{E}(Y_i|x_i) = \mu_i = b'(\theta_i) \qquad \& \qquad \text{Var}(Y_i|\theta_i) = a(\phi)b''(\theta_i).$$

In a standard GLM, it is also assumed that the expected values may be modeled as

$$\mu_i = h(\eta_i),$$

where $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ is the linear predictor. However, restricting the predictor to be a linear combination of the covariates is often not sufficient. Therefore semiparametric predictors have been introduced (see for example Hastie and Tibshirani, 1986; Wood, 2017; Fahrmeir et al., 2013) that combine linear effects of some covariates x_{i1}, x_{i2}, ... with smooth, nonlinear effects of continuous covariates x_{ir}, x_{ir+1}, ... leading to a predictor of the form

$$\begin{aligned}\eta_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + \\ s_r(x_{ir}) + s_{r+1}(x_{ir+1}) + \ldots\end{aligned}$$

One convenient way to specify the nonlinear effects s_j is based on B-splines where the effects $s_j(x_{ij})$ are approximated by sums of several B-spline basis functions $B_{l_j}^{(d)}(x_{ij})$ (of a pre-specified degree d) evaluated at x_{ij}, scaled by the basis coefficients γ_{l_j} such that

$$s_j(x_{ij}) = \sum_{l_j=1}^{L_j} B_{l_j}^{(d)}(x_{ij})\gamma_{l_j}.$$

The derivative of a B-spline can be routinely calculated based on the local polynomial structure of B-splines. To determine the derivative, we start with defining a set of new basis functions

$$\dot{B}_{l_j}(x_{ij}) = (d-1)\left(\frac{B_{l_j}^{(d-1)}(x_{ij})}{\kappa_{l_j+d-1} - \kappa_{l_j}} - \frac{B_{l_j+1}^{(d-1)}(x_{ij})}{\kappa_{l_j+d} - \kappa_{l_j+1}}\right),$$

where $B_{l_j}^{(d-1)}(x_{ij})$ are B-spline basis functions determined based on the same set of knots κ_{l_j} as the original basis functions $B_{l_j}^{(d)}(x_{ij})$ but with the degree decreased by one. Combining $\dot{B}_{l_j}(x_{ij})$ with the original coefficients results in the derivative (for details see de Boor, 1978)

$$s_j'(x_{ij}) = \frac{\mathrm{d}s_j(x_{ij})}{\mathrm{d}x} = \sum_{l_j=1}^{L_j} \dot{B}_{l_j}(x_{ij})\gamma_{l_j}. \tag{1}$$

In the following, we suppress the degree d of the basis functions to simplify notation. While pure B-spline fits depend crucially on the number L_j and positioning of the basis functions, P-splines as introduced by Eilers and Marx (1996) avoid this dependency by considering a large number of basis functions subject to the smoothness condition that neighboring basis coefficients should not differ too much. This is achieved by adding the penalty terms $\lambda_j \sum_{l_j=3}^{L_j}(\gamma_{l_j} - 2\gamma_{l_j-1} + \gamma_{l_j-2})^2$ for each smooth effect to the fit criterion where $\lambda_j \geq 0$ is the smoothing parameter determining the impact of the penalty on the estimation result. The penalty may be written in matrix notation as $\lambda_j \boldsymbol{\gamma}_j^\top \boldsymbol{K}_j \boldsymbol{\gamma}_j$, with $\boldsymbol{\gamma}_j = (\gamma_{j_1}, \ldots, \gamma_{j_{L_j}})^\top$ the vector coefficients of one smooth effect. $\boldsymbol{K}_j$ is the penalty matrix to determine the second order differences.
To simplify the notation, we generically define the semiparametric predictor as $\eta_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma}$ where

- $\boldsymbol{Z}$ is the design matrix built jointly from the linear effects and the evaluated basis functions. The ith row of $\boldsymbol{Z}$ is defined as

$$\boldsymbol{z}_i^\top = (1, x_{i1}, \ldots, B_{l_r}(x_{ir}), B_{l_r+1}(x_{ir}), \ldots)$$

- $\boldsymbol{\gamma}$ is the complete vector of coefficients

$$\boldsymbol{\gamma}^\top = (\beta_0, \beta_1, \ldots, \gamma_{l_r}, \gamma_{l_r+1}, \ldots)$$

- $\boldsymbol{K}$ is a blockdiagonal matrix summarizing the penalties $\lambda_j \boldsymbol{K}_j$ of the individual effects on the diagonal. Unpenalized coefficients have diagonal elements with value 0.

In summary, including a semiparametric predictor in GLMs results in GAMs. The likelihood of GAMs is similar to the one of GLMs where only the penalty needs to be augmented when fitting the model. Hence, the penalized log-likelihood is defined as

$$l(\boldsymbol{\gamma}, \lambda) = \log \left(\prod_{i=1}^{n} f(y_i, \boldsymbol{\gamma}) \right) - \frac{1}{2} \boldsymbol{\gamma}^\top \boldsymbol{K} \boldsymbol{\gamma}.$$

The estimation scheme for GAMs stays the same Fisher scoring algorithm as in the standard GLM. Thus the coefficients are estimated using iteratively weighted least squares with working responses

$$y_i^{(k)} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)} + \frac{y_i - h(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)})}{h'(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)})}$$

and working weights

$$w_i^{(k)} = \frac{\left(h'(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)}) \right)^2}{\mathrm{Var}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)})}$$

for the kth iteration. The coefficients are now estimated including the penalty via penalized iteratively weighted leasted squares, i.e.

$$\boldsymbol{\gamma}^{(k)} = \left(\boldsymbol{Z}^\top \boldsymbol{W}^{(k)} \boldsymbol{Z} + \boldsymbol{K} \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{W}^{(k)} \boldsymbol{y}^{(k)},$$

with $\boldsymbol{W}^{(k)}$ representing the diagonal matrix of the working weights. To optimize the smoothing parameters λ_j, several approaches such as generalized cross-validation (GCV) can be considered (see Wood, 2017, for details). In Wood (2017) also details on the derivation of the standard GLM Fisher scoring and for the derivation of the GAM are displayed. Besides P-splines, several other types of smooth functions may be included in the same way, e.g. tensor product splines or Gaussian Markov random fields (see Fahrmeir et al., 2013, for details).

2.2 Monotonic P-splines

For the estimation of the response function, we are interested in monotonically increasing P-splines. For their estimation several approaches have been introduced. Bollaerts et al. (2006) use an extra penalty that adds a heavy penalty for negative differences of neighboring coefficients. However, this approach lacks identifiability if more than one smooth covariate is used. An alternative approach overcoming this restriction was introduced by Pya and Wood (2015), leading to shape constrained P-splines (SCOP-splines). In the case that there is only one covariate x, these SCOP-splines are defined like standard B-splines as

$$\Psi(x_i) = \sum_{l=1}^{L} B_l(x_i) \xi_l = \boldsymbol{B}_i^\top \boldsymbol{\xi},$$

where $\boldsymbol{B}_i^\top = (B_1(x_i), \ldots, B_L(x_i))$ is the vector of evaluated basis functions at observation i and $\boldsymbol{B}$ is the corresponding matrix for all observations. To fulfill the condition of a monotonic increase, i.e. $\Psi'(x) > 0 \Leftrightarrow \xi_l \leq \xi_{l+1} \forall l$, they reparameterize the coefficients $\boldsymbol{\xi}$ such that

$$\boldsymbol{\xi} = \boldsymbol{U} \tilde{\boldsymbol{\nu}}$$

where

$$\boldsymbol{\nu} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_L \end{pmatrix}, \tilde{\boldsymbol{\nu}} = \begin{pmatrix} \nu_1 \\ \exp(\nu_2) \\ \vdots \\ \exp(\nu_L) \end{pmatrix}, \boldsymbol{U} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ & & & \ddots & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$

The monotone spline is then defined as

$$\Psi(x_i) = \boldsymbol{B}_i^\top \boldsymbol{U}\tilde{\boldsymbol{\nu}}.$$

So we optimize the penalized log-likelihood

$$l(\boldsymbol{\nu}, \lambda_\nu) = l(\boldsymbol{\nu}) - \frac{1}{2}\boldsymbol{\nu}^\top \boldsymbol{K}_\nu \boldsymbol{\nu},$$

where $\boldsymbol{K}_\nu$ is the corresponding penalty matrix including the smoothing parameter λ_ν. In order to estimate the coefficients, we apply the reparameterization via $\boldsymbol{U}\tilde{\boldsymbol{\nu}}$ inside of $l(\boldsymbol{\nu})$. Details of the estimation procedure are described in Pya and Wood (2015). This method is implemented in the `scam` package (Pya, 2017).
Similarly as in Equation (1) we get the derivative of the SCOP-spline as

$$\Psi'(x_i) = \frac{\mathrm{d}\Psi(x_i)}{\mathrm{d}x} = \dot{\boldsymbol{B}}_i^\top \boldsymbol{U}\tilde{\boldsymbol{\nu}}, \tag{2}$$

where $\dot{\boldsymbol{B}}_i^\top$ is the vector of evaluated basis functions $\dot{B}_l(x_i)$ as before.

