



LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK



Gerhard Tutz & Sebastian Petry

Nonparametric Estimation of the Link Function Including Variable Selection

Technical Report Number 085, 2010
Department of Statistics
University of Munich

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



Nonparametric Estimation of the Link Function Including Variable Selection

Gerhard Tutz & Sebastian Petry
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
{tutz, petry}@stat.uni-muenchen.de

July 9, 2010

Abstract

Nonparametric methods for the estimation of the link function in generalized linear models are able to avoid bias in the regression parameters. But for the estimation of the link typically the full model, which includes all predictors, has been used. When the number of predictors is large these methods fail since the full model can not be estimated. In the present article a boosting type method is proposed that simultaneously selects predictors and estimates the link function. The method performs quite well in simulations and real data examples.

Keywords: Single-Index Models, P-splines, Choice of Link Function, Variable Selection, Nonparametric Estimation of Link Function.

1 Introduction

In generalized linear models (GLMs), for given data (y_i, \mathbf{x}_i) , $i = 1, \dots, n$, the conditional expectation of $y_i | \mathbf{x}_i$, $\mu_i = E(y_i | \mathbf{x}_i)$, is modeled by

$$g(\mu_i) = \eta_i \quad \text{or} \quad \mu_i = h(\eta_i),$$

where $\eta_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}$ is the linear predictor, $g(\cdot)$ is the link function and $h(\cdot) = g^{-1}(\cdot)$ is the response function.

Usually it is assumed that the response function $h(\cdot)$ is fixed and known, for example $h(\cdot) = \exp(\cdot)$ yields the loglinear model which represents the canonical

link model if responses follow a Poisson distribution. In applications typically the link function is unknown and frequently the canonical link function is used. But it is well known that misspecification of the link function can lead to substantial bias in the regression parameters (see Czado and Santner (1992) for binomial responses). That may be avoided by flexible modelling of the link.

When responses are metrically scaled a flexible generalization of classical approaches is the so-called single-index model. It assumes that $h(\cdot)$ is unknown and has to be estimated by nonparametric techniques. The model may be seen as a special case of projection pursuit regression, which assumes that μ_i has additive form $h_1(\mathbf{x}_i^T \boldsymbol{\beta}_1) + \dots + h_m(\mathbf{x}_i^T \boldsymbol{\beta}_m)$ with unknown functions h_1, \dots, h_m , which transform the indices $\mathbf{x}_i^T \boldsymbol{\beta}_j$, see Friedman and Stützle (1981). In single index models only one index, $\mathbf{x}_i^T \boldsymbol{\beta}$, is assumed. The main difference between a single index model and a GLM is that in the former the transformation function $h(\cdot)$ is not restricted whereas in GLMs it is assumed that $h(\cdot)$ is strictly monotone and hence invertible. Although single index models are useful in dimension reduction, strict monotonicity, as assumed in GLMs, is very helpful when parameters are to be interpreted. Therefore we will focus on monotonic response functions. Then nonparametric estimation of the function $h(\cdot)$ may be seen as estimation of the unknown link function in a GLM.

Estimation of the unknown link function when the underlying distribution is from a simple exponential family was considered for example by Weisberg and Welsh (1994), Ruckstuhl and Welsh (1999) and Muggeo and Ferrara (2008). Weisberg and Welsh (1994) proposed to estimate regression coefficients using the canonical link and then estimate the link via kernel smoothers given the estimated parameters. Then parameters are reestimated. Alternating between estimation of link and parameters yields consistent estimates. But all these approaches do not select predictors.

The main advantage of the presented approach is that it combines estimation of the link function with variable selection. In the last decade the traditional forward/backward procedures for the selection of variables have been widely replaced by regularized estimation methods that implicitly select predictors, among them the Lasso (Tibshirani (1996)), which was adapted to GLMs by Park and Hastie (2006), the Dantzig selector (James and Radchenko (2008)), SCAD (Fan and Li (2001)) and boosting approaches (Bühlmann and Hothorn (2007), Tutz and Binder (2006)). However, in all of these procedures selection is always based on a known response function. If the assumed response function is wrong the performance of these selection procedures can be strongly affected. For illustration let us consider a small simulation study.

We fitted a Poisson model with the true response function having sigmoidal form $h_T(\eta) = 10/(1 + \exp(-5 \cdot \eta))$, see Fig. 2. The parameter vector of length $p = 20$ was $\boldsymbol{\beta}^T = (0.2, 0.4, -0.4, 0.8, 0, \dots, 0)$ and covariates were drawn from a normal distribution $\mathbf{X} \sim N(\mathbf{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \{\sigma_{ij}\}_{i,j \in \{1, \dots, p\}}$ where $\sigma_{ij} = 0.5$, $i \neq j$, $\sigma_{ii} = 1$. We generated $N = 50$ data sets with $n = 200$ observations and

fitted the model by using the usual maximum likelihood (ML) procedure based on the canonical log-link (without variable selection). In addition, we applied three alternative fitting methods that include variable selection: the nonparametric flexible link procedure derived in Section 2.2, the lasso for generalized linear models (Lokhorst, Venables, Turlach, and Maechler (2007)) and a boosting procedure (Hothorn, Bühlmann, Kneib, Schmid, and Hofner (2009)). The latter procedure is based on componentwise boosting, which is also the selection procedure used in the flexible link procedure. While the flexible link procedure selects a link function, ML estimates as well as lasso and boosting use the canonical link. It is seen in the upper four panels from Figure 1 that lasso and boosting, which include variable selection perform distinctly better than classical maximum likelihood fitting. But the best results are obtained if the link function is estimated nonparametrically. In particular the parameters of predictors that are not influential are estimated more stable and closer to zero. The dominance of the flexible procedure is also seen in the two lower panels from Figure 1, which shows the mean squared error for the estimation of the parameter vector and the predictive deviance on an independently drawn test data set with $n = 1000$. In Figure 2 the estimated response functions and the true response function $h_{\mathbf{T}}(\cdot)$ are shown. For more details see Section 3.

For normally distributed responses various estimation methods for single-index models have been proposed. One popular technique is based on average derivative estimation (see Stoker (1986), Powell, Stock, and Stoker (1989), Hristache, Juditsky, and Spokoiny (2001)). Alternatively M -estimation has been applied, which considers the unknown link function as an infinite dimensional nuisance parameter (see e.g. Klein and Spady (1993)). Other authors focus more on the estimation of $h(\cdot)$. Based on kernel regression techniques, Härdle, Hall, and Ichimura (1993) investigated the optimal amount of smoothing in single-index models when simultaneously estimating $\boldsymbol{\beta}$ and the bandwidth. Yu and Ruppert (2002) suggested to use penalized regression splines. They also allow for partially linear terms in the model and report more stable estimates compared to earlier approaches based on local regression (e.g. Carroll, Fan, Gijbels, and Wand (1997)). Tutz and Leitenstorfer (2009) proposed a boosted version of the penalized regression splines approach, but without variable selection. More recently, Gaïffas and Lecue (2007) proposed an aggregation algorithm with local polynomial fits and investigated optimal convergence rates. Bayesian approaches were proposed by Antoniadis, Gregoire, and McKeague (2004). More general distribution models have been considered by Weisberg and Welsh (1994) who proposed an algorithm that alternates between the estimation of $\boldsymbol{\beta}$ and $h(\cdot)$.

In the following we will extend the penalized regression splines approach used by Yu and Ruppert (2002) and Tutz and Leitenstorfer (2009) for netric response to the more general case of GLMs and in particular incorporate variable selection. In Section 2 the estimation procedure is given, in Section 3 the method is compared to competitors. In Section 4 a modified version that allows to reduce

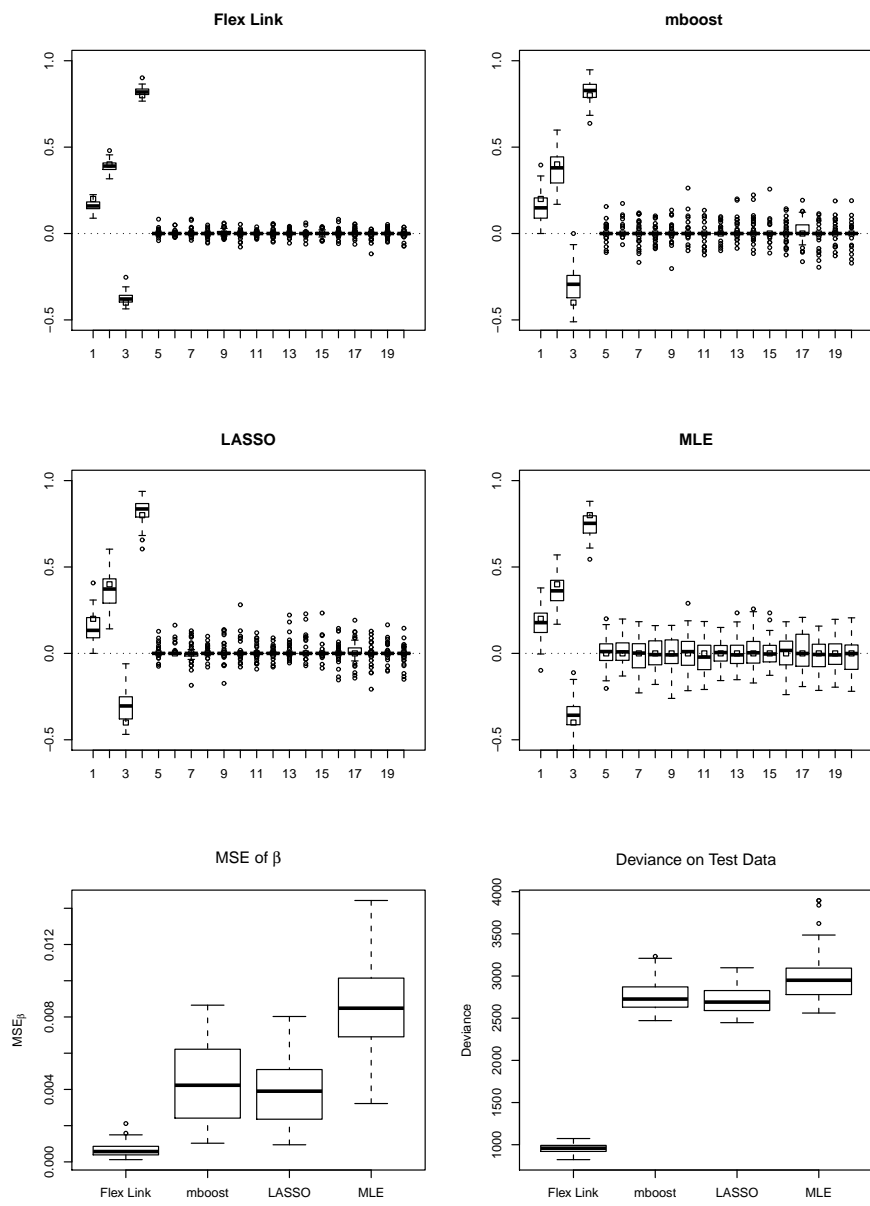


FIGURE 1: *Estimates of coefficient vector in simulation study for flexible link, boosting, lasso and ML and the mean squared error for parameter vector and predictive deviance for simulation setting.*

the false positives is introduced. Applications are given in Section 5.

