LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Florian Leitenstorfer & Gerhard Tutz

Estimation of Single-Index Models Based on Boosting Techniques

Technical Report Number 034, 2008
Department of Statistics
University of Munich

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

Estimation of Single-Index Models Based on Boosting Techniques

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

June 19, 2008

Abstract

In single-index models the link or response function is not considered as fixed. The data determine the form of the unknown link function. In order to obtain a flexible form of the link function we specify the link function as an expansion in basis function and propose to estimate parameters as well as the link function by weak learners within a boosting framework. It is shown that the method is a strong competitor to existing methods. The method is investigated in simulation studies and applied to real data.

Keywords: Single-Index Models, Boosting, P-splines, Choice of Link Function.

1 Introduction

In standard linear regression as well as in generalized linear models, for given data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, the conditional expectation of $y_i|\mathbf{x}_i$ is modeled by $E(y_i|\mathbf{x}_i) = h(\mathbf{x}_i'\boldsymbol{\beta})$. Usually it is assumed that the response function $h(.)$ is fixed and known, for example $h(.) = \mathrm{id}(.)$ yields the classical linear model. A flexible generalization of classical approaches is the so-called single-index model. Here, $h(.)$ is assumed to be unknown and has to be estimated by nonparametric techniques, whereas the parameter vector $\boldsymbol{\beta}$ (also called index vector) is identifiable up to a constant of proportionality. Such a model may be seen as a special case of projection pursuit regression, see Friedman and Stützle (1981) or as an alternative to additive models (Hastie and Tibshirani 1990).

Several approaches to the estimation of single-index models have been proposed in the literature. One popular techniques is based on average derivative estimation, which exploits the fact that the average gradient of $h(\mathbf{x}_i'\boldsymbol{\beta})$ is proportional to $\boldsymbol{\beta}$. This gradient may be estimated by using nonparametric techniques

(see e.g. Stoker (1986) or Powell, Stock, and Stoker (1989)). The method may yield unstable estimates or even fail when the covariate dimension is high. Hristache, Juditsky, and Spokoiny (2001) developed an iterative algorithm to resolve this drawback. Another approach is based on M-estimation, which considers the unknown link function as an infinite dimensional nuisance parameter (see e.g. Klein and Spady (1993)). In all these aforementioned procedures the focus is on accurate estimation of the index vector $\boldsymbol{\beta}$.

Other authors focus more on the estimation of $h(.)$. Based on kernel regression techniques, Härdle, Hall, and Ichimura (1993) investigate the optimal amount of smoothing in single-index models when simultaneous estimating $\boldsymbol{\beta}$ and the bandwidth. Weisberg and Welsh (1994) proposed an algorithm that alternates between the estimation of $\boldsymbol{\beta}$ and $h(.)$. More recently, Yu and Ruppert (2002) suggested to use penalized regression splines (P-splines, see Eilers and Marx (1996) for details) to estimate $h(.)$. 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)). From a Bayesian point of view, Antoniadis, Gregoire, and McKeague (2004) considered the P-spline approach.

Recently, boosting approaches became more and more important in nonparametric regression. For instance, Bühlmann and Yu (2003) estimated additive models using boosting whereas Tutz and Leitenstorfer (2007) tackled monotonicity restrictions in monotonic regression with similar techniques. In the following, we present a boosting algorithm for estimating single-index models which uses an alternating scheme in the sense of Weisberg and Welsh (1994) where the estimation of $h(.)$ is obtained by a P-spline approach. The advantage of our approach is that with slight modifications, it is able to do variable selection. Thus one can estimate single-index models in cases where the number of covariates is high compared to sample size, where more traditional approaches become unstable or even fail to produce an estimate. As the presented examples will show, the proposed procedures produce accurate estimates for both the index-vector $\boldsymbol{\beta}$ and the smooth function $h(.)$.