The third alternative is using constrained least squares. For monotonic splines, we again set up the P-spline as

$$\hat{h}(x_i) = \sum_{l=1}^{L} B_l(x_i)\nu_l = \boldsymbol{B}_i^\top \boldsymbol{\nu}.$$

The aim is to optimize the penalized least squares criterion (PLS)

$$(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\nu})^\top (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\nu}) + \boldsymbol{\nu}^\top \boldsymbol{K}_\nu \boldsymbol{\nu},$$

with second order difference penalty $\boldsymbol{K}_\nu$ (including λ_ν) subject to the constraint that the coefficients are increasing, i.e. $\nu_l \leq \nu_{l+1}$. This can be done using inequality constraints via quadratic programming. An approach to solve this was introduced by Wood (1994) and is implemented in the `pcls` function of the `mgcv` package.

3 Generalized Additive Models with Flexible Response Functions

3.1 Indirect Estimation of the Response Function (FlexGAM1)

Our first approach for combining flexible estimates for the response function with additive predictors is inspired by an earlier proposal by Muggeo and Ferrara (2008) for flexible response functions in GLMs (which itself is an extension of the paper of Yu and Ruppert, 2002). Muggeo and Ferrara (2008) combine the standard response function with a smooth transformation of the linear predictor, leading to

$$g(\mu_i) = \Psi(\eta_i) \quad \Leftrightarrow \quad \mu_i = g^{-1}(\Psi(\eta_i)) = h(\Psi(\eta_i))$$

where Ψ is estimated as a monotone P-spline, i.e. in our setting a SCOP-spline, and h is the canonical response function.
Replacing the linear predictor $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ with a semiparametric predictor $\eta_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma}$ leads to our first proposed approach called *FlexGAM1* in the following. The penalized log-likelihood now depends on the coefficients of the covariate effects $\boldsymbol{\gamma}$ and the coefficients of the estimated response function Ψ: $l(\boldsymbol{\gamma}, \boldsymbol{\nu})$

For the maximization of this likelihood, we use a similar type of algorithm as proposed by Muggeo and Ferrara (2008) for the linear case. In particular, we rely on an iterative procedure with fixing either $\boldsymbol{\gamma}^{(k)}$ to estimate $\Psi^{(m)}$ (outer loop (m)) or fixing $\Psi^{(m)}$ to estimate $\boldsymbol{\gamma}^{(k)}$ (inner loop (k)). Thereby the outer loop is a GAM with one single smooth function $\Psi^{(m)}$ of the continuous covariate $\boldsymbol{\eta}^{(k)}$. Thus $\Psi^{(m)}$ is estimated with the standard tools of SCOP-splines. The single covariate is the semiparametric predictor $\boldsymbol{\eta}^{(k)}$ that changes during the iterations. The inner loop is then optimizing the profile likelihood with fixed $\Psi^{(m)}$. This is done with the usual iteratively weighted least squares algorithm. Based on similar ideas as in the standard GLM (see Wood, 2017, for example), the working weights and working responses can be derived. However, the chain rule has to be applied to consider the response function as $h(\Psi(\eta_i)^{(m)})$. Finally, we get the following working responses $y_i^{(k)}$ and working weights $w_i^{(k)}$

$$y_i^{(k+1)} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k)} + \frac{y_i - h\left(\Psi^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k)})\right)}{h'\left(\Psi^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k)})\right) \Psi'^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k)})}$$

$$w_i^{(k+1)} = \frac{\left(h'\left(\Psi^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k)})\right) \Psi'^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k)})\right)^2}{\text{Var}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k)})}$$

where Ψ' is the derivative of the SCOP-spline as defined in Equation (2).
Algorithm 1 in the Appendix gives a detailed description of the resulting fitting scheme. We also provide details on the derivation of the IWLS updates for binomial, Poisson, Gaussian and gamma distributed data in the supplementary material.
An important issue in models with flexible response functions is the inclusion of appropriate identifiability constraints. In our case, these constraints are as follows:

1. We require at least two continuous covariates in the model specification. Otherwise, the flexible response function and the flexible covariate effect can not be separated from each other.

2. The intercept has to be removed, i.e. $\gamma_0 = 0$. Alternatively, a shift on the x-axis of the response function is compensated by the intercept.

3. All smooth effects have to be centered around zero, i.e. $\sum_{i=1}^{n} s_j(x_{ij}) = 0$. This is done to prevent that one covariate spline is shifted on the y-axis and that shift is compensated by a shift of another covariate spline.

4. The predictor has to be scaled, i.e. $\sum_{i=1}^{n} \eta_i = 0$ and $\frac{1}{n}\sum_{i=1}^{n} \eta_i^2 = 1$. Otherwise, the predictor can be shifted and stretched arbitrarily and the effect is compensated by a shift or stretching in the response function.

5. The response function has to be monotonically increasing, i.e. $\Psi'(\eta) > 0$.

Condition (3) is incorporated in the setup of the smooth effects by including a QR-decomposition in their basis functions (for details see Wood, 2017). We consider condition (4) via scaling of the inner predictor

$$\eta_i = \frac{\eta_i - \text{mean}(\boldsymbol{\eta})}{\text{sd}(\boldsymbol{\eta})}$$

in each step of the algorithm. The monotonicity condition (5) is considered as an identifiability restriction, otherwise the estimated function $h(\Psi(\eta))$ will not be a valid response function in a strict sense (compare McCullagh and Nelder, 1989, p. 27). This restriction reduces the flexibility of the approach, but it allows us to interpret the covariate effects in a traditional way. Thus an increase of the predictor induces an increase of the expected value. Moreover, the reduced flexibility also stabilizes the estimation and prevents from weird effects at the edge of the parameter space. Furthermore, a small simulation study also showed, that using monotone splines results in smoother estimates of the response function.

In this paper, we follow the approach of Muggeo and Ferrara (2008) when scaling the predictor to achieve identifiability, but alternatives are possible. Tutz and Petry (2016), for example, apply constraints on the variance of each effect, which could be included in our method by scaling each effect s_j with $\frac{s_j}{\sum_i \sum_j s_j^2(x_{ij})}$. In our simulation study, both scalings achieved similar results. Following Li and Racine (2007, p. 251f.), coefficients are scaled to have $||\boldsymbol{\gamma}|| = 1$. However, this ignores the ties in the coefficients of the smooth effects. Thus we decided to follow the approach of Muggeo and Ferrara (2008), since it allows for the inclusion of other smooth effects like Gaussian Markov random fields more easily.

3.2 Direct Estimation of the Response Function (FlexGAM2)

In the paper of Muggeo and Ferrara (2008) and our FlexGAM1 approach, the combination of the traditional response function h and a transformation function Ψ is used to estimate the fitted values

$$\mu_i = \mathbb{E}[Y_i|\boldsymbol{x}_i] = h(\Psi(\eta_i)).$$

This is done to get a flexible response function while simultaneously ensuring that the response function maps the predictor to the right parameter space, e.g. $0 \leq \hat{\mu}_i \leq 1$ for binary data. By applying an outer response function, we however implicitly keep a distributional assumption. Therefore we aim at removing this implicit assumption and fit the response function completely flexible (*FlexGAM2*) such that

$$\mu_i = \mathbb{E}[Y_i|\boldsymbol{x}_i] = \hat{h}(\eta_i).$$

However, $\hat{h}$ should still be a valid response function and we incorporate corresponding restrictions via constrained least squares. Furthermore, we want to be able to deal with nonlinear effects such that we combine our procedure with semiparametric predictors.
Similar to the standard GAM, we can derive the estimation procedure. We just exchange the response function to be estimated directly using a strictly monotonically increasing P-spline which also considers restrictions on the fitted values

$$\hat{h}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}) = \sum_{l=1}^{L} B_l(\boldsymbol{z}_i^\top \boldsymbol{\gamma})\nu_l.$$

For the estimation, we use the constraint least squared approach as introduced by Wood (1994). This leads to a the similar penalized log-likelihood as in the standard GAM, but besides the coefficients of the covariate effects $\boldsymbol{\gamma}$ the coefficients $\boldsymbol{\nu}$ of the response function also have to be estimated, which results in the penalized log-likelihood: $l(\boldsymbol{\gamma}, \boldsymbol{\nu})$.
We optimize this log-likelihood using a two stage procedure as in FlexGAM1. Here, in the outer loop, we estimate the monotonically increasing P-spline $\hat{h}^{(m)}$, while in the inner loop we apply a standard IWLS algorithm as in the usual GAM but with the estimated response functions instead of the pre-specified ones. This results in the following working elements:

$$y_i^{(k)} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)} + \frac{y_i - \hat{h}^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)})}{\hat{h}'^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)})}$$

$$w_i^{(k)} = \frac{\left(\hat{h}'^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)})\right)^2}{\text{Var}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(k-1)})}$$

The complete algorithm is displayed in the Appendix in Algorithm 2, while the derivation for binomial, Poisson, Gaussian and gamma distributed data is displayed in the electronic appendix. The derivatives $\hat{h}'$ are calculated as introduced in Equation (1).
Moreover we require the same constraints as for FlexGAM1 to ensure identifiably. In addition, we need some further constraints to get a valid response function $\hat{h}(\cdot) = \sum_{l=1}^{L} B_l(\cdot)\nu_l$:

- The response function should be strictly monotonically increasing ($\frac{\partial h}{\partial \eta} > 0$), i.e.

$$\nu_l < \nu_{l+1} \Leftrightarrow \nu_{l+1} - \nu_l \geq \delta_{min}$$

  where δ_{min} is a small positive number.