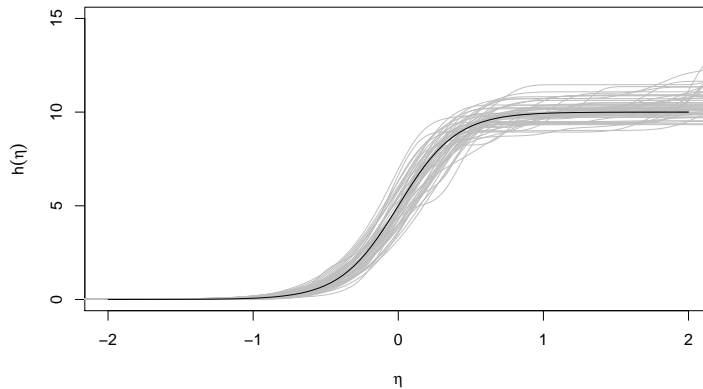


FIGURE 2: True (black) and the estimated (grey) response functions for simulation setting.

2 Estimation

2.1 Data Generating and Approximating Model

In the following it is assumed that the *data generating model* is

$$E(y_i|\mathbf{x}_i) = \mu_i = h_T(\eta_i),$$

where $h_T(\cdot)$ is the unknown true transformation function and $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ is the linear predictor. Given \mathbf{x}_i the y_i are (conditionally) independent observations from a simple exponential family

$$f(y_i|\theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}, \quad (1)$$

where θ_i is the natural parameter of the family, ϕ is a scale or dispersion parameter and $b(\cdot), c(\cdot)$ are specific functions corresponding to the type of the family. For uniqueness we will assume that for the true parameter $\|\boldsymbol{\beta}\| = 1$ holds and that the linear predictor η_i contains no intercept. Thus, the magnitude of $\|\boldsymbol{\beta}\|$ and the intercept are absorbed into h_T .

The *approximating model* that is fitted has the form

$$\mu_i = h_0(h(\eta_i)),$$

where $h_0(\cdot)$ is a fixed transformation function, which has to be chosen. The function $h(\cdot)$ is considered as unknown and has to be estimated. Typically, the choice of $h_0(\cdot)$ depends on the distribution of the response. When the response is binary a canonical choice is the logistic distribution function. The main advantage of specifying a fixed link function is that it may be selected such that the predictor

is automatically mapped into the admissible range of the mean response. For example, the logistic distribution function has values from $[0, 1]$, which is appropriate for binary responses. Thus, in contrast to existing procedures, which estimate $h_T(\cdot)$ directly, we estimate the inner function $h(\cdot)$.

The function $h(\cdot)$ will be approximated by expansion in basis functions

$$h(\eta_i) = \sum_{s=1}^k \alpha_s \phi_s(\eta_i) = \Phi_i^T \boldsymbol{\alpha}, \quad (2)$$

where ϕ_1, \dots, ϕ_k denote the basis functions. As basis functions we use natural B-splines of degree 3 (compare Dierckx (1993)), which are provided by the `fda` package in R. One problem with basis functions is that a sequence of knots $\{\tau_j\}_1^k$ has to be placed in a certain domain $[\eta_{\min}, \eta_{\max}]$ where the response function is to be estimated. Since the parameter vector is normalized by setting $\|\boldsymbol{\beta}\| = 1$, one can infer from the Cauchy-Schwarz-inequality that the range of $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, $i \in \{1, \dots, n\}$ is restricted to $[-u, u]$ where $u = \max_{i=1, \dots, n} \{\|\mathbf{x}_i\|_2\}$. We will use equidistant knots on $[-u, u]$. As in P-spline regression Eilers and Marx (1996), a high number of knots is used and the smoothness of the function estimate is controlled by appropriate penalization. As penalty term for the estimation of $\boldsymbol{\alpha}$ we use the integral of the squared second derivation of the approximated response function $h(\cdot)$ given by (2), $\int_{-u}^u \left(\frac{d^2}{d\eta^2} h(\eta)\right)^2 d\eta$, which can be given in matrix form as $\mathbf{P}_h = \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}$ with symmetric matrix $\mathbf{K} = (k_{ij})$, $k_{ij} = \int_{-u}^u \left(\frac{d^2}{d\eta^2} \phi_i(\eta)\right) \left(\frac{d^2}{d\eta^2} \phi_j(\eta)\right) d\eta$.

2.2 Estimation of Parameters Including Variable Selection

Componentwise boosting techniques have been successfully used to select relevant predictors in classical linear and generalized linear models (see for example the overview given by Bühlmann and Hothorn (2007)). The basic principle is to update within one step only one single component, in our case one coefficient of the predictor. With the link function being unknown also the coefficients of basis functions have to be estimated. In contrast to the selection procedure for the components of $\boldsymbol{\beta}$ the estimation of the coefficients of basis functions includes no selection step. Since the underlying link function is assumed to be smooth estimates are updated by using penalized estimation.

We will use likelihood-based boosting techniques, which aim at the maximization of the log-likelihood $l(\boldsymbol{\alpha}, \boldsymbol{\beta})$. As usual in boosting no explicit penalty on the log-likelihood is specified. Regularization is obtained implicitly by stopping the iteration procedure. The specific advantages of boosting techniques concerning the tradeoff between bias and variance have been derived by Bühlmann and Yu (2003). Moreover, it has been shown that in special cases boosting is very similar to lasso regularized estimates (see Efron, Hastie, Johnstone, and Tibshirani

(2004)). The penalization techniques that are used here follow the same principles as likelihood-based boosting outlined in Tutz and Binder (2006).

Computation of estimates uses boosting techniques in two stages, once for the estimation of the parameter vector $\boldsymbol{\beta}$ and once for the estimation of the vector of basis coefficients $\boldsymbol{\alpha}$. Before giving the algorithm we will consider the two stages (and initialization) separately. For simplicity we will use matrix notation with \mathbf{X} denoting the design matrix of predictors, and $\hat{\boldsymbol{\beta}}^{(l)}$, $\hat{\boldsymbol{\eta}}^{(l)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(l)}$ denoting the parameter estimate and the fitted predictor in the l th step. Moreover, $\boldsymbol{\Phi}^{(l)} = (\boldsymbol{\Phi}_1^{(l)}, \dots, \boldsymbol{\Phi}_n^{(l)})^T$ with $\boldsymbol{\Phi}_i^{(l)} = (\phi_1(\hat{\eta}_i^{(l)}), \dots, \phi_k(\hat{\eta}_i^{(l)}))^T$ is the current design matrix for the basis functions.

Initialization

We need two initialization values, $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\alpha}^{(0)}$. For $\boldsymbol{\beta}^{(0)}$ we choose $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p$. The initialization value for the coefficient vector of the basis functions $\boldsymbol{\alpha}^{(0)}$ is generated by approximating h by a linear function, $s \cdot \eta + t$, where $t = h_0^{-1}(\bar{y})$ and the slope is chosen as a small value ($s = 0.0001$).

Boosting for Fixed Predictor

For fixed predictor $\hat{\boldsymbol{\eta}}^{(l-1)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(l-1)}$ the estimation of the response function corresponds to fitting the model $\boldsymbol{\mu} = h_0((\boldsymbol{\Phi}^{(l-1)})^T \hat{\boldsymbol{\alpha}}^{(l-1)} + (\boldsymbol{\Phi}^{(l-1)})^T \hat{\boldsymbol{a}}^{(l)})$ where $(\boldsymbol{\Phi}^{(l-1)})^T \hat{\boldsymbol{a}}^{(l-1)}$ is a fixed offset that represents the previously fitted value. One step of penalized Fisher scoring has the form

$$\hat{\boldsymbol{a}}^{(l)} = \nu_h \left((\boldsymbol{\Phi}^{(l-1)})^T \hat{\mathbf{D}}^{(l-1)} (\hat{\boldsymbol{\Sigma}}^{(l-1)})^{-1} \hat{\mathbf{D}}^{(l-1)} \boldsymbol{\Phi}^{(l-1)} + \lambda_h \mathbf{P}_h \right)^{-1} \cdot (\boldsymbol{\Phi}^{(l-1)})^T \hat{\mathbf{D}}^{(l-1)} (\hat{\boldsymbol{\Sigma}}^{(l-1)})^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}) \quad (3)$$

where

$$\hat{\mathbf{D}}^{(l-1)} = \text{diag} \left\{ \frac{\partial h_0(\hat{h}^{(l-1)}(\hat{\eta}_i^{(l-1)}))}{\partial h^{(l-1)}(\eta)} \right\}_{i=1}^n \quad (4)$$

is the estimate of the derivative matrix evaluated at the estimate of the previous step $h_0(\hat{h}^{(l-1)}(\eta))$ and

$$(\hat{\boldsymbol{\Sigma}}^{(l-1)}) = \text{diag} \left\{ \sigma^2(\hat{h}^{(l-1)}(\hat{\eta}_i^{(l-1)})) \right\}_{i=1}^n \quad (5)$$

is the matrix of variances evaluated at $h_0(\hat{h}^{(l-1)}(\eta))$. \mathbf{P}_h is the penalty matrix which penalizes the second derivation of the estimated (approximated) response function. The shrinkage parameter, which makes the procedure a weak learner, is fixed by $\nu_h = 0.1$.