2 Estimation of a single-index model by boosting techniques

In the following, we focus on a single-index model with Gaussian errors, i.e. for scalar responses y_i and p-dimensional covariates $\mathbf{x}_i$, $i = 1, \ldots, n$, we assume

$$y_i = h(\mathbf{x}_i'\boldsymbol{\beta}) + \epsilon_i, \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $h(.)$ is a univariate smooth function. In order to make the problem identifiable, $||\boldsymbol{\beta}|| = 1$ is postulated, where $||.||$ denotes the Euclidean norm. Note that $\boldsymbol{\beta}$ contains no intercept; it is included in $h(.)$.

Following Yu and Ruppert (2002), we suggest to estimate $h(.)$ by using penalized regression splines (P-splines). Therefore, $h(.)$ is expanded into m basis

functions, i.e.

$$h(\eta_i) = \sum_{j=1}^{m} \alpha_j B_j(\eta_i),$$

where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$. We use truncated power series basis functions $B_j(.)$ of degree q, which have also been used for example by Ruppert (2002) in P-spline regression. Thus the functions have the form $B_1(\eta) = 1, B_2(\eta) = \eta, B_{q+1}(\eta) = \eta^q$, $B_{q+j}(\eta) = |\eta - t_j|_+, j > 1$, where $t_1, t_2, \ldots$ are fixed knots. Therefore a sequence of knots $\{t_j\}$ has to be placed in a certain domain $[\eta_{\min}, \eta_{\max}]$. With $\tilde{m}$ denoting the number of interior knots, the number of basis functions is determined by $m = \tilde{m}+q+1$. In P-spline regression, usually a rather high number of equidistant knots is used (say $\tilde{m} = 20$ or 40) and the smoothness of the function estimate is controlled by appropriate penalization. Following Ruppert (2002), we suggest to penalize the squared coefficients that belong to the truncated powers, i.e. $\sum_{j=q+2}^{m} \alpha_j$. We prefer the truncated power series over B-splines, since the former is more convenient when flexibility of knot locations is desired (see below). In contrast P-spline regression with B-splines requires an equally spaced knot mesh when simple difference based penalties are used (see Eilers and Marx (1996)). Using matrix notation, let the response vector be given by $\mathbf{y}' = (y_1, \ldots, y_n)$ and the design matrix by $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(p)})$, where $\mathbf{x}_{(j)} = (x_{1j}, \ldots, x_{nj})'$ denotes the jth covariate, $j = 1, \ldots, p$. Then, an estimator of the single-index model (1) is formulated as minimizer of the penalized least squares criterion

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{y} - \mathbf{B}(\boldsymbol{\eta})\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{B}(\boldsymbol{\eta})\boldsymbol{\alpha}) + \lambda_P \boldsymbol{\alpha}'\mathbf{P}\boldsymbol{\alpha}, \tag{2}$$

where $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{B}(\boldsymbol{\eta}) = (B_1(\boldsymbol{\eta}), \ldots, B_m(\boldsymbol{\eta})) = (\mathbf{1}, \boldsymbol{\eta}, \ldots, \boldsymbol{\eta}^q, (\boldsymbol{\eta} - \tau_1)_+^q, \ldots, (\boldsymbol{\eta} - \tau_{\tilde{m}})_+^q)$, $B_j(\boldsymbol{\eta}) = (B_j(\eta_1), \ldots, B_j(\eta_n))'$, $\mathbf{P} = \text{diag}\{\mathbf{0}_{q+1}, \mathbf{1}_{\tilde{m}}\}$ and λ_P is a penalization parameter.

Yu and Ruppert (2002) suggest to solve (2) by using common nonlinear least squares routines. We present a novel approach, which alternates between the estimation of $\boldsymbol{\beta}$ and $h(.)$ by applying boosting techniques. Developed in the machine learning community for classification purposes (e.g. Schapire 1990), boosting became increasingly popular in statistics in the last years. Friedman (2001) showed that it may be seen as a optimization technique in function space. The approach has been extended to regression modeling with continuous response settings (e.g. Bühlmann and Yu (2003)). The basic idea is to fit a function iteratively by fitting in each stage a "weak" learner to the current residual; see Bühlmann and Hothorn (2008) for a nice overview on boosting techniques.