- For binomial, Poisson or gamma distributed data the response function should be positive ($\hat{h}(\cdot) \geq 0$): Since an unscaled B-spline basis function is always non-negative ($B_l(\cdot) \geq 0$), this is achieved by

$$\nu_l \geq 0 \quad \forall l.$$

- The maximal value of the response function should be one ($\hat{h}(\cdot) \leq 1$) for binomial distributed data: Since an unscaled B-spline sums up to one ($\sum_{l=1}^{L} B_l(\cdot) = 1$), this is achieved by

$$\nu_l \leq 1 \quad \forall l.$$

In summary, we have several linear inequality constraints on the coefficients $\boldsymbol{\nu}$, which we can translate into $\boldsymbol{A\nu} \geq \boldsymbol{b}$, where $\boldsymbol{A}$ and $\boldsymbol{b}$ are the matrix and the vector of the linear inequality constraints respectively. Therefore we can apply the `pcls` function of the `mgcv` package, which deals with least squares problems under inequality constraints and quadratic penalties (Wood, 1994), to estimate $\hat{h}^{(m)}$ in the outer loop.

3.3 Numerical Details of FlexGAM1 and FlexGAM2

So far we have defined both, FlexGAM1 and FlexGAM2, for fixed smoothing parameters λ_j and λ_ν. In practice, these smoothing parameters have to be optimized. Since both stages of the algorithms depend on each other and therefore on the smoothing parameters of both stages, an optimization within these stages cannot achieve the best results. We therefore propose to optimize all smoothing parameters jointly from outside the algorithm, i.e. to define one set of fixed smoothing parameters, estimate the model, evaluate its prediction error and then check the next set of smoothing parameters. The possible smoothing parameter sets are evaluated by standard optimization procedures. We compute the prediction error via an ordinary 5-fold cross-validation with true separation in training and validation data sets. The error is thereby estimated as the predictive deviance. In our setting, a GCV criterion would not be applicable straightforward since the definition of the effective degrees of freedom in this interdependent two stage procedure is non-trivial.
In the estimation procedure, it regularly occurs that the algorithm does not converge, since a small difference in one of the estimates induces another small change in the other estimates. This micro-oscillation also occurs in standard GLM estimation via the Fisher-Scoring/ IWLS algorithm, mostly when not using the conjugate link function (see for example Marschner et al., 2011). Thereby a nondecreasing deviance can be detected. Generally, in these cases step halving is applied (see for example Jørgensen, 1984, Marschner et al., 2011 or Yu et al., 2017). So we adapt the approach by only accepting the new $\boldsymbol{\gamma}^{(k)}$ in the inner loop if they reduce the penalized deviance. If the deviance is nondecreasing, a new proposal of $\boldsymbol{\gamma}^{(k)}$ is calculated as

$$\boldsymbol{\gamma}^{(k)} = \frac{1}{2}\left(\boldsymbol{\gamma}^{(k)} + \boldsymbol{\gamma}^{(k-1)}\right).$$

However, step halving is only applicable in the inner loop, which is a modification of the standard IWLS. Still, micro-oscillation also occurs in the outer loop. Therefore we include an extra stopping criterion for the outer loop, if the penalized deviance does not decrease. To force the algorithm to always start iterating, the outer stop criterion is not used in the first outer loop, while the step halving of the inner loop is always possible except in the very first step. Overall step halving solved the convergence problem in our algorithms.

An additional possibility for adjustments is the choice of the initial parameters. Generally we propose to use a standard GAM with canonical response function. However, this model could be either estimated including the intercept, or excluding the intercept. In the simulation study, the models with intercept in the initial model provided better results, while in the empirical examples the estimates without intercept in the initial model performed better. Checking both possibilities reduces the risk of being stuck in a local optimum.

3.4 Uncertainty Quantification

In addition to providing point estimates, determining measures of uncertainty for the estimated coefficients is also of high relevance. Since we apply cubic P-splines to model the response function, we can differentiate the likelihood two times continuously. Therefore the Fisher regularity conditions (see Held and Sabanés Bové, 2014, p. 80) are fulfilled and we can make use of the standard asymptotics of ML-estimates (compare Fahrmeir et al., 2013, p. 662)

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N(\boldsymbol{\theta}, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\theta}})),$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\nu}})^\top$ are the ML-estimates based on the algorithms given above and $\boldsymbol{F}(\hat{\boldsymbol{\theta}})$ is the expected Fisher information derived for all coefficients jointly (for the formulas of $\boldsymbol{F}(\hat{\boldsymbol{\theta}})$ see supplementary material Section A). To get valid asymptotics, we need enough data, such that the coefficients are asymptotically unbiased. Based on the distribution of the coefficients, we can build the standard confidence intervals for the coefficients of the linear effects. For the display of the variation of smooth effects we make use of the approach of Marra and Wood (2012), i.e. we build the confidence intervals for a smooth function $s_j(x_{ij})$ as

$$\hat{s}_j(x_{ij}) \pm z_{1-\alpha/2}\sqrt{[\boldsymbol{V}_{s_j}]_{ii}}$$

where $[\boldsymbol{V}_{s_j}]_{ii}$ is the ith diagonal element of the covariance matrix $\boldsymbol{V}_{s_j} = \boldsymbol{Z}_j \boldsymbol{F}_j^{-1} \boldsymbol{Z}_j^\top$. Here $\boldsymbol{Z}_j$ is the model matrix for the jth component and $\boldsymbol{F}_j^{-1}$ are the corresponding elements of the inverse of the expected Fisher information matrix. Additionally, the confidence intervals for the estimated response function are given similarly as

$$h\left(\Psi(\eta_i) \pm z_{1-\alpha/2}\sqrt{[\boldsymbol{V}_\eta]_{ii}}\right) \quad \text{respectively}$$
$$\hat{h}(\eta_i) \pm z_{1-\alpha/2}\sqrt{[\boldsymbol{V}_\eta]_{ii}}$$

where $[\boldsymbol{V}_\eta]_{ii}$ is the ith diagonal element of the covariance matrix $\boldsymbol{V}_\eta = \boldsymbol{Z}_\eta \boldsymbol{F}_\eta^{-1} \boldsymbol{Z}_\eta^\top$. Here $\boldsymbol{Z}_\eta$ is the model matrix for the predictor η and $\boldsymbol{F}_\eta^{-1}$ are the elements for the spline of the predictor in the inverse of the expected Fisher information matrix.
Since the penalization is included in the Fisher information, we need to restrict the smoothing parameters to be small enough such that the Fisher information is not dominated by the penalty and the matrix stays invertible.

4 Simulation Study

We validate the suggested methods for flexible response functions by conducting a simulation study with $n = 1000$ observations and $N = 100$ replications for both binomial and Poisson data. Since the main part of our research concerns estimation of the response function, we provide a three step procedure to simulate the data. First, we simulate the predictor η. Here we generate two different scenarios comprising either only linear effects or considering smooth effects. The linear type is designed to compare our approach with the traditional single index models while the smooth type shows the benefits of our combination with additive predictors. The predictors in the simulation study are specified as follows:

$$x_1, x_2, x_3, x_4 \sim U[0, 1]$$
$$\eta = 3x_1 + 4x_2 - 4x_3 - 3x_4 \qquad \text{(Linear)}$$
$$\eta = -4 + 2\sin(6x_1) + 2\exp(x_2) + 2x_3 - 2x_4 \qquad \text{(Smooth)}$$

see Figure 1 for a graphic representation of the covariate effects. Furthermore, we simulate the covariates x_j as being independent for the linear case while they are correlated with $\rho = 0.5$ in the smooth design. The predictors are the same for binomial and Poisson data.

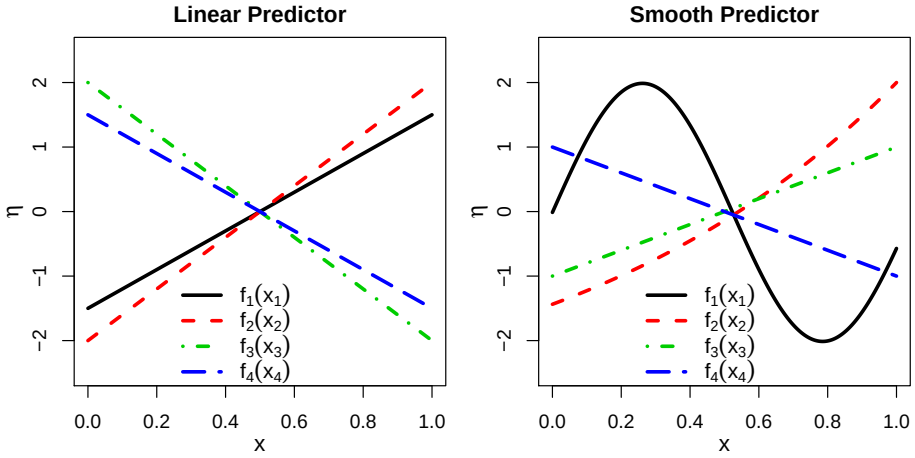

Figure 1: Covariate effects in the simulation study.