Componentwise Boosting for Fixed Response Function

Let $h(\cdot)$ be fixed and the design matrix have the form $\mathbf{X} = (\mathbf{x}_1 | \dots | \mathbf{x}_p)$ with corresponding response vector $\mathbf{y} = (y_1, \dots, y_n)^T$. Componentwise boosting means to update one parameter within one boosting step. Therefore one fits the model $\boldsymbol{\mu} = h_0(h(\mathbf{X}\widehat{\boldsymbol{\beta}}^{(l-1)} + \mathbf{x}_j\beta_j))$, where $\mathbf{X}\widehat{\boldsymbol{\beta}}^{(l-1)}$ is a fixed offset and only the variable \mathbf{x}_j is included in the model. Then penalized Fisher scoring for parameter β_j has the form

$$\widehat{\beta}_j^{(l)} = \nu_p \left(\mathbf{x}_j^T \widehat{\mathbf{D}}_\eta^{(l-1)} (\widehat{\boldsymbol{\Sigma}}^{(l-1)})^{-1} \widehat{\mathbf{D}}_\eta^{(l-1)} \mathbf{x}_j \right)^{-1} \mathbf{x}_j^T \widehat{\mathbf{D}}_\eta^{(l-1)} (\widehat{\boldsymbol{\Sigma}}^{(l-1)})^{-1} (\mathbf{y} - \widehat{\boldsymbol{\mu}}^{(l-1)}), \quad (6)$$

where $\nu_p = 0.1$ and

$$\begin{aligned} \widehat{\mathbf{D}}_\eta^{(l-1)} &= \text{diag} \left\{ \frac{\partial h_0(\widehat{h}^{(l-1)}(\widehat{\eta}_i^{(l-1)}))}{\partial \eta} \right\}_{i=1}^n \\ &= \text{diag} \left\{ \frac{\partial h_0(\widehat{h}^{(l-1)}(\widehat{\eta}_i^{(l-1)}))}{\partial h^{(l-1)}(\eta)} \cdot \frac{\partial \widehat{h}^{(l-1)}(\widehat{\eta}_i^{(l-1)})}{\partial \eta} \right\}_{i=1}^n \end{aligned} \quad (7)$$

is the matrix of derivatives evaluated at the values of the previous iteration and

$$\widehat{\boldsymbol{\Sigma}}^{(l-1)} = \text{diag} \left\{ \sigma^2(h_0(\widehat{h}(\widehat{\eta}_i^{(l-1)}))) \right\}_{i=1}^n \quad (8)$$

is the variance from the previous step.

The basic algorithm given in the following computes updates of the parameter vector and the coefficients of the basic functions. In each step it is decided which update is best and only one is executed. Thus in each step either the parameter vector or the coefficients of the basic functions are refitted.

Algorithm: FlexLink

Step 1 (Initialization)

Set $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, $\widehat{\boldsymbol{\eta}}^{(0)} = \mathbf{0}$. Compute $\widehat{\mathbf{D}}^{(0)} = \text{diag} \left\{ h_0(\widehat{\eta}_i^{(0)}) \right\}_{i=1}^n$, $(\widehat{\boldsymbol{\Sigma}}^{(0)}) = \text{diag} \left\{ \sigma^2(\widehat{\eta}_i^{(0)}) \right\}_{i=1}^n$ and determine $\boldsymbol{\alpha}^{(0)}$ as described previously.

Step 2 (Iteration)

For $l = 1, 2, \dots, M$

1. *Predictor update*

- Compute for every $j \in \{1, \dots, p\}$ the penalized estimate $\widehat{\boldsymbol{\beta}}_j^{(l)} = (0, \dots, \widehat{b}_j^{(l)}, \dots, 0)$ based on one-step Fisher scoring (6) and determine the candidate update

$$\boldsymbol{\beta}_j^{(l)} = \widehat{\boldsymbol{\beta}}^{(l-1)} + \widehat{\boldsymbol{b}}_j^{(l)}.$$

- Compute $\widehat{\boldsymbol{\beta}}_j^{(l)} = \boldsymbol{\beta}_j^{(l)} / \|\boldsymbol{\beta}_j^{(l)}\|$ and the corresponding negative log-likelihood function $-l(\mathbf{y}, h_0(\widehat{h}^{(l-1)}(\mathbf{X}\widehat{\boldsymbol{\beta}}_j^{(l)})))$.
- Choose the parameter vector $\widehat{\boldsymbol{\beta}}_{opt}^{(l)}$ which minimizes the negative log-likelihood function and set $\widehat{\boldsymbol{\beta}}^{(l)} = \widehat{\boldsymbol{\beta}}_{opt}^{(l)}$

2. Response function update

- Compute $\widehat{\boldsymbol{\alpha}}^{(l)}$ as described in (3) and set $\widehat{\boldsymbol{\alpha}}^{(l)} = \widehat{\boldsymbol{\alpha}}^{(l-1)} + \widehat{\mathbf{a}}^{(l)}$
- Compute $\widehat{h}^{(l)}(\boldsymbol{\eta}^{(l-1)}) = \Phi\widehat{\boldsymbol{\alpha}}^{(l)}$ and $l_{\boldsymbol{\alpha}} = -l(\mathbf{y}, h_0(\widehat{h}^{(l)}(\boldsymbol{\eta}^{(l-1)})))$.

3. Update choice

- If $l_{\boldsymbol{\alpha}} < l_{\boldsymbol{\beta}}$ then $\boldsymbol{\alpha}^{(l)}$ is updated and $\widehat{\boldsymbol{\beta}}$ remains unchanged, $\widehat{\boldsymbol{\beta}}^{(l)} = \widehat{\boldsymbol{\beta}}^{(l-1)}$.
- If $l_{\boldsymbol{\alpha}} \geq l_{\boldsymbol{\beta}}$ then $\widehat{\boldsymbol{\beta}}^{(l)}$ is updated and $\widehat{\boldsymbol{\alpha}}$ remains unchanged, $\widehat{\boldsymbol{\alpha}}^{(l)} = \widehat{\boldsymbol{\alpha}}^{(l-1)}$.

Further Details

If the transformation $h_T(\cdot)$ in the generating model is considered as response function it has to be monotone. The approximating transformation is given by $h_0(\widehat{h}(\cdot))$ where the outer function $h_0(\cdot)$ is already a monotonically increasing link function. In order to obtain a monotonically increasing response function $h_0(\widehat{h}(\cdot))$ we have to restrict the estimation of $\widehat{h}(\cdot)$ by a monotonicity constraint.

A sufficient condition for the B-Spline basis expansion to be monotonically increasing is that the components of the coefficient vector $\boldsymbol{\alpha}$ are ordered such that $\alpha_i \leq \alpha_{i+1}$ holds. In boosting methods this inequation must hold after every update step, $\alpha_i^{(l-1)} + a_i^{(l)} \leq \alpha_{i+1}^{(l-1)} + a_{i+1}^{(l)}$. Therefore we constrain every update step $\widehat{\boldsymbol{\alpha}}$ to be from

$$\mathcal{A} = \{\widehat{\boldsymbol{\alpha}}^{(l)} : \widehat{a}_2^{(l)} - \widehat{a}_1^{(l)} \geq \widehat{a}_1^{(l-1)} - \widehat{a}_2^{(l-1)}, \dots, \widehat{a}_k^{(l)} - \widehat{a}_{k-1}^{(l)} \geq \widehat{a}_{k-1}^{(l-1)} - \widehat{a}_k^{(l-1)}\}. \quad (9)$$

Monotone functions can be obtained in several ways. After computing $\widehat{\boldsymbol{\alpha}}^{(l)}$ in the l th step one can monotonize the components by use of isotone regression, provided for example by the R-routine `isoreg`. Alternatively, one can solve the

optimization problem that is behind the Fisher step in (3) with the additional restriction that $\hat{\mathbf{a}}$ is from \mathcal{A} . Therefore one minimizes

$$\mathbf{a}^T \Phi^T \widehat{\mathbf{W}} \Phi \mathbf{a} - 2 \Phi^T \widehat{\mathbf{W}} \left(\widehat{\mathbf{D}}^{(l-1)} \right)^{-1} (\mathbf{y} - \widehat{\boldsymbol{\mu}}^{(l-1)}), \text{ s.t. } \mathcal{A}$$

where $\widehat{\mathbf{W}} = \widehat{\mathbf{D}}^{(l-1)} (\widehat{\boldsymbol{\Sigma}}^{(l-1)})^{-1} \widehat{\mathbf{D}}^{(l-1)}$. Solutions can be obtained by use of the R-package `quadprog` (see Turlach (2009)) which is able to solve a quadratic optimization problem with linear constraints. Results are very similar. In our applications we use `quadprog`. For the use of similar constraints see also Gertheiss et al. (2009).

In step 3 of the algorithm a selection step is included in which it is determined if the coefficients of parameters or the link function is updated. We tried several alternatives but updating one of the sets of coefficients turned out to be most efficient.

Choice of Tuning Parameter

There are two tuning parameter in the model: the number of boosting iterations m which mainly steers variable selection and λ_h which controls the smoothness of the response function and the number of response function updates. For determining the appropriate tuple of tuning parameters $\boldsymbol{\pi} = (m, \lambda_h)$ we use K-fold cross validation (CV). There are several reasons to use this procedure and not to work with information-based criteria as used for example by Tutz and Leitnerstorfer (2009). On the one hand Hastie (2007) suggests to use CV in boosting procedures because the effective degrees of freedom can be underestimated by using the trace of the hat-matrix. On the other hand the trace of the hat matrix does not capture the complexity of a SIM because the complexity must be measured on two stages, first the complexity of the predictor, i.e. the number of influential covariables, and second the rawness of the estimated response function, i.e. the trace of its hat-matrix. In addition, the two restrictions (monotonicity of the response function and normalization of $\boldsymbol{\beta}$) make the problem of finding appropriate hat matrices more difficult.

In K-fold cross validation the data set is splitted K -times into a test data set of size n/K and a training data set of size $n - n/K$. For every tuple of tuning parameters $\boldsymbol{\pi}$ the model is fitted on the κ -th training data set obtaining $\hat{\boldsymbol{\gamma}}_{\kappa}^{\boldsymbol{\pi}} = (\hat{\boldsymbol{\alpha}}_{\kappa}^{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}_{\kappa}^{\boldsymbol{\pi}})$. Then the deviance on the κ -th test set $\text{Dev}(\mathbf{y}_{\text{test}}^{\kappa}, \hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\gamma}}_{\kappa}^{\boldsymbol{\pi}} | \mathbf{X}_{\text{test}}^{\kappa}, \boldsymbol{\pi}))$ is computed. The final $\boldsymbol{\pi}_{\text{opt}}$ is determined by

$$\boldsymbol{\pi}_{\text{opt}} = \underset{\boldsymbol{\pi} \in \mathcal{M} \times \Lambda}{\text{argmin}} \left\{ \sum_{\kappa=1}^K \text{Dev}(\mathbf{y}_{\text{test}}^{\kappa}, \hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\gamma}}_{\kappa}^{\boldsymbol{\pi}} | \mathbf{X}_{\text{test}}^{\kappa}, \boldsymbol{\pi})) \right\} \quad (10)$$

In the following we choose $m \in \mathcal{M} := \{1, \dots, 1000\}$ and $\lambda_h \in \Lambda := \{0.01, 0.1, 1, 10, 100\}$.