In the present setting the objective is minimization of criterion (2) by means of boosting. That means boosting techniques are applied in two stages, namely once for the estimation of the index vector $\boldsymbol{\beta}$ and once for the estimation of the vector of basis coefficients $\boldsymbol{\alpha}$. For the latter, partial derivation of (2) with respect to $\boldsymbol{\alpha}$ leads to an L_2-type boosting algorithm as proposed by Bühlmann and Yu (2003). That means in each iteration the current residuals $\mathbf{y} - \hat{h}^{\text{old}}(\boldsymbol{\eta})$ with a

P-spline as weak learner (λ_P has be to sufficiently large in order to obtain a weak learner) are refitted. More concrete, starting with $\hat{h}^{\text{old}}(\boldsymbol{\eta}) = \mathbf{B}(\boldsymbol{\eta})\hat{\boldsymbol{\alpha}}^{\text{old}}$ one obtains the improved estimate by computing $\hat{\mathbf{a}} = (\mathbf{B}(\boldsymbol{\eta})'\mathbf{B}(\boldsymbol{\eta}) + \lambda_P\mathbf{P})^{-1}\mathbf{B}(\boldsymbol{\eta})'[\mathbf{y} - \hat{h}^{\text{old}}(\boldsymbol{\eta})]$, yielding the updated coefficient vector

$$\hat{\boldsymbol{\alpha}}^{\text{new}} = \hat{\boldsymbol{\alpha}}^{\text{old}} + \hat{\mathbf{a}}$$

and therefore the new response function

$$\hat{h}^{\text{new}}(\boldsymbol{\eta}) = \mathbf{B}(\boldsymbol{\eta})\hat{\boldsymbol{\alpha}}^{\text{new}} = \mathbf{B}(\boldsymbol{\eta})[\hat{\boldsymbol{\alpha}}^{\text{old}} + \hat{\mathbf{a}}]. \tag{3}$$

However, in this procedure $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ have to be specified. Thus, since estimation of $\boldsymbol{\beta}$ and $h(.)$ is interdependent, one needs to interlock estimation of $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ adequately. Therefore, in each iteration of an L_2-type boosting algorithm for estimation of $h(.)$, we suggest to estimate $\boldsymbol{\beta}$ also by boosting techniques, that means in a stepwise manner by use of a weak learner. For given estimate $\hat{\boldsymbol{\alpha}}$ (and a previous estimate of $\boldsymbol{\beta}$) a weak learner is used to update the estimate of $\boldsymbol{\beta}$. For the derivation of a weak learner it is useful to consider the modified criterion

$$Q^*(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}) = Q(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}) + \lambda_R\boldsymbol{\beta}'\boldsymbol{\beta},$$

where λ_R is a ridge-type penalty on $\boldsymbol{\beta}$. Partial derivation subject to $\boldsymbol{\beta}$ leads to

$$\frac{\partial Q^*(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{X}'\mathbf{D}(\boldsymbol{\eta})[\mathbf{y} - \hat{h}(\boldsymbol{\eta})] + 2\lambda_R\boldsymbol{\beta}, \tag{4}$$

with $\hat{\mathbf{D}}(\boldsymbol{\eta}) = \text{diag}\{\partial\hat{h}(\eta_i)/\partial\eta\}_{i=1}^n$. Computation of the expectation of the derivative of (4) with respect to $\boldsymbol{\beta}$ and standard Fisher scoring techniques as in generalized linear models (see e.g. McCullagh and Nelder (1989)), yields an application the one-step Fisher scoring estimate for $\boldsymbol{\beta}$,

$$\hat{\mathbf{b}} = (\mathbf{X}'\hat{\mathbf{D}}(\boldsymbol{\eta})^2\mathbf{X} + \lambda_R\mathbf{I})^{-1}\mathbf{X}'\hat{\mathbf{D}}(\boldsymbol{\eta})(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}), \tag{5}$$

where $\hat{\boldsymbol{\mu}}^{(l-1)}$ is the current estimate of the mean. With these building blocks, we are able to give a boosting algorithm that alternates between estimation of $\boldsymbol{\beta}$ and $h(.)$, as suggested by Weisberg and Welsh (1994). Defining $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ with $\mu_i = h(\eta_i)$, it has the following form.