In the second step, we determine the expected response values by applying the response function h on the predictor. The response functions for the binomial case are

$$\begin{aligned} h_{Logit}(\eta_i) &= \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \\ h_{Gamma}(\eta_i) &= \text{cdf of } \Gamma(\eta_i + 2, \text{shape} = 2, \text{rate} = \sqrt{2}) \\ h_{Bimodal}(\eta_i) &= \frac{0.25}{1+\exp(-7.5\eta_i - 10)} \\ &\quad + \frac{0.75}{1+\exp(-7.5\eta_i + 10)} \end{aligned}$$

while the response functions for the Poisson data are specified as

$$\begin{aligned} h_{Log}(\eta_i) &= \exp(\eta_i/2) \\ h_{Pois1}(\eta_i) &= \frac{10}{1+\exp(-1.5\eta_i)} \\ h_{Pois2}(\eta_i) &= \frac{10}{1+\exp(-3.75\eta_i - 7.5)} + \\ &\quad \frac{10}{1+\exp(-3.75\eta_i + 7.5)} \end{aligned}$$

The functions $h_{Bimodal}$, h_{Pois1} and h_{Pois2} are similar to response functions of the simulation study in Tutz and Petry (2012). Furthermore, we include the logit and the log response functions as a benchmark, where our flexible methods are not necessary and the ordinary GAM is sufficient. All response functions are visualized in Figure 2.
As the third step, we use the expected values $\mu_i = h(\eta_i)$ as parameters for the simulation of the responses y_i:

$$y_i \sim B(1, \mu_i) \qquad \text{and} \qquad y_i \sim Po(\mu_i)$$

All estimations are done in `R` (R Core Team, 2017). As models, we consider the following alternatives:

GAM: Classical GAM based on code of the R-package `mgcv`.

IC: Single index model of Ichimura (1993) with code of the R-package `np`.

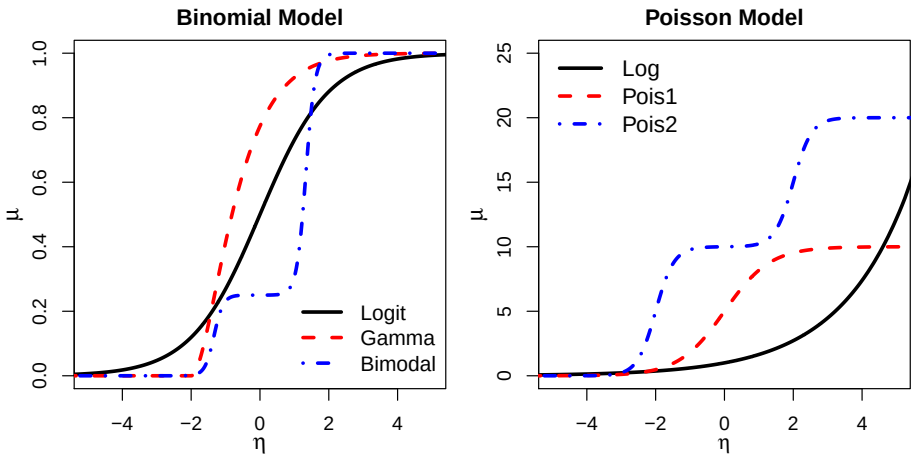

Figure 2: Response functions in the simulation study.

SIBoost: Single index boosting model of Tutz and Petry (2012) for the linear case and Tutz and Petry (2016) for the smooth case. The code of the linear case is attached to their paper, while the code for the smooth case was provided to us by the authors.

FlexGAM1: Indirect estimation of the response function as described in Section 3.1 with code of our `FlexGAM` package (see supplementary material).

FlexGAM2: Direct estimation of the response function as described in Section 3.2 with code of our `FlexGAM` package (see supplementary material).

The initial models of FlexGAM1 and FlexGAM2 are estimated including the intercept.
In the following, we compare two different settings. First, we focus on the comparison with the existing methods by applying the linear predictors to generate the data and to analyze the data (Section 4.1). Second, we benchmark the error that occurs by generating the data with the smooth predictor, but only applying the linear predictor in the single index models, while the logit model, the model of Tutz and Petry (2016) and our models use the semiparametric predictor (Section 4.2). As measure for the goodness of fit, we use the bias, the root mean squared error (RMSE) and the predictive deviance. The bias and the RMSE are estimated from the difference between the true underlying expected values of the data generating process μ_i and the fitted values $\hat{\mu}_i$ while the predictive deviance is estimated by applying a validation data set of the same data generating process (but with $n = 10.000$ observations) to the estimated models and comparing the responses y_i with the fitted values $\hat{\mu}_i$:

$$Bias = \frac{1}{n}\sum_{i=1}^{n} \mu_i - \hat{\mu}_i$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\mu_i - \hat{\mu}_i)^2}$$

$$Pdeviance = \begin{cases} 2\sum_{i=1}^{n} (1-y_i)\log\left(\frac{1}{1-\hat{\mu}_i}\right) + y_i \log\left(\frac{1}{\hat{\mu}_i}\right) \\ 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \end{cases}$$

The area under the receiver operator characteristic curve (AUC) would be another measure for the classification performance of a binomial regression model. However, in the literature (see for example Lobo et al., 2008; Hand, 2009, and citations therein) a fundamental discussion about the use of the AUC is going on. Furthermore, the AUC is based on a binary classification derived from a cut off for the predicted probabilities and will therefore only react to changes in probabilities close to that cut off. This results in a rather low sensitivity with respect to changes in the response

function as illustrated in the supplementary material. We therefore focus on the predictive deviance in the following which rewards improvements in predicting the success probability rather then rewarding the correct classification of the response.

4.1 Linear Predictor

4.1.1 Binomial Data

Most single index approaches are able to deal with linear predictors, therefore we build a simulation design based on linear predictors to check whether our approaches are able to compete with the existing ones. The resulting RMSE is plotted in the first row of Figure 3 where the red horizontal line indicates the median RMSE of the logit model as a benchmark.
These figures show that all single index models give more accurate results than the logit model if indeed the data are simulated with non-standard response functions. However, in the case that we have logistically distributed data, the standard GLM fits better than all single index models, except the FlexGAM1 approach originally proposed by Muggeo and Ferrara (2008). Overall, FlexGAM1 fits best in all data settings. Except for the pure logistic data, both approaches FlexGAM1 and FlexGAM2 behave similarly. All estimated response functions are plotted in the supplementary material. By analyzing the estimated response curves, it is visible that the kernel methods regularly result in very wiggly estimates despite that fact that the bandwidth is optimized in the `np` package. Pre-specifying the bandwidth appropriately solves the problem of wiggliness, but the model fit does not change relevantly compared to our models, especially in the simulation design with smooth predictors.

4.1.2 Poisson Data

Besides the binomial model, Poisson data are also of interest in this paper. Therefore we present in the second row of Figure 3 their estimated RMSE for the case with linear predictors. Similar to the binomial case, the data generating process with log link deals as a benchmark. Based on the resulting RMSE, we can conclude that the FlexGAM1 model has a similar behavior as the standard GLM for this benchmark data setting, while the other models have a small drawback. For the other two data settings, the original GLM is not sufficient and the P-spline based methods result in smaller RMSE. The method based on Ichimura (1993) yields competitive results in this linear setting but the estimated response functions are not necessarily monotonically increasing.

4.2 Smooth Predictor

The new approaches proposed in this paper have the big advantage of being able to also deal with nonlinear covariate effects. To show this, we conduct the simulation study with the smooth predictor specified above. So there are two possible misspecifications, either the predictor is fixed to be only linear (IC), even if we have nonlinear covariate effects and otherwise the response function is fixed to a specified response function (GAM) and not able to capture the skewness of the response function. Only the new approaches (FlexGAM1, FlexGAM2) and the boosting approach of Tutz and Petry (2016) (SIBoost) are able to deal with both error types. As it can be seen from Figure B.22 in the supplementary material, all methods are approximately unbiased, even if the predictor is misspecified.

4.2.1 Binomial Data

In the third row of Figure 3, the estimated RMSE for models based on binomial data and smooth predictors is displayed. In the case of logistically distributed data, we see that the traditional single index model gets far worse RMSE than the GAM that uses the semiparametric predictors and our new approaches are within the range of the standard GAM. Beyond the data following a latent logistic distribution, the new approaches result in better RMSE than the traditional GAM, while the standard single index model results in worse RMSE.
The bias and the RMSE are measures for the goodness of fit on the given data sets. However, the predictive performance of the models is also essential. Therefore we compare the estimated