3 Simulation Studies

Measures of Model Assessment

Some care should be taken when estimates are compared. We assume $\mu_i = h_T(\mathbf{x}_i^T \boldsymbol{\beta})$ where $h_T(\cdot)$ is the unknown true transformation function and for the true parameter (without intercept) $\|\boldsymbol{\beta}\| = 1$ holds and the magnitude of $\|\boldsymbol{\beta}\|$ as well as the intercept are absorbed into $h_T(\cdot)$. Let the generating model without restrictions be given by $\mu_i = h_G(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}_0)$ with unrestricted parameter vector $\boldsymbol{\beta}_0$, where h_G is any monotone function. Then the model can always be rewritten in the corresponding standardized true response function $h_T(\cdot)$ by

$$\mu_i = h_G(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}_0) = h_G(\beta_0 + \|\boldsymbol{\beta}_0\|(\mathbf{x}_i^T \boldsymbol{\beta}_0 / \|\boldsymbol{\beta}_0\|)) = h_T(\eta),$$

with $\eta = \beta_0 + \|\boldsymbol{\beta}_0\|\eta$, $\eta = \mathbf{x}_i^T \boldsymbol{\beta}$, $\boldsymbol{\beta} = \boldsymbol{\beta}_0 / \|\boldsymbol{\beta}_0\|$. In particular when a given link function like the canonical link is used, estimates cannot be compared directly to the parameters $\|\boldsymbol{\beta}_0\|$ for some generating link function $h_G(\cdot)$. Therefore estimated parameters are also standardized and one considers $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{can} / \|\hat{\boldsymbol{\beta}}_{can}\|$, where $\boldsymbol{\beta}_{can}$ is the estimate resulting from the canonical link model.

Comparisons in this article always refer to corresponding standardized estimates $\hat{\boldsymbol{\beta}}$. Therefore the difference between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ is measured by

$$MSE_{\boldsymbol{\beta}} = \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2,$$

where $\|\boldsymbol{\beta}\| = 1$ and $\|\hat{\boldsymbol{\beta}}\| = 1$. In addition, the accuracy of prediction is investigated by use of the predictive deviances on an independent test set

$$\text{Dev}(\text{test}) = \text{Dev}(\mathbf{y}_{\text{test}}, \hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\gamma}}^{\boldsymbol{\pi}_{opt}} | \mathbf{X}_{\text{test}}, \boldsymbol{\pi}_{opt})).$$

The number of observations in the test data set is chosen by $n_{\text{test}} = 5 \cdot n_{\text{train}}$.

Procedures and Results

We compare our procedure with three procedures:

- The boosting procedure `mboost` with canonical link function (see Hothorn, Bühlmann, Kneib, Schmid, and Hofner (2009)).
- L1 penalized GLM `lasso` with canonical link function (see Lokhorst, Venables, Turlach, and Maechler (2007) for Poisson and Friedman, Hastie, and Tibshirani (2008) for the binary case).
- The ML-estimator with canonical link function for the full model.

Further we present a modified version of FlexLink which truncates small coefficients to zero. This modification of the FlexLink is explained in Section 4.

The predictor matrix was generated as a $N(\mathbf{0}_p, \mathbf{\Sigma})$ -distribution with $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j \in \{1, \dots, p\}}$ where $\sigma_{ij} = 0.5$, $i \neq j$, $\sigma_{ii} = 1$. We use two parameter vectors with $p = 20$,

$$\begin{aligned}\boldsymbol{\beta}_a &= (0.2, 0.4, -0.4, 0.8, 0, \dots, 0)^T, \\ \boldsymbol{\beta}_b &= (0.5, 0.5, -0.5, -0.5, 0, \dots, 0)^T,\end{aligned}$$

to generate $\eta = \mathbf{X}\boldsymbol{\beta}$. As distributions of the response we consider normal, Poisson and binomial distribution. Further we consider two different response functions for every distribution. So 12 different simulation settings were investigated. They are denoted in the following way, $\langle \text{dis} \rangle \langle \text{resp} \rangle \langle \text{beta} \rangle$. For example, the setting Bin2b has binomial distributed response, uses the second response function and $\boldsymbol{\beta}_b$ is the true parameter vector. The true response functions that are used in the following, a approximation by the canonical link of it and the 50 estimated response function are shown in Figure 3. The estimated response function are for the case $\boldsymbol{\beta}_a$.

(1) *Normal Distribution*

In the Normal case we use the response functions

1. $h_T(\eta) = 3 \cdot \eta^3$
2. $h_T(\eta) = \text{sgn}(\eta)5 \cdot \sqrt[3]{\eta}$

which are shown in the first row of Figure 3. In addition an approximation of $h_T(\cdot)$ to the canonical response function, which in this case is linear, is shown. Therefore, $h_{can}(\eta) = a + b \cdot \eta$ is computed where a and b are chosen to minimize $\int_{-2}^2 (h(\eta) - h_{can}(\eta))^2 d\eta$. The approximation is shown by the grey line. For the first first response function the error term is $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, 9\mathbf{I})$ and for the second $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I})$. The results of the simulations are shown in Figure 4 and summarized in Table 1. We included a modified version of Flex Link, called Flex Link (cut) which will be introduced in the next section. Performance in terms of MSE and predictive deviance is about the same as for Flex Link. Comparison to the other procedures favours Flex Link which distinctly outperforms LASSO and mboost in all settings.

(2) *Poisson Distribution*

In the Poisson case we consider the response functions:

1. $h_T(\eta) = \frac{10}{1 + \exp(-5 \cdot \eta)}$
2. $h_T(\eta) = \frac{10}{1 + \exp(-10 \cdot \eta - 10)} + \frac{10}{1 + \exp(-10 \cdot \eta + 10)}$

They are shown in the second row from Figure 3. Also the approximation of $h_T(\cdot)$ by the canonical response function is given. The results of the simulations are

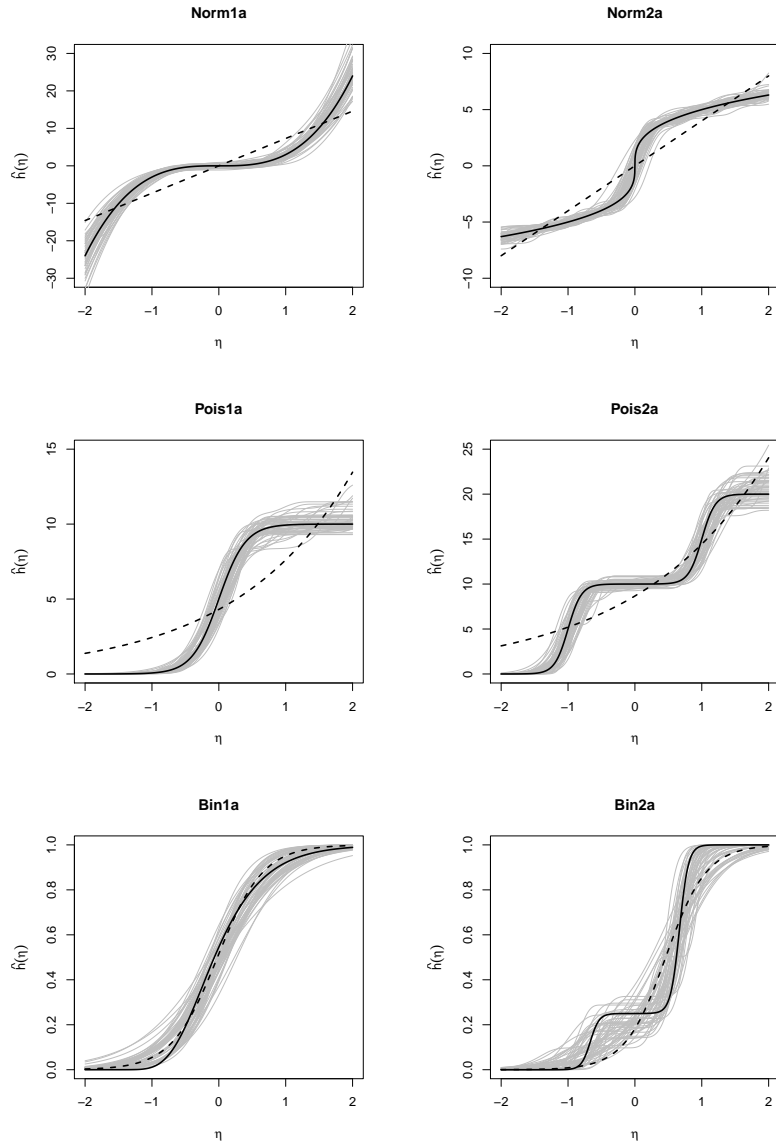


FIGURE 3: True response functions (black lines), approximating canonical response functions (dashed lines) and estimated response functions (grey) of simulation study.

shown in Figure 5 and summarized in Table 1. Flex Link outperforms LASSO and mboost even more distinctly than in the normal distribution case.

(3) Binomial Distribution

For binomial responses the true response functions are

1. $h_T(\eta) = \exp(-\exp(-2 \cdot \eta - 0.5))$,

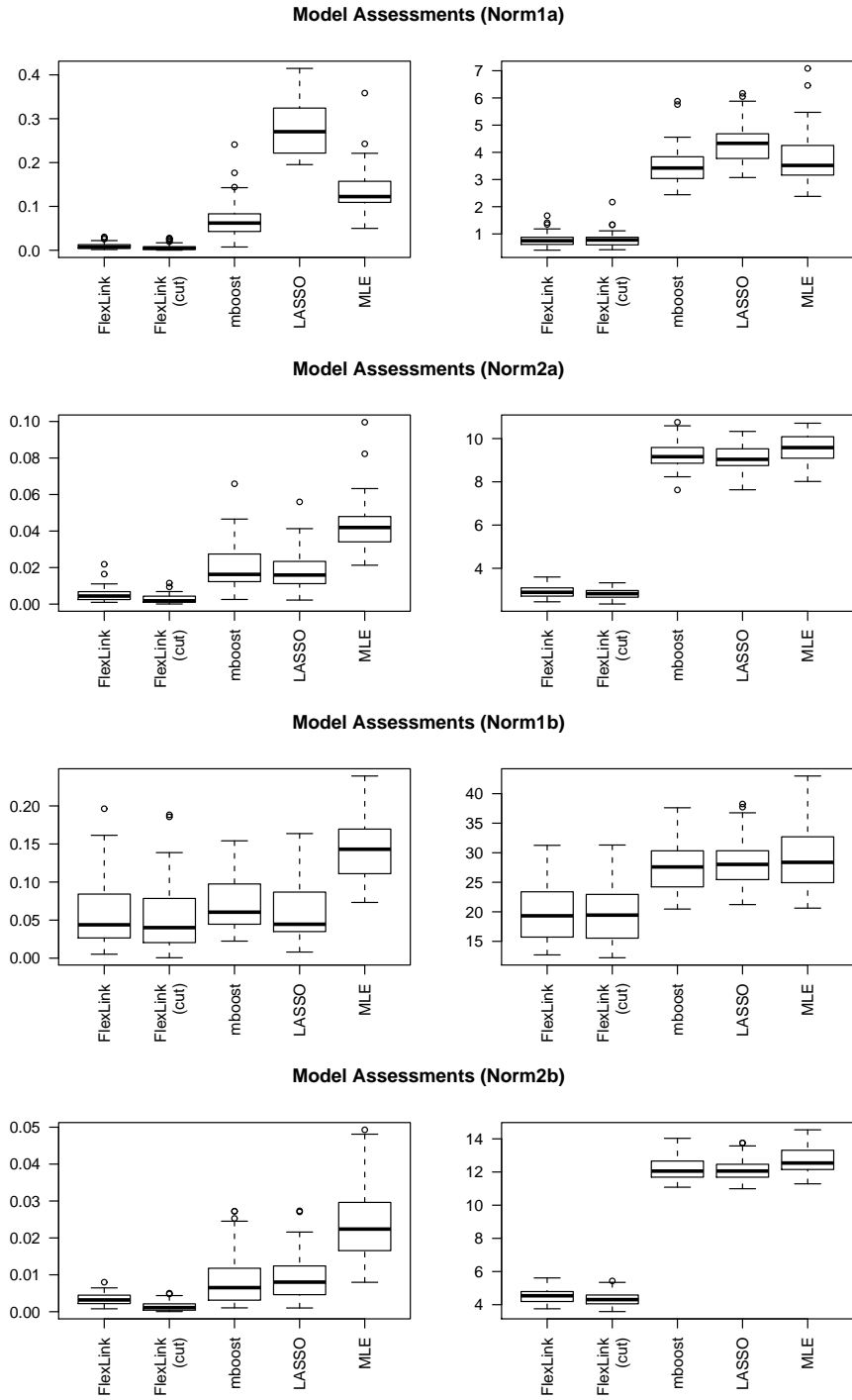


FIGURE 4: Boxplots of model assessment measurements MSE_{β} (left) and Dev_{test} (right) in the normal case.

$$2. h_T(\eta) = \frac{0.25}{1 + \exp(-15 \cdot \eta - 10)} + \frac{0.75}{1 + \exp(-15 \cdot \eta + 10)}.$$

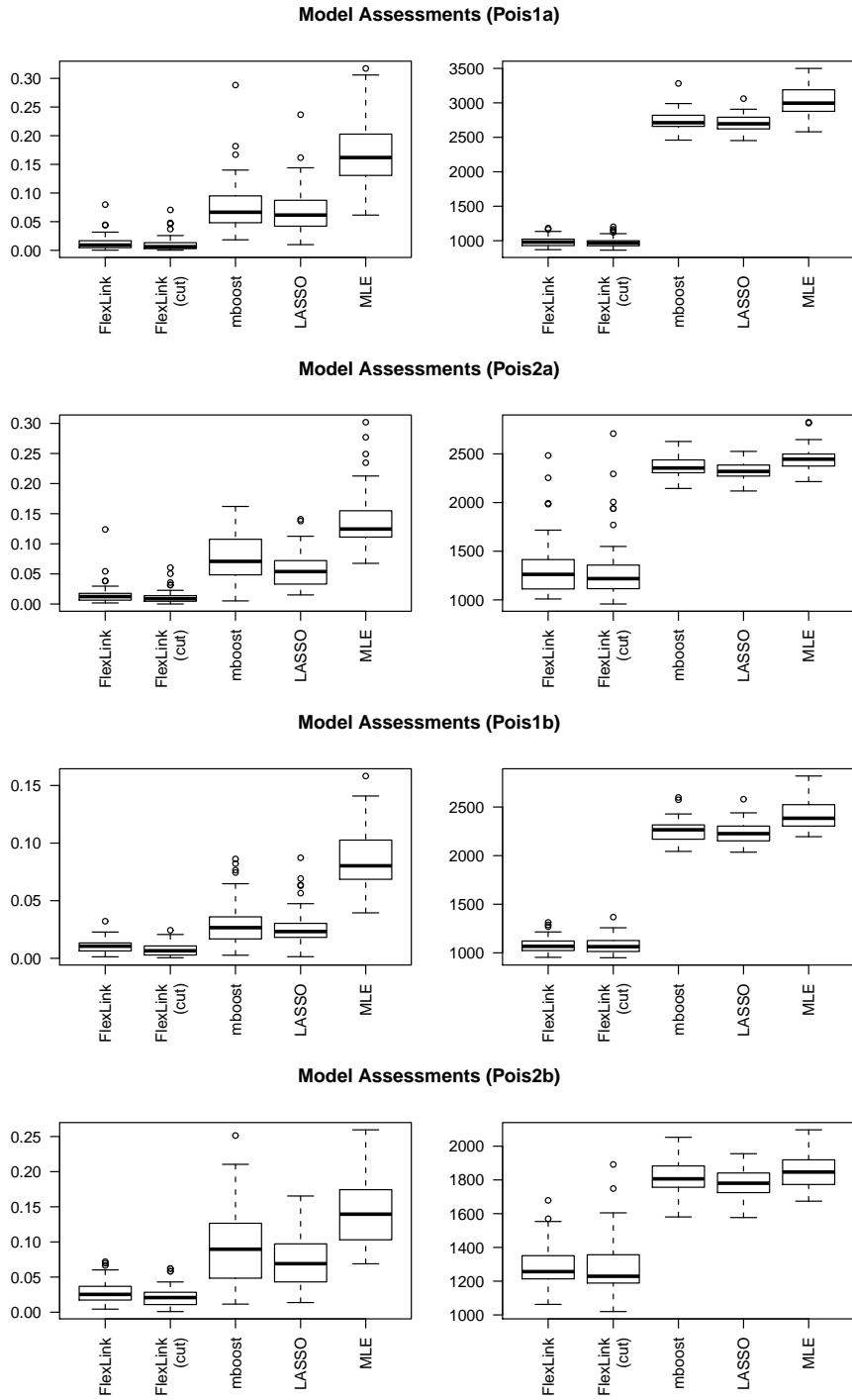


FIGURE 5: Boxplots of model assessment measurements MSE_{β} (left) and Dev_{test} (right) in the Poisson case.

Figure 3 shows the response function and the approximating canonical response function. The second response function corresponds to the Gumbel-link and can

be approximated by the canonical logit-link quite well. The binomial case is a challenge for the estimation of the unknown link function because the information in 0-1 observations is weak and the true link function does not differ so much from the canonical link. It is seen that there is not so much difference among the selection procedures but all yield better results than MLE. We found that increasing the number of boosting iterations beyond the optimal number m_{opt} yields quite good approximation of the link function but selection of relevant predictors suffers.

		FlexLink	FlexLink (cut)	mboost	LASSO	MLE
Normal distribution						
Norm1a	MSE_{β}	0.0072	0.0046	0.0620	0.2705	0.1224
	Dev(test)	0.7493	0.7778	3.4226	4.3310	3.5184
Norm2a	MSE_{β}	0.0044	0.0018	0.0163	0.0159	0.0419
	Dev(test)	2.8895	2.8265	9.1622	9.0386	9.5827
Norm1b	MSE_{β}	0.0439	0.0401	0.0604	0.0446	0.1431
	Dev(test)	19.3224	19.4391	27.5937	28.0246	28.3730
Norm2b	MSE_{β}	0.0032	0.0011	0.0065	0.0080	0.0224
	Dev(test)	4.5385	4.3072	12.0554	12.0587	12.5424
Poisson distribution						
Pois1a	MSE_{β}	0.0092	0.0063	0.0664	0.0615	0.1619
	Dev(test)	979.60	966.41	2711.83	2696.37	2995.60
Pois2a	MSE_{β}	0.0123	0.0088	0.0708	0.0539	0.1246
	Dev(test)	1262.12	1218.33	2354.94	2320.07	2445.86
Pois1b	MSE_{β}	0.0105	0.0065	0.0266	0.0232	0.0803
	Dev(test)	1067.16	1063.71	2265.70	2226.17	2384.33
Pois2b	MSE_{β}	0.0253	0.0208	0.0896	0.0691	0.1395
	Dev(test)	1256.51	1229.08	1806.42	1780.39	1846.76
Binomial distribution						
Bin1a	MSE_{β}	0.0761	0.0804	0.0797	0.0843	0.1798
	Dev(test)	813.30	809.57	802.25	796.64	886.73
Bin2a	MSE_{β}	0.0732	0.0734	0.0905	0.0800	0.2197
	Dev(test)	760.35	761.97	818.05	789.27	1336.62
Bin1b	MSE_{β}	0.0836	0.0788	0.0610	0.0719	0.1515
	Dev(test)	981.58	982.99	967.42	979.92	1070.33
Bin2b	MSE_{β}	0.0904	0.0948	0.0939	0.0930	0.1938
	Dev(test)	988.73	998.56	1029.98	986.38	1411.32

TABLE 1: Medians of the model assessment measures for the settings of the simulation study.