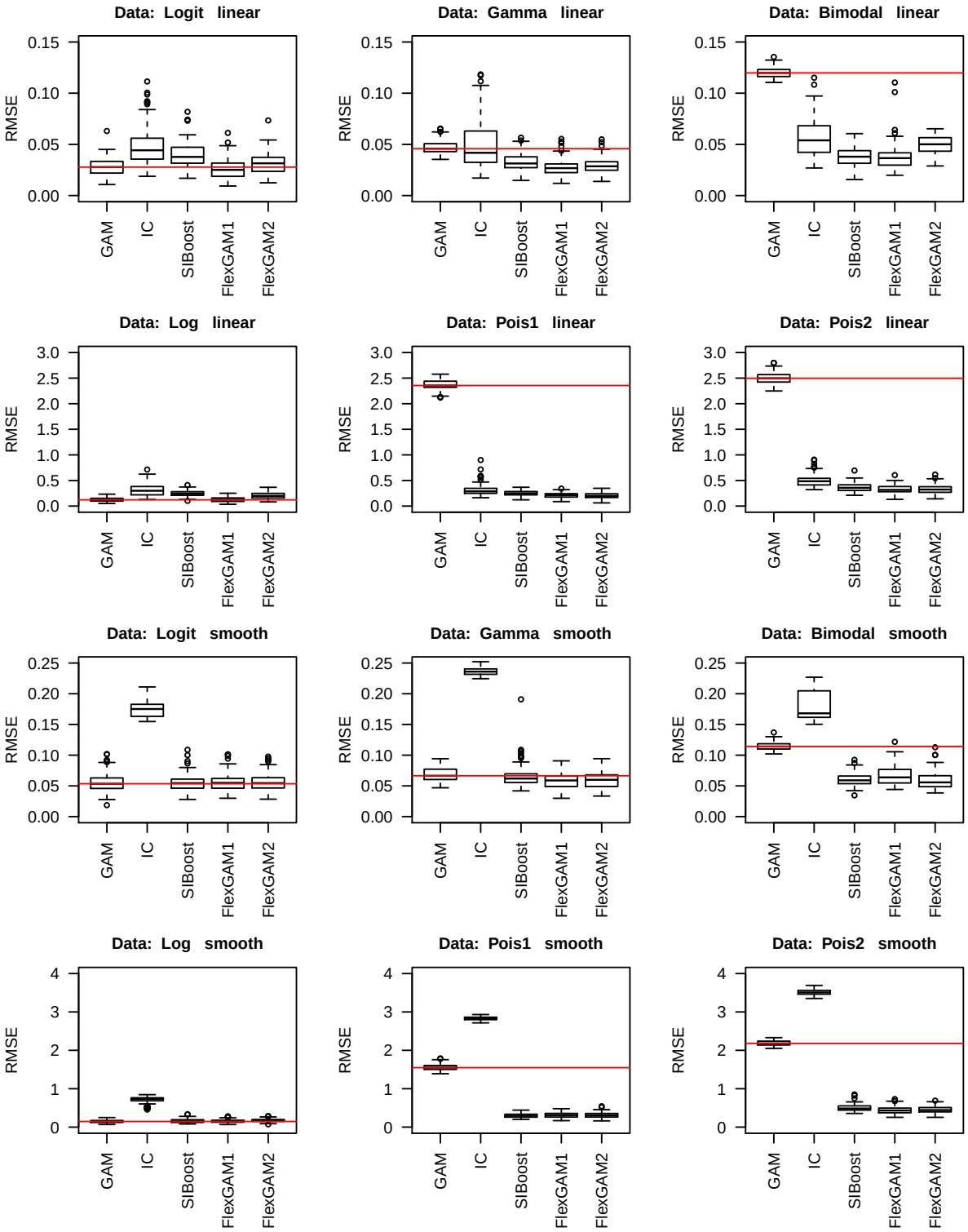

Figure 3: RMSE of binomial and Poisson models with linear and smooth predictors.

predictive deviances. Since these yield a similar pattern as with the RMSE, we show the results in the supplementary material Figure B.23.

Besides the goodness-of-fit criteria, the estimated response functions are of special interest. The estimated functions of the logit model as well as FlexGAM1 and FlexGAM2 models for the logistic and the bimodal data are given in Figure 4 (rows 1 and 2). The other functions are displayed in the supplementary material. From the pictures, we can conclude that our methods estimate the response functions correctly. In the logistic data, the FlexGAM2 method provides results with more variation than FlexGAM1. FlexGAM1 has the logit model as a limit, since a high penalization of the transformation function Ψ results in the identity function. Therefore we obtain

better estimates in the logistic setting. On the other side, the indirect estimation of the response function is not as flexible as the direct one, due to the slight distributional input of the logit link. This results in a worse estimation of the response function in the bimodal case. However, the response function estimated via FlexGAM1 also captures the effects a lot better than the logit model. Additionally to the estimated response functions, the estimated covariate effects are of interest. Therefore the estimated covariate effects of x_1 for the logit, FlexGAM1 and FlexGAM2 model for the logistic and the bimodal data are plotted in Figure 4 (rows 3 and 4). The other estimated effects are given in the supplementary material. We achieve comparability between the models by rescaling the results of the logit model with $\eta = \frac{\eta - \bar{\eta}}{\text{sd}(\eta)}$, as well as the underlying effect (red dashed line). From Figure 4 we conclude that our methods as well as the logit model identify the underlying effects correctly. However, our methods penalize the splines a bit more such that we get smoother results.

4.2.2 Poisson Data

Similar to the binomial data setting, the classical single index models are not able to deal with the nonlinear structure of the predictor in the Poisson case. However, the flexible approaches FlexGAM1 and FlexGAM2 both capture the covariate effects and the response function. Therefore their RMSE is lower than the one of the standard GAM, as shown in the forth row of Figure 3. Further index of the goodness of fit and the estimated response functions and covariate effects are plotted in the supplementary material.

4.3 Models without Monotonicity Constraint

Additionally to the models discussed above, we also applied our approaches without monotonicity constraints (*FlexGAM1n*, *FlexGAM2n*) in the simulation study. They show similar results in terms of the goodness-of-fit criteria. However, rather wiggly response functions are estimated and we therefore show the results of the non-monotonic estimates only in the electronic appendix.

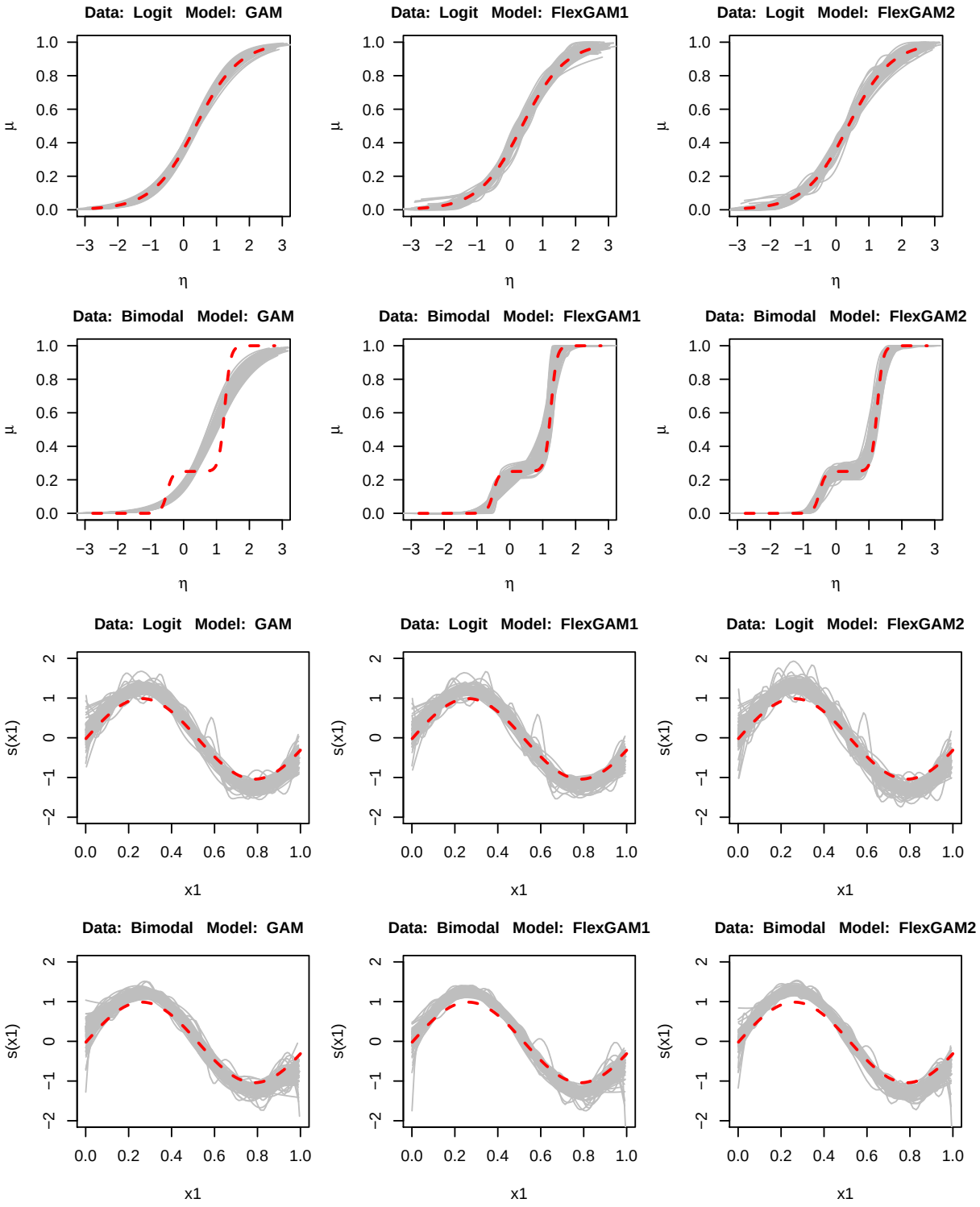

Figure 4: Estimated response functions and effects for x_1 of the logit, FlexGAM1 and FlexGAM2 model for the logistic and the bimodal data (grey, solid), with the true underlying function (red, dashed). The predictors are scaled simultaneously $\left(\eta = \frac{\eta - \bar{\eta}}{\mathrm{sd}(\eta)}\right)$.

5 Application

5.1 Mortality Rate in São Paulo

To illustrate our methods, we apply them to the data set used exemplary in Tutz and Petry (2016), where the mortality rate of 65 year old persons living in the area of São Paulo (Brazil) is estimated for the years 1994 to 1997. The original data set is available at `http://www.ime.usp.br/~jmsinger/Polatm9497.zip`. We make use of the subset of Leitenstorfer and Tutz (2007) which is also used in Tutz and Petry (2016). The sample size is $n = 1351$ and the considered variables are described in Table 1.

Variable	Explanation
`RES65`	Number of daily deaths caused by respiratory reasons (0-12).
`TEMPO`	Time in days (1461 in total).
`SO2ME.2`	The 24-hours mean of SO_2 concentration (in $\mu g/m^3$) over all monitoring measurement stations, led by 2 days.
`TMIN.2`	The daily minimum temperature, led by 2 days.
`UMID`	The daily relative humidity.
`CAR65`	Cardiologically caused deaths per day.
`OTH65`	Other (non respiratory or cardiological) deaths per day.

Table 1: Variables to model the death rate in São Paulo.

As response variable, we take the number of deaths caused by respiratory diseases. All other variables are used as smooth covariates applying P-splines with 20 inner knots. We compare the standard GAM model (*GAM*), the boosting model of Tutz and Petry (2016) (*SIBoost*) and our two approaches *FlexGAM1* and
FlexGAM2. For GAM, SIBoost and FlexGAM1 we choose the log-link. The smoothing parameters were optimized for GAM with the GCV criterion, for FlexGAM1 and FlexGAM2 via cross-validation, while we choose $\lambda_h = 1$ and $\lambda_f = 0.01$ for SIBoost as in Tutz and Petry (2016). As initial model for FlexGAM1 and FlexGAM2 we chose the standard Poisson model without intercept, since these models had a lower predictive deviance. The estimated response functions are displayed in Figure 5. They show that assuming a pure log-link does not give sufficient results, since the estimated curves show faster increasing behavior for positive predictors and slower for negative predictors. Here we scaled the predictors via $\left(\eta = \frac{\eta - \bar{\eta}}{\mathrm{sd}(\eta)}\right)$ to be comparable.

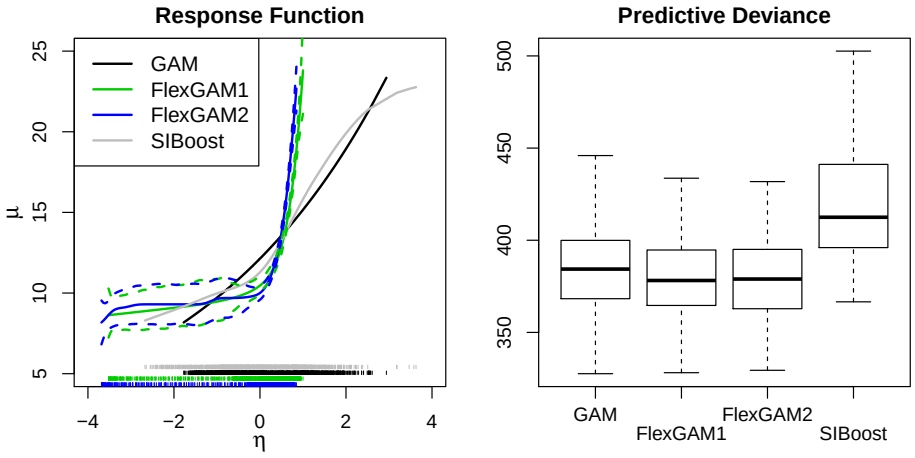

Figure 5: Estimated effects for the death rate data. On the left the estimated response function of the full data set, with confidence intervals in dashed lines, are displayed. Therefore the effects are standardized $\left(\frac{\eta - \bar{\eta}}{\mathrm{sd}(\eta)}\right)$. On the right the predictive deviance of 50 models with randoms splits of the data set.

We checked the predictive behavior of the models by estimating 50 models based on random splits of the data set, with sample size of 1000 for the training data set. To speed up the analysis,

the smoothing parameters in the random splits were fixed to the output of the original model with the full data set. The resulting predictive deviance is plotted in Figure 5. Thereof we may conclude that FlexGAM1 and FlexGAM2 give similar results, which are better than the one by pure GAM. In contrast to the paper by Tutz and Petry (2016) SIBoost resulted in worse predictive patterns than the standard GAM. However, in our design the standard GAM had a lower predictive deviance, than in the original paper, which might explain the problem.
Additionally, we provide the estimates of the covariate effects in Figure 6. Here, we can see that the variable `TEMPO` has the largest impact (consider the different scaling of the y-axis). It shows the seasonal dependence of the mortality rate. Besides with increasing SO_2 concentration the mortality rate increases, while it decreases with increasing humidity. Differences between the models should be analyzed with care, since the models are scaled. So only differences in the shapes can be analyzed, while the absolute value can only be considered jointly with the response function. All shifting and stretching effects are compensated by the response function and also by the centering of the splines. The upwards shift of the `TEMPO` variable for small values for example depends on the large negative values for $1000 <$ `TEMPO` < 1300. However, the divergence from the usual seasonal pattern is a specific pattern of the new approaches. This effect might be explained by some "unobserved" covariate. For the other covariates only the cardiological deaths (`CAR65`) show a slight change in the new models. Generally the new models put even more emphasis on the seasonal effect, since compared to this effect the other covariates decline in their impact.

5.2 Credit Scoring

As a second example and to apply our new methods to binomial data, we use credit scoring data of a German bank. The data was published in Fahrmeir et al. (1996) and is available online at `https://data.ub.uni-muenchen.de/23/`. Here we use a sample size of $n = 1000$, but we truncated 12 outliers of the continuous covariates. Table 2 describes the variables used in the model.

Variable	Explanation
`credit`	Whether a person repaid its credit (response).
`moral`	Whether the person has a good previous payment behavior.
`guarantor`	Whether the credit is secured by other persons (0 = non, 1 = other person involved, 2 = guarantor).
`duration`	Duration of the credit (0-60).
`amount`	Amount of the credit (250-14896).
`age`	Age of the person (19-70).

Table 2: Variables to model the credit scoring rate.

Similar to the example above, we estimate a model with P-splines with 20 inner knots for the smooth covariates (`duration`, `amount`, `age`). In addition, the categorical covariates (`moral`, `guarantor`) were also included. As initial model for FlexGAM1 and FlexGAM2 we chose the standard logit model without intercept to take care of the categorical covariates in the design matrix. Since the code for *SIBoost* provided by the authors does only support continuous covariates, we cannot estimate their model in this setting. Hence we only compare the three models *GAM*, *FlexGAM1* and *FlexGAM2*, with logit link if necessary. Therefore we first estimate a model with optimized smoothing parameters based on the full data set and afterwards 50 models based on training data sets of size 800. The resulting response functions of the full models are given in Figure 7 next to the estimated predictive deviance of the random split models. Here the estimated response functions are scaled again according to $\eta = \frac{\eta - \bar{\eta}}{\mathrm{sd}(\eta)}$.
The estimated response functions describe a flat area for η between -4 and -1. This behavior shows that there is some unobserved heterogeneity in the model. Furthermore, it indicates that using the logit link is not sufficient. Here we occasionally find that the pointwise confidence intervals are not monotonically increasing. This results from the increasing uncertainty associated with a smaller number of observations on the outer part of the parameter space. Estimating simultaneous confidence intervals instead of pointwise intervals could solve this problem. From the values of the