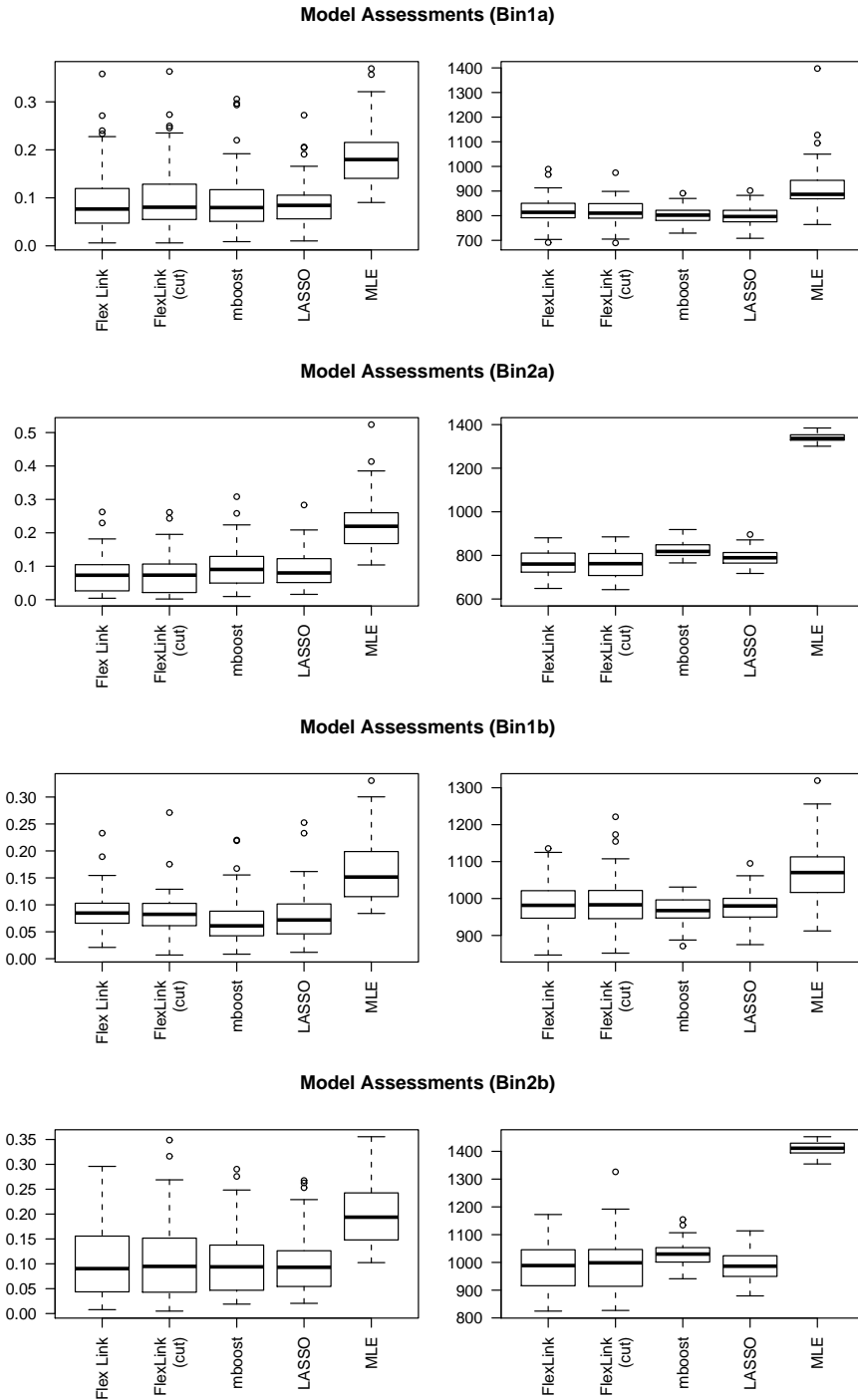


FIGURE 6: Boxplots of model assessment measurements MSE_{β} (left) and Dev_{test} (right) in the binomial case.

4 Modified Estimator and Selection of Predictors

MSE and predictive deviance are important criteria in the comparison of fitting procedures. However, in selection procedures the performance should also refer

to the precision of the selection. Criteria by which selection can be measured are in particular hit rate (proportion of correctly identified influential variables) and false positives (proportion of non-influential variables dubbed influential).

One problem with simple boosting procedures is that some predictors are selected just once or twice. The corresponding estimated parameters are very small but are unequal zero. Performance of selection can be easily improved by cutting off these small values. In the procedure called Flex Link (cut) we use a truncated version of $\hat{\beta}$. The components of estimate $\hat{\beta}$ are compared to $1/p$, where p is the number of predictors, and components that are smaller than $1/p$ are set to zero. Then the new estimate is re-standardized to have Euclidian norm 1. When used in the cross-validation procedure (10) one obtains the new optimal tuning parameter $\tilde{\pi}_{opt}$.

As is seen from Table 2, which gives the means of hits and false positive rates for all settings, that the truncated version of Flex Link shows distinct improvement. False positive rates are much smaller, hit rates are in most cases the same as in simple FlexLink, or slightly smaller. Comparison to mboost and LASSO are strongly in favour of FlexLink. The effect is illustrated in Figure 7 where hits and false positive rates for one setting are plotted in a ROC-type way. The best performance would be the point (false positive rate, hit rate)=(0,1). Among the considered procedures FlexLink (cut) shows the best approximation to the optimal point.

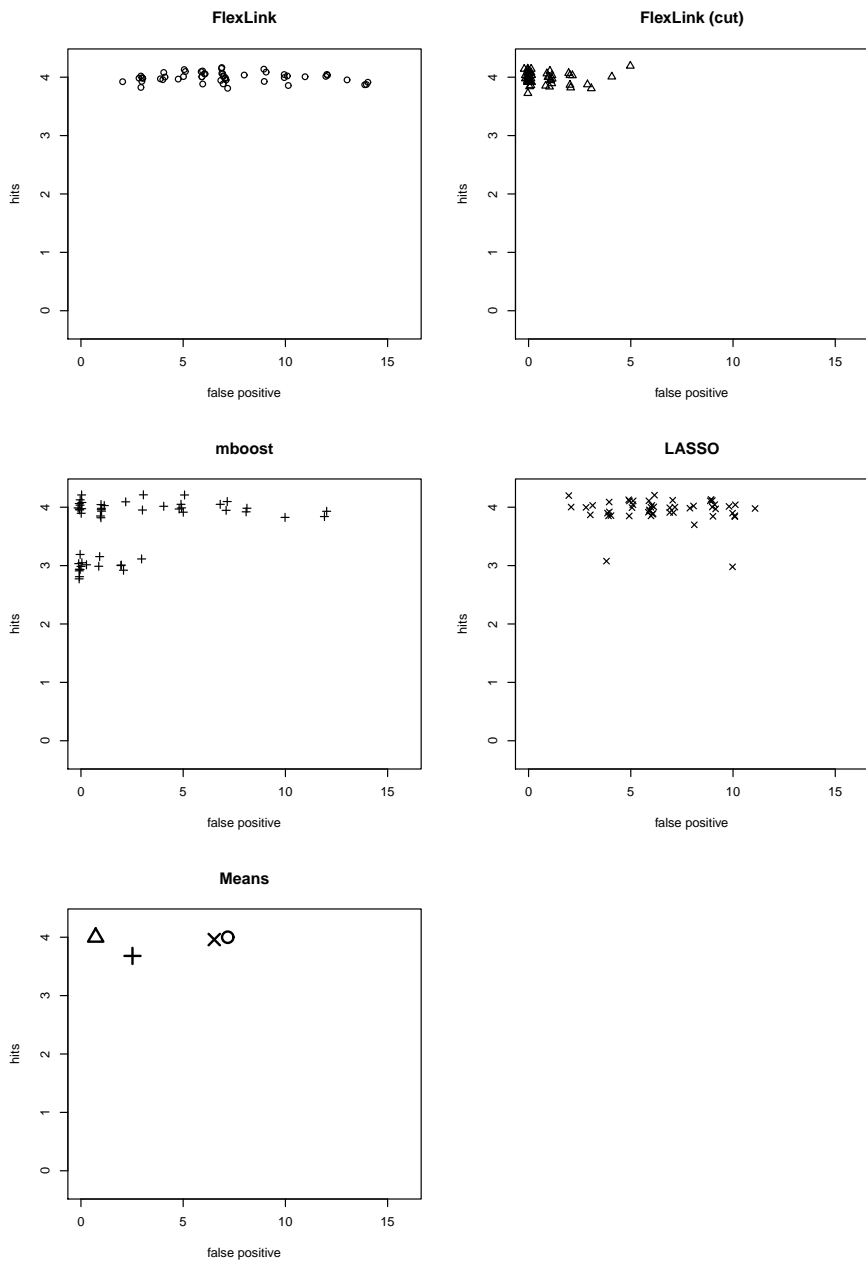


FIGURE 7: Hits and false positive rates for setting *Pois2a* with jittered values. Last panel shows the means over simulations.

	FlexLink		FlexLink (cut)		mboost		LASSO	
	hits	false pos.	hits	false pos.	hits	false pos.	hits	false pos.
Normal case								
<i>Norm1a</i>	1.000	0.504	1.000	0.015	0.980	0.355	0.530	0.004
<i>Norm2a</i>	1.000	0.445	1.000	0.006	1.000	0.369	1.000	0.406
<i>Norm1b</i>	1.000	0.575	1.000	0.158	1.000	0.376	1.000	0.165
<i>Norm2b</i>	1.000	0.504	1.000	0.000	1.000	0.355	1.000	0.431
Poisson case								
<i>Pois1a</i>	0.995	0.370	0.990	0.038	0.965	0.265	0.980	0.326
<i>Pois2a</i>	1.000	0.449	1.000	0.045	0.920	0.158	0.990	0.408
<i>Pois1b</i>	1.000	0.439	1.000	0.025	1.000	0.283	1.000	0.309
<i>Pois2b</i>	1.000	0.665	1.000	0.104	1.000	0.223	1.000	0.456
Binomial case								
<i>Bin1a</i>	0.915	0.244	0.870	0.100	0.960	0.366	0.975	0.409
<i>Bin2a</i>	0.980	0.306	0.955	0.133	0.975	0.390	0.985	0.430
<i>Bin1b</i>	1.000	0.304	1.000	0.128	1.000	0.401	1.000	0.451
<i>Bin2b</i>	1.000	0.394	1.000	0.160	1.000	0.405	0.446	1.000

TABLE 2: Means of the hits and false positive rates.

5 Applications

5.1 Medical Care Data

In this section we consider the health care data from Dep and Trivedi (1997). The original data is from the US National Medical Expenditure Survey and is available from the data archive of the Journal of Applied Econometrics (<http://www.econ.queensu.ca/jae/1997-v12.3/deb-trevidi/>). We use the `data.frame` from Zeileis (2006). The response variable that is considered is the *number of physician office visits (ofp)*, which potentially depends on the variables given in Table (3). In our investigation we use only male patients, which reduces the sample size to $n = 1778$ from the total available sample of 4406 individuals.

We compare the same estimating procedures as in Section 3. For measuring the prediction performance 25 splits into a training data set of $n_{train} = 1185$ and a test data set of $n_{test} = 593$ were used. The tuning parameter λ_h of the FlexLink is determined by 5-fold cross validation to $\lambda_h = 100$. Figure 8 shows the predictive deviances in the test data and the fitted link functions. For illustration we give the the estimated response functions in Figure 10 (for male patients which visit physician office maximum 30 times). It is seen that the link function for the

flexible model differs from the canonical link in particular for large values of the linear predictor. While the canonical link still increases distinctly the flexible link is very flat. The estimated link functions are very stable across splits. It is seen from Figure 8 that prediction for the flexible model with variable selection distinctly outperforms the competitors. From Table 4 it is seen that the flexible link procedure reduces the number of coefficients.