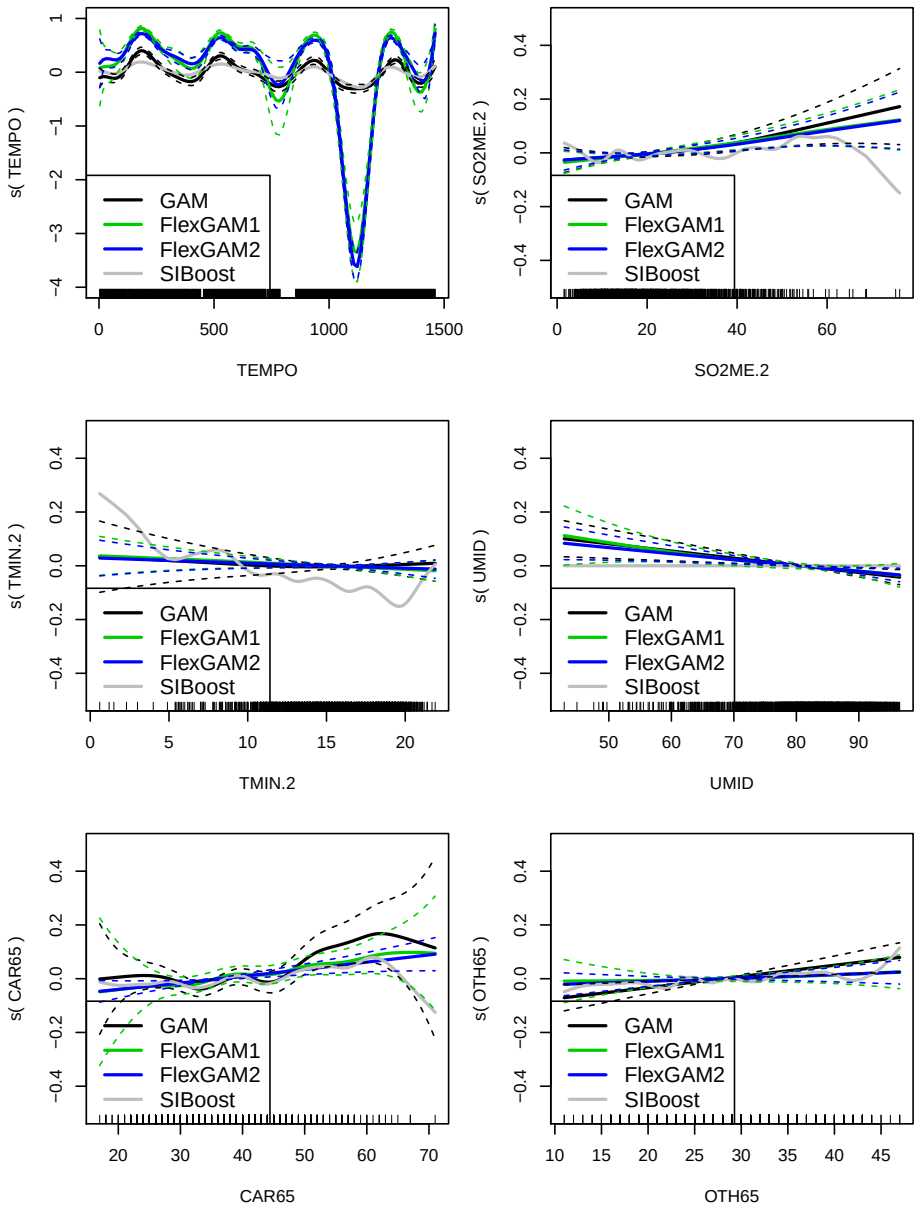

Figure 6: Estimated covariate effects for the death rate data.

predictive deviance we conclude that our approaches are able to better deal with the unobserved parts, so the deviance is lower.
Besides the estimated response function also the estimated covariate effects are of interest. Therefore we show in Figure 8 the estimates of the smooth effects. With increasing duration, the probability of paying back the credit declines. Contrarily with increasing age the probability increases until the person ages 40, than the effect is constant. Small credits and larger credits have a higher probability of default, while the medium sized credits are rather surely payed back, if the other covariates stay constant. Here rather small differences between the estimates of the logit model (black) and the new models (green, blue) occur. However, in the middle of the parameter space for `amount` and `duration` the new models show higher values, while they decrease faster for higher values.

6 Conclusion

Based on our simulation study and the empirical examples, we conclude that estimating the response function along with a flexible predictor often leads to a better model fit. We have proposed two approaches for estimating the response function, where one is more flexible while the other has

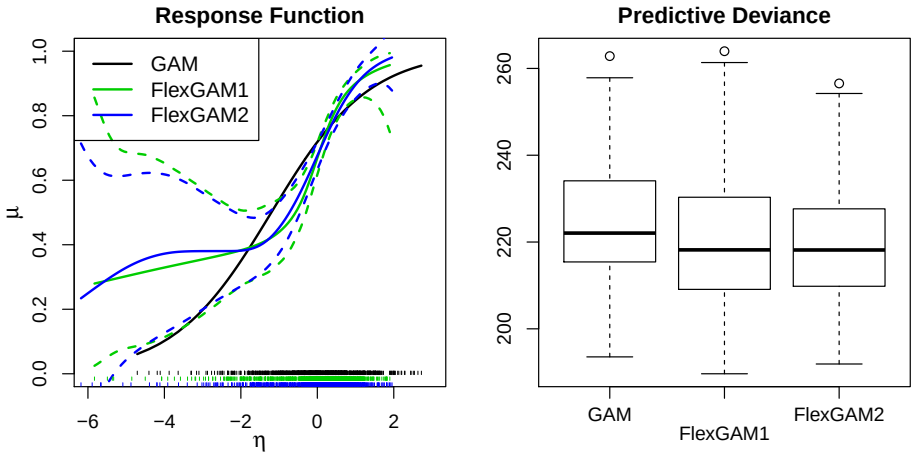

Figure 7: Estimated effects for the credit scoring data. On the left the estimated response function of the full data set, with confidence intervals in dashed lines, are displayed. Therefore the effects are standardized $\left(\frac{\eta-\bar{\eta}}{\mathrm{sd}(\eta)}\right)$. On the right the predictive deviance of 50 models with randoms splits of the data set.

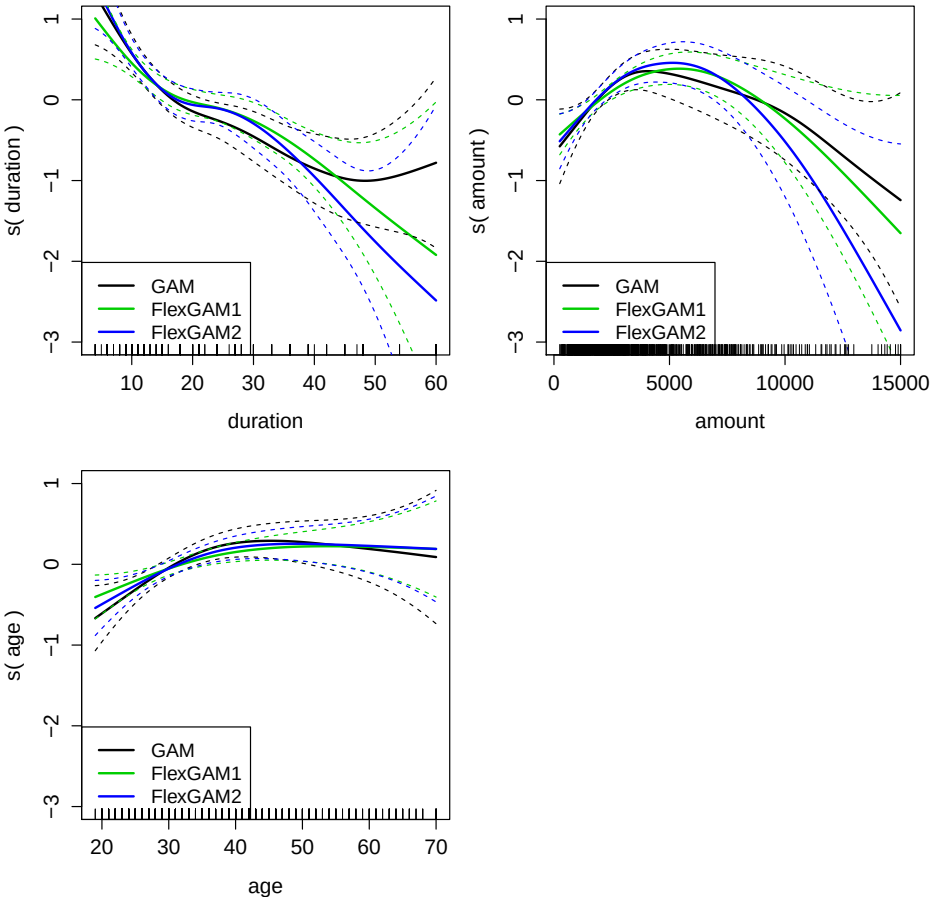

Figure 8: Estimated covariate effects for the credit scoring data.

the advantage of having the canonical response function as the natural limit. Both are working very well in the simulation studies so the user may decide which property she/he prefers. An important step in our approach is first of all the identifiability which we achieve with the introduction of several constraints, and second the correct estimation of the smoothing parameters. If the smoothing parameters are misspecified, the predictive properties decline. Therefore we accept the computational intensive cross-validation for simultaneously determining the smoothing parameters

of the response function and the semiparametric predictor. Alternatives for the estimation of the smoothing parameters using cross-validation could be, first of all, an adaption of the approach by Wood and Fasiolo (2017) to our combined likelihoods such that the estimation of the covariate effects and of the smoothing parameters is done jointly. Second, a theoretically well-grounded estimation of the degrees of freedom could be established in on our interdependent likelihoods such that the generalized cross-validation criterion could be applied and the number of model fits could be reduced. Both alternatives are left for further research.
The flexibility of the approach provides several benefits, but it has a drawback, namely the correct specification of the model. This specification has a relevant impact on the goodness-of-fit of each GAM. Therefore several approaches to select models with additive structure have been proposed (see Marra and Wood, 2011, for an overview). The flexible response function may capture some unobserved effects, however structured approaches on model selection for GAM with flexible responses are left for further research.

Acknowledgements

We thank Sebastian Petry for providing the code to the paper of Tutz and Petry (2016), such that we could compare our method with the boosting approach. We also want to thank two anonymous referees and an associate editor for their helpful comments improving this paper. Moreover we acknowledge financial support by the German Research Foundation (DFG), grant KN 922/4-2.

References

Bollaerts, K., P. H. Eilers, and I. Mechelen (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology 59*(2), 451–469.

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22*(4), 477–505.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Carroll, R. J., J. Fan, I. Gijbels, and M. P. Wand (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association 92*(438), 477–489.

Czado, C. and T. J. Santner (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference 33*(2), 213–231.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer Verlag.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–121.

Fahrmeir, L., A. Hamerle, and G. Tutz (1996). *Multivariate Statistische Verfahren*. Walter de Gruyter GmbH & Co KG.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression: Models, Methods and Applications*. Springer Science & Business Media.