Label	Explanation
<code>exclhlth</code>	= 1 if self-perceived health is excellent
<code>poorhlth</code>	= 1 if self-perceived health is poor
<code>numchron</code>	number of chronic conditions (cancer, heart attack, gall bladder problems, emphysema , arthritis, diabetes, other heart disease)
<code>adldiff</code>	= 1 if the person has a condition that limits activities of daily living
<code>noreast</code>	= 1 if the person lives in northeastern US
<code>midwest</code>	= 1 if the person lives in the midwestern US
<code>west</code>	= 1 if the person lives in the western US
<code>age</code>	age in years divided by 10
<code>black</code>	= 1 if the person is African American
<code>married</code>	= 1 if the person is married
<code>school</code>	number of years of education
<code>faminc</code>	family income in \$10 000
<code>employed</code>	= 1 if the person is employed
<code>privins</code>	= 1 if the person is covered by private health insurance
<code>medicaid</code>	= 1 if the person is covered by Medicaid

TABLE 3: *Variable description for health care data*

The estimated parameters are given in Table 4. Since data are strongly overdispersed ($\hat{\Phi} = 7.736$) we give quasi-likelihood estimates (QLE) instead of the maximumlikelihood estimates. It is seen that all covariates with p-values smaller than 0.05 for QLE were selected by FlexLink an FlexLink (cut). The latter procedures select two more covariates, covariate 7 and 10. In Figure 9 we show the error bars across 300 bootstrap samples. The circles mark the parameter estimate from Table 4 and the whiskers are the 0.975- and 0.025-quantiles determined by bootstrapping. We used simple pairwise bootstrap. The data contains $n = 1778$ pairs (y_i, \mathbf{x}_i) , $i = 1, \dots, n$, where y_i is the response value and \mathbf{x}_i is the corresponding vector of covariables. We sample $b = 300$ bootstrap samples. Each bootstrap sample is sampled by drawing n pairs (y_i, \mathbf{x}_i) with replacement. We achieve $(\mathbf{y}_b^*, \mathbf{X}_b^*)$, $b = 1, \dots, 300$, bootstrap samples with n observations whereby some observations are equal. Then we fit models on $(\mathbf{y}_b^*, \mathbf{X}_b^*)$, $b = 1, \dots, 300$, and

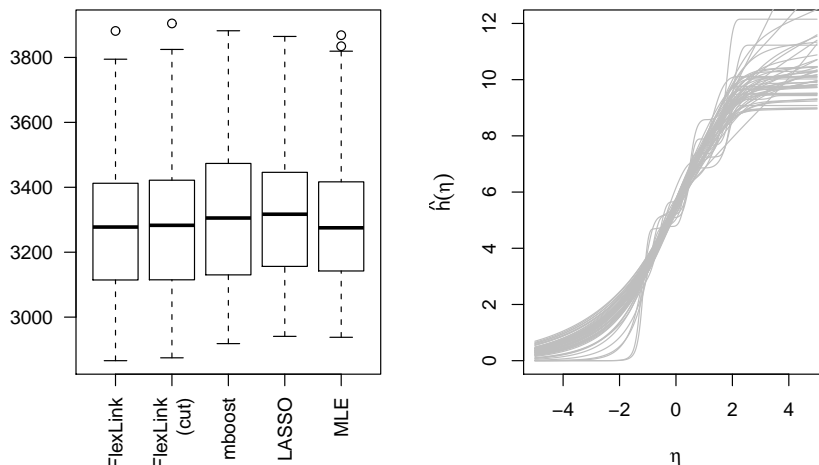


FIGURE 8: *Left panel: Boxplots of predictive deviance on test data sets across 50 random splits. Right panel: Estimated response function for health care data across 50 random splits.*

Number	Variable	FlexLink	FlexLink(cut)	mboost	LASSO	QLE (p-value)
1	exclhlth	0.0000	0.0000	-0.2023	-0.0382	-0.1979 (0.050)
2	poorhlth	0.2083	0.3614	0.3203	0.2546	0.3134 (0.000) *
3	numchron	0.9492	0.8849	0.7042	0.7971	0.6854 (0.000) *
4	adldiff	0.0000	0.0000	0.0250	0.0000	0.0258 (0.766)
5	noreast	0.0000	0.0000	-0.1623	-0.0916	-0.1318 (0.177)
6	midwest	0.0000	0.0000	0.0451	0.0000	0.0636 (0.483)
7	west	0.0419	0.0866	0.0725	0.0062	0.0840 (0.359)
8	age	0.0000	0.0000	-0.0592	0.0000	-0.0142 (0.875)
9	black	0.0000	0.0000	-0.1693	0.0000	-0.1276 (0.208)
10	married	0.0364	0.0878	0.1166	0.0734	0.1420 (0.119)
11	school	0.1725	0.1965	0.3949	0.4817	0.4224 (0.000)*
12	faminc	0.0000	0.0000	0.0000	0.0000	-0.0064 (0.939)
13	employed	0.0000	0.0000	0.0138	0.0000	0.0254 (0.772)
14	privins	0.1508	0.1801	0.3318	0.2292	0.3555 (0.001)*
15	medicaid	0.0000	0.0000	0.1196	0.0000	0.1462 (0.124)

TABLE 4: *Parameter estimates for medical care data set.*

achieve the corresponding estimates $\widehat{\beta}_b^*$. Finally we computed the quantiles of the distribution of the components of estimates $\widehat{\beta}_b^*$, $b = 1, \dots, 300$.

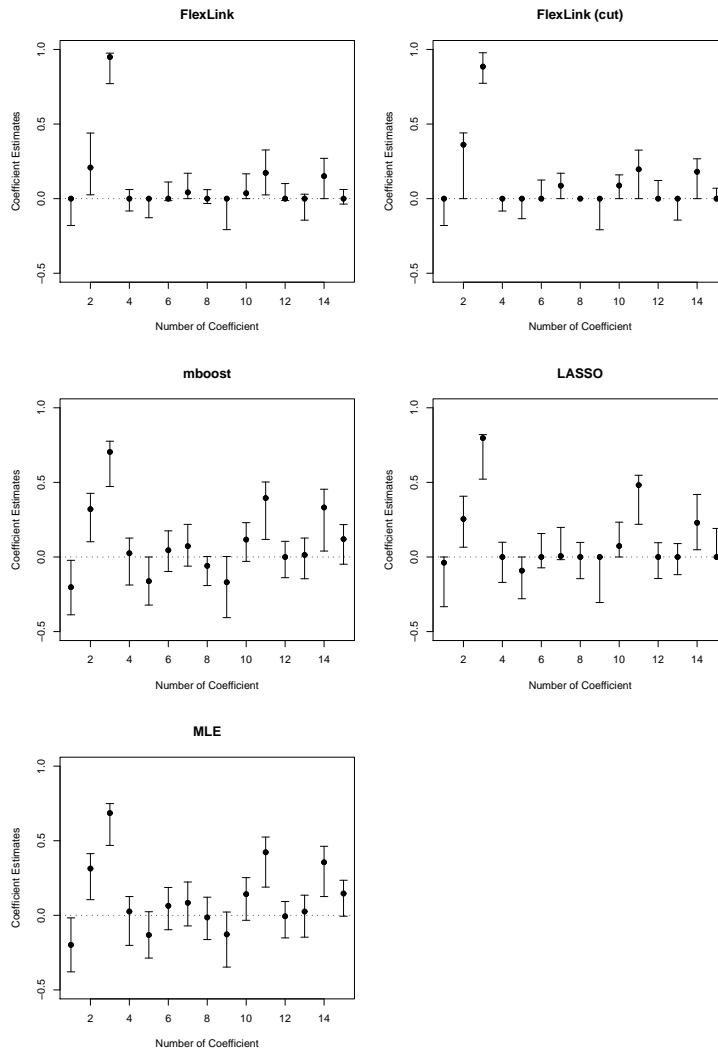


FIGURE 9: *The error bar plot across 300 bootstrap samples.*

It is remarkable that mboost selects nearly all variables. The LASSO and the FlexLink select a similar set of variables. Further by estimating the response function flexibly there is a tendency that the smaller values of y_i are accumulate on the left side what seems to be reasonable for an increasing response function. This effect can not be find for the other procedures.

5.2 Noisy Miner Data

In this section we consider the noisy miner data from Maron (2007), which are available at <http://www.sci.usq.edu.au/staff/dunn/>

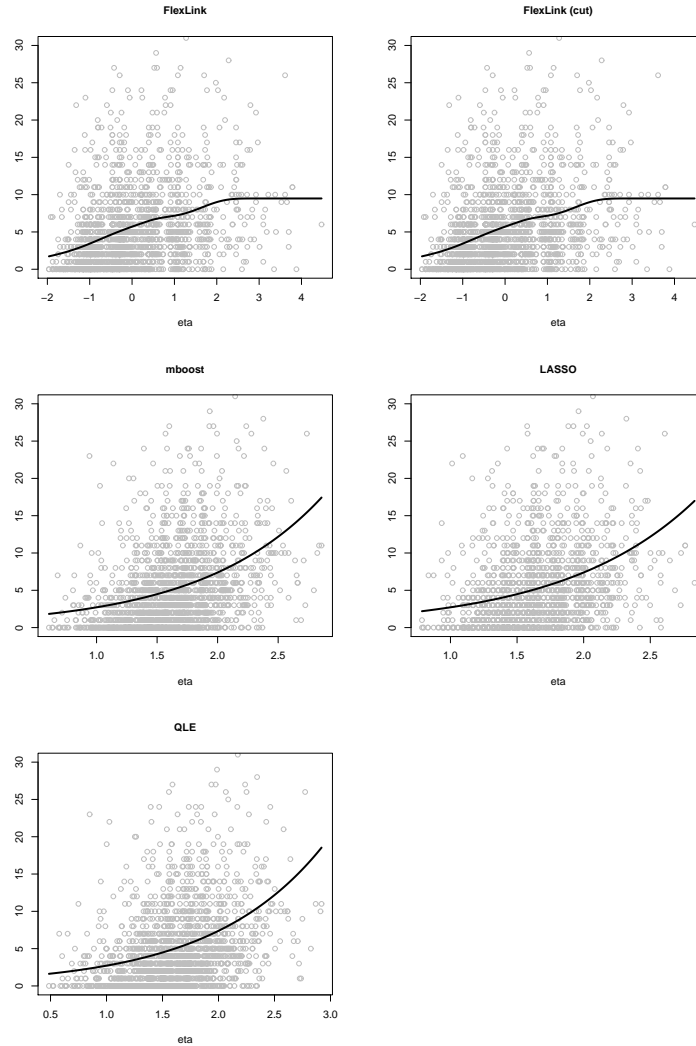


FIGURE 10: Response function of the four procedures with optimal tuning parameter determined by 5-fold cross validation and the QLE . Grey circles mark the observed response values y_i at the estimated value $\hat{\eta}_i$.

[Datasets/tech-glms.html](#). The data set has an biological background. Three 20 minutes surveys were conducted in each of 31 2ha belt transects in buloke woodland patches within the Wimmera Plains of western Victoria, Australia. The considered response is the number of species from a list of birds (*number of declining species*, in short *declinerab*). It is of particular interest how the number of species is determined by the presence of the noisy miner, which is an aggressive competitor. The collected explanatory variables are given in Table (5). Figure 11 shows the fitted response function together with the approximated canonical link function. It is seen that in particular for large

Number	Label	Explanation
1	eucs	number of eucalypts in each <i>2ha</i> transect
2	area	area [<i>ha</i>] of remnant patch vegetation in which the transect was located
3	grazed	whether the area was grazed (= 1) or not (= 0)
4	shrubs	whether shrubs were present (= 1) or not (= 0)
5	buloke	number of buloke trees in each transect
6	timber	number of pieces of fallen timber
7	finelitt	percentage of fallen litter on the ground
8	minerab	number of observed noisy miners

TABLE 5: *Variable description of the noisy miner data.*

values of the linear predictor the two link functions differ strongly; in that area the flexible response function is much steeper than the canonical response function. The prediction performance is measured by using 50 splits into a training data set of $n_{train} = 21$ and a test data set of $n_{test} = 10$. The tuning parameter λ_h of the FlexLink was determined by 5-fold cross validation. Figure 11 shows the predictive deviances in the test data. It is seen that prediction for the flexible model with variable selection distinctly outperforms the competitors. Table 6 shows parameter estimates for the various models. It turned out that Flex Link selects one variable, namely the number of noise miners, which seems to be responsible for the decrease in species. In contrast mboost selects five predictors and lasso three. Since the data are strongly overdispersed ($\hat{\Phi} = 4.647$) we used the quasi-likelihood estimator (QLE) instead of the MLE. QLE also suggests that only one variable in the linear predictor is relevant. Figure 12 shows the estimates for the 50 random splits by boxplots. It illustrates that the estimates are very stable for the flexible link procedure.

Number	Variable	FlexLink	FlexLink (cut)	mboost	LASSO	QLE (p-value)
1	eucs	0.0000	0.0000	0.0000	0.0000	0.0557 (0.264)
2	area	0.0000	0.0000	-0.0494	-0.0726	-0.0249 (0.168)
3	grazed	0.0000	0.0000	0.0000	0.0000	-0.5010 (0.430)
4	shrubs	0.0000	0.0000	0.1074	0.0000	-0.2569 (0.717)
5	buloke	0.0000	0.0000	0.0938	0.0505	0.0032 (0.345)
6	timber	0.0000	0.0000	0.0720	0.0000	0.0024 (0.896)
7	finelitt	0.0000	0.0000	0.0000	0.0000	0.0023 (0.910)
8	minerab	-1.0000	-1.0000	-0.9859	-0.9961	-0.8242 (0.001)*

TABLE 6: *Standardized parameter estimates for the whole noisy miner data set normed to length equal to 1.*

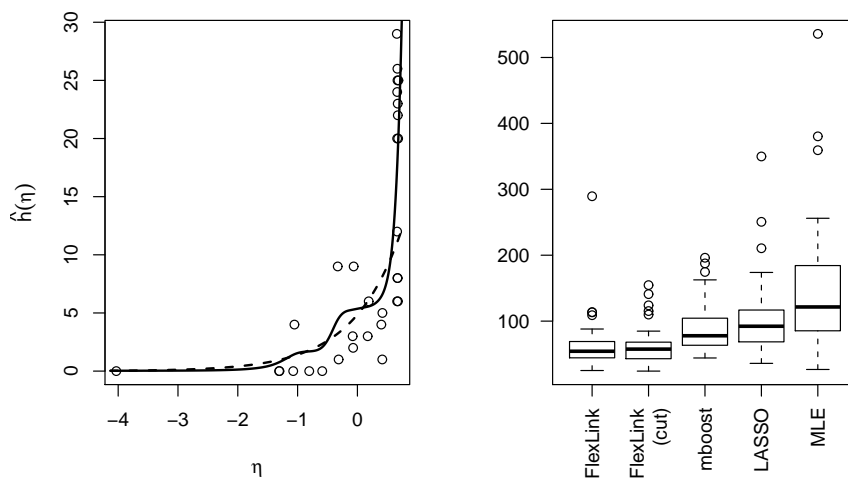


FIGURE 11: *Estimated response function for the noisy miner data and approximation by the canonical response function (left panel), box plots for deviances over 50 random splits (right panel)*

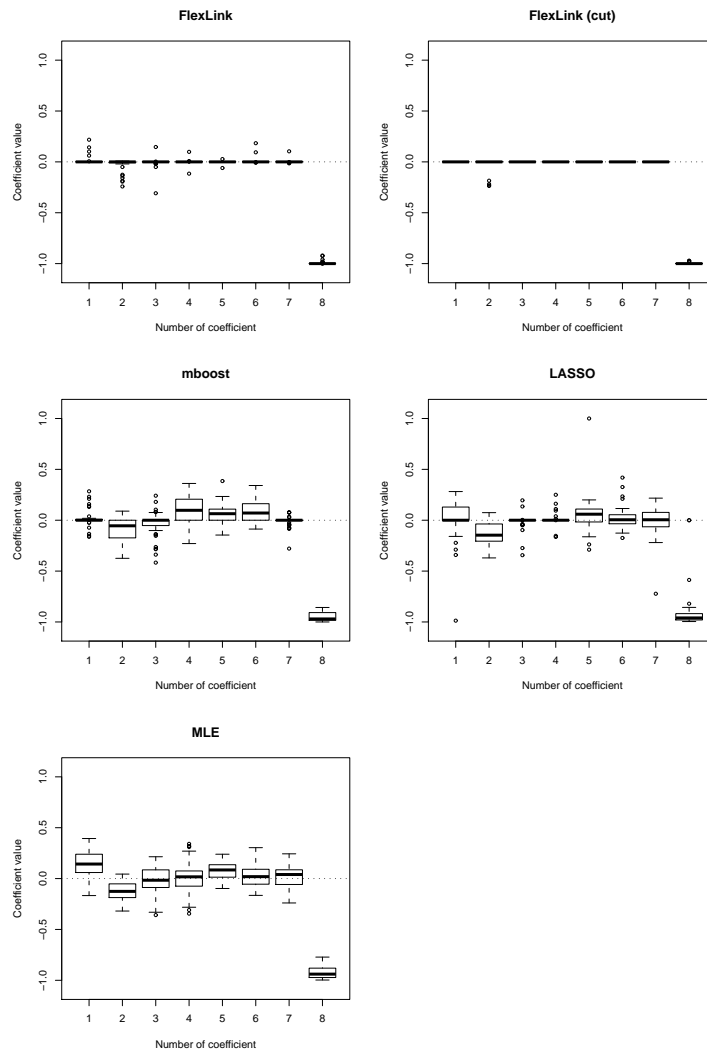


FIGURE 12: *The error bar plot across 300 bootstrap samples.*

6 Concluding Remarks

A flexible estimation of the response function combined with variable selection is proposed. It has been demonstrated that the method improves parameter estimation and prediction in the presence of irrelevant variables. The method works for generalized linear models, improvement is usually strong but less impressive for binary responses where information is weak. The modified version FlexLink (cut) shows much better variable selection performance without suffering in accuracy concerning estimation and prediction.

We focussed on the estimation of link functions for generalized linear models and therefore included a monotonicity restriction for the response function. In future work the monotonicity restriction shall be dropped resulting in generalized single-index models (compare Cui, Härdle, and Zhu (2009)). Then information-based criteria like AIC can be used since the hat matrix of boosting can be derived (for AIC in single-index models compare Naik and Tsai (2001)). The use of information-based criteria is attractive because it could reduce computational costs.

Acknowledgments

This work was partially supported by DFG Project TU62/4-1 (AOBJ: 548166).

References

- Antoniadis, A., G. Gregoire, and I. W. McKeague (2004). Bayesian estimation in single-index models. *Statistica Sinica* 14, 1147–1164.
- Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science* 22, 477–505.
- Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association* 98, 324–339.
- 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, 477–489.
- Cui, X., W. K. Härdle, and L. Zhu (2009). Generalized single index models: The efm approach. Discussion Paper 50, SFB 649, Humboldt University Berlin, Economic Risk.
- Czado, Y. and T. Santner (1992). The effect of link misspecification on binary regression inference. *Journal of statistical planning and inference* 33, 213–231.

- Dep, P. and P. K. Trivedi (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* 12, 313–336.
- Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Oxford Science Publications.
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics* 32, 407–499.
- Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and Penalties. *Statistical Science* 11, 89–121.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalize likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.
- Friedman, J., T. Hastie, and R. Tibshirani (2008). *glmnet: Lasso and elastic-net regularized generalized linear models*. R package version 1.1.
- Friedman, J. H. and W. Stützle (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* 76, 817–823.
- Gaiffas, S. and G. Lecue (2007). Optimal rates and adaptations in the single-index model using aggregation. *Electronic Journal of Statistics* 1, 538–573.
- Gertheiss, J., S. Hogger, C. Oberhauser, and G. Tutz (2009). Selection of ordinally scaled independent variables. Technical Report 62, Department of Statistics LMU Munich.
- Härdle, W., P. Hall, and H. Ichimura (1993). Optimal smoothing in single-index models.
- Hastie, T. (2007). Comment: Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* 22(4).
- Hothorn, T., P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner (2009). *mboost: Model-Based Boosting*. R package version 2.0-0.
- Hristache, M., A. Juditsky, and V. Spokoiny (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics* 29, 595–623.
- James, G. M. and P. Radchenko (2008). A Generalized Dantzig selector with Shrinkage Tuning. *Biometrika* , 127–142.
- Klein, R. L. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica* 61, 387–421.
- Lokhorst, J., B. Venables, B. Turlach, and M. Maechler (2007). *lasso2: L1 constrained estimation aka ‘lasso’*. R package version 1.2-6.
- Maron, M. (2007). Threshold effect of eucalypt density on an aggressive avian competitor. *Biological Conservation* 136, 100–107.

- Muggeo, V. M. R. and G. Ferrara (2008). Fitting generalized linear models with unspecified link function: A p-spline approach. *Computational Statistics & Data Analysis* 52(5).
- Naik, P. A. and C.-L. Tsai (2001). Single-index model selection. *Biometrika Trust* 88, 821–832.
- Park, M. Y. and T. Hastie (2006). An l1 regularization-path algorithm for generalized linear models. *Preprint, Department of Statistics, Stanford University*.
- Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–1430.
- Ruckstuhl, A. and A. Welsh (1999). Reference bands for nonparametrically estimated link functions. *Journal of Computational and Graphical Statistics* 8(4), 699–714.
- Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* 54, 1461–1481.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* 58, 267–288.
- Turlach, B. A. (2009). *quadprog: Functions to solve Quadratic Programming Problems*. R package version 1.4-11, S original by Berwin A. Turlach, R port by Andreas Weingessel.
- Tutz, G. and H. Binder (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics* 62, 961–971.
- Tutz, G. and F. Leitenstorfer (2009). Estimation of single-index models based on boosting techniques. , to appear. *Statistical Modelling, to appear*.
- Weisberg, S. and A. H. Welsh (1994). Adapting for the missing link. *Annals of Statistics* 22, 1674–1700.
- Yu, Y. and D. Ruppert (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 97, 1042–1054.
- Zeileis, A. (2006). Object-oriented computation of sandwich estimator. *Journal of Statistical Software* 16(9).