Friedman, J. H. and W. Stuetzle (1981). Projection pursuit regression. *Journal of the American Statistical Association 76*(376), 817–823.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning 77*(1), 103–123.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science 1*(3), 297–310.

Held, L. and D. Sabanés Bové (2014). *Applied Statistical Inference*. Berlin/Heidelberg: Springer Verlag.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics 58*(1), 71–120.

Jørgensen, B. (1984). The delta algorithm and GLIM. *International Statistical Review/Revue Internationale de Statistique 52*(3), 283–300.

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society 61*(2), 387–421.

Koenker, R. and J. Yoon (2009). Parametric links for binary choice models: A fisherian–bayesian colloquy. *Journal of Econometrics 152*(2), 120–130.

Leitenstorfer, F. and G. Tutz (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics 8*(3), 654–673.

Leitenstorfer, F. and G. Tutz (2011). Estimation of single-index models based on boosting techniques. *Statistical Modelling 11*(3), 203–217.

Li, Q. and J. S. Racine (2007). *Nonparametric econometrics: theory and practice*. Princeton: Princeton University Press.

Lobo, J. M., A. Jiménez-Valverde, and R. Real (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography 17*(2), 145–151.

Marra, G. and S. N. Wood (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis 55*(7), 2372–2387.

Marra, G. and S. N. Wood (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics 39*(1), 53–74.

Marschner, I. C. et al. (2011). glm2: fitting generalized linear models with convergence problems. *The R Journal 3*(2), 12–15.

Marx, B. D. (2015). Varying-coefficient single-index signal regression. *Chemometrics and Intelligent Laboratory Systems 143*, 111–121.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). Chapman & Hall.

Muggeo, V. M. and G. Ferrara (2008). Fitting generalized linear models with unspecified link function: A P-spline approach. *Computational Statistics & Data Analysis 52*(5), 2529–2537.

Pya, N. (2017). *scam: Shape Constrained Additive Models*. R package version 1.2-2.

Pya, N. and S. N. Wood (2015). Shape constrained additive models. *Statistics and Computing 25*(3), 543–559.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Tutz, G. and S. Petry (2012). Nonparametric estimation of the link function including variable selection. *Statistics and Computing 22*(2), 545–561.

Tutz, G. and S. Petry (2016). Generalized additive models with unknown link function including variable selection. *Journal of Applied Statistics 43*(15), 2866–2885.

Wang, J.-L., L. Xue, L. Zhu, Y. S. Chong, et al. (2010). Estimation for a partial-linear single-index model. *The Annals of Statistics 38*(1), 246–274.

Weisberg, S. and A. Welsh (1994). Adapting for the missing link. *The Annals of Statistics 22*(4), 1674–1700.

Wood, S. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing 15*(5), 1126–1133.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2 ed.). CRC Press.

Wood, S. N. and M. Fasiolo (2017). A generalized fellner-schall method for smoothing parameter optimization with application to tweedie location, scale and shape models. *Biometrics*.

Yu, Y. and D. Ruppert (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association 97*(460), 1042–1054.

Yu, Y., C. Wu, and Y. Zhang (2017). Penalised spline estimation for generalised partially linear single-index models. *Statistics and Computing 27*(2), 571–582.

7 Appendix

Algorithm 1 (FlexGAM1)

Start:

$$\hat{\boldsymbol{\gamma}}^{(0)} = \texttt{gam(}\boldsymbol{y} \sim \boldsymbol{x}_1 + \boldsymbol{x}_2 + s(\boldsymbol{x}_r) + s(\boldsymbol{x}_{r+1}) + \ldots\texttt{,}\ \texttt{family=...)}$$

$$\boldsymbol{\eta}_1^{(0)} = \boldsymbol{Z}\hat{\boldsymbol{\gamma}}^{(0)}$$

$$\Rightarrow \qquad \boldsymbol{\eta}^{(0)} = \frac{\boldsymbol{\eta}_1^{(0)} - mean(\boldsymbol{\eta}_1^{(0)})}{sd(\boldsymbol{\eta}_1^{(0)})}$$

Outer (m) :

$$\Psi^{(m)}(\boldsymbol{\eta}^{(k-1)}) = \texttt{scam(}\boldsymbol{y} \sim s(\boldsymbol{\eta}^{(k-1)}\texttt{, bs="mpi")},\ \texttt{family=...)}$$

Inner (k) :

$$\boldsymbol{y}^{(k)} = \boldsymbol{\eta}^{(k-1)} + \frac{\boldsymbol{y} - h(\Psi^{(m)}(\boldsymbol{\eta}^{(k-1)}))}{h^{'}(\Psi^{(m)}(\boldsymbol{\eta}^{(k-1)}))\ \Psi'^{(m)}(\boldsymbol{\eta}^{(k-1)})}$$

$$\boldsymbol{w}^{(k)} = \frac{\left(h^{'}(\Psi^{(m)}(\boldsymbol{\eta}^{(k-1)}))\ \Psi'^{(m)}(\boldsymbol{\eta}^{(k-1)})\right)^2}{Var(\boldsymbol{\eta}^{(k-1)})}$$

$$\hat{\boldsymbol{\gamma}}^{(k)} = \left(\boldsymbol{Z}^\top \boldsymbol{W}^{(k)} \boldsymbol{Z} + \boldsymbol{K}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{W}^{(k)} \boldsymbol{y}^{(k)}$$

via `mgcv::gam`

$$\boldsymbol{\eta}_1^{(k)} = \boldsymbol{Z}\hat{\boldsymbol{\gamma}}^{(k)}$$

$$\Rightarrow \qquad \boldsymbol{\eta}^{(k)} = \frac{\boldsymbol{\eta}_1^{(k)} - mean(\boldsymbol{\eta}_1^{(k)})}{sd(\boldsymbol{\eta}_1^{(k)})}$$

The inner iteration is done until the convergence of $\boldsymbol{\gamma}^{(k)}$, meaning $\frac{||\boldsymbol{\gamma}^{(k)} - \boldsymbol{\gamma}^{(k-1)}||}{||\boldsymbol{\gamma}^{(k)}||} < \varepsilon_1$. Then the outer iteration is repeated. The inner and outer loops are iterated until the coefficients of $\Psi^{(m)}$ are constant, meaning $\frac{||\boldsymbol{\nu}^{(m)} - \boldsymbol{\nu}^{(m-1)}||}{||\boldsymbol{\nu}^{(m)}||} < \varepsilon_2$.

Algorithm 2 (FlexGAM2)

Start:

$$\hat{\gamma}^{(0)} = \texttt{gam(}\boldsymbol{y} \sim \boldsymbol{x}_1 + \boldsymbol{x}_2 + s(\boldsymbol{x}_r) + s(\boldsymbol{x}_{r+1}) + \ldots\texttt{,}\ \texttt{family=}\ldots\texttt{)}$$

$$\boldsymbol{\eta}_1^{(0)} = \boldsymbol{Z}\hat{\gamma}^{(0)}$$

$$\Rightarrow \qquad \boldsymbol{\eta}^{(0)} = \frac{\boldsymbol{\eta}_1^{(0)} - mean(\boldsymbol{\eta}_1^{(0)})}{sd(\boldsymbol{\eta}_1^{(0)})}$$

Outer (m) :

$$\hat{h}^{(m)}(\boldsymbol{\eta}^{(k-1)}) = \texttt{pcls(}\boldsymbol{y} \sim s(\boldsymbol{\eta}^{(k-1)}\texttt{, bs="ps")}\ \texttt{/}\ \boldsymbol{A\nu} \geq \boldsymbol{b}\texttt{)}$$

Inner (k) :

$$\boldsymbol{y}^{(k)} = \boldsymbol{\eta}^{(k-1)} + \frac{\boldsymbol{y} - \hat{h}^{(m)}(\boldsymbol{\eta}^{(k-1)})}{\hat{h}'^{(m)}(\boldsymbol{\eta}^{(k-1)})}$$

$$\boldsymbol{w}^{(k)} = \frac{\left(\hat{h}'^{(m)}(\boldsymbol{\eta}^{(k-1)})\right)^2}{Var(\boldsymbol{\eta}^{(k-1)})}$$

$$\hat{\gamma}^{(k)} = \left(\boldsymbol{Z}^\top \boldsymbol{W}^{(k)} \boldsymbol{Z} + \boldsymbol{K}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{W}^{(k)} \boldsymbol{y}^{(k)}$$

via `mgcv::gam`

$$\boldsymbol{\eta}_1^{(k)} = \boldsymbol{Z}\hat{\gamma}^{(k)}$$

$$\Rightarrow \qquad \boldsymbol{\eta}^{(k)} = \frac{\boldsymbol{\eta}_1^{(k)} - mean(\boldsymbol{\eta}_1^{(k)})}{sd(\boldsymbol{\eta}_1^{(k)})}$$

Again the inner iteration is done until the convergence of $\boldsymbol{\gamma}^{(k)}$, Then the outer iteration is repeated. The inner and outer loops are iterated until the coefficients of $\hat{h}^{(m)}$ are constant, meaning $\frac{||\boldsymbol{\nu}^{(m)} - \boldsymbol{\nu}^{(m-1)}||}{||\boldsymbol{\nu}^{(m)}||} < \varepsilon_2$.