Algorithm: SIBoost

Step 1 (Initialization)

Set $\hat{\boldsymbol{\alpha}}^{(0)} = (0, 0, 0, \ldots, 0)'$, $\hat{\boldsymbol{\beta}}^{(0)} = (0, \ldots, 0)'$ $\hat{\boldsymbol{\eta}}^{(0)} = (0, \ldots, 0)'$, $\hat{\boldsymbol{\mu}}^{(0)} = (0, \ldots, 0)'$ and $\mathbf{D}_0 = \mathbf{I}$.

Step 2 (Iteration)

For $l = 1, 2 \ldots,$

1. *Estimation of the parametric term* $\boldsymbol{\beta}$

   Compute the penalized estimate based on one-step Fisher scoring from (5),

   $$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{D}_{l-1}^2\mathbf{X} + \lambda_R\mathbf{I})^{-1}\mathbf{X}'\mathbf{D}_{l-1}{}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}). \tag{6}$$

   Set $\hat{\tilde{\boldsymbol{\beta}}}^{(l)} = \hat{\boldsymbol{\beta}}^{(l-1)} + \hat{\mathbf{b}}$ and normalize the estimate to length one by computing $\hat{\boldsymbol{\beta}}^{(l)} = \hat{\tilde{\boldsymbol{\beta}}}^{(l)}/||\hat{\tilde{\boldsymbol{\beta}}}^{(l)}||$. Set $\hat{\boldsymbol{\eta}}^{(l)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(l)}$.

2. *Estimation of smooth function* $h(.)$

   Compute B-spline basis function of $\hat{\boldsymbol{\eta}}^{(l)}$, $\mathbf{B}^{(l)} = \mathbf{B}(\hat{\boldsymbol{\eta}}^{(l)})$, by using an equidistant knot mesh in a pre-specified interval $[\eta_{\min}, \eta_{\max}]$ and update $\tilde{\boldsymbol{\mu}}^{(l)} = \boldsymbol{B}^{(l)}\hat{\boldsymbol{\alpha}}^{(l-1)}$. The least squares fit of the P-Spline to the current residuals is then given by

   $$\hat{\mathbf{a}} = (\mathbf{B}^{(l)'}\mathbf{B}^{(l)} + \lambda_P\mathbf{P})^{-1}\mathbf{B}^{(l)'}(\mathbf{y} - \tilde{\boldsymbol{\mu}}^{(l-1)}).$$

   Set

   $$\hat{\boldsymbol{\alpha}}^{(l)} = \hat{\boldsymbol{\alpha}}^{(l-1)} + \hat{\mathbf{a}}, \quad \hat{\boldsymbol{\mu}}^{(l)} = \mathbf{B}^{(l)}\hat{\boldsymbol{\alpha}}^{(l)} \tag{7}$$

   and $\mathbf{D}_l = \text{diag}\{\frac{\partial}{\partial\eta}\hat{h}(\hat{\eta}_i^{(l)})\}_{i=1}^n$, where $\frac{\partial}{\partial\eta}\hat{h}(\hat{\eta}_i^{(l)}) = \sum_{k=1}^m \frac{\partial}{\partial\eta}B_k(\hat{\eta}_i^{(l)})\hat{\alpha}_k^{(l)}$.

Boosting algorithms are usually regularized via the number of iterations, i.e. in order to avoid overfitting, an appropriate stopping criterion is necessary. Cross-validation techniques might be applied, but in boosting computational costs can be prohibitively high. Alternatively, one may use AIC-type model selection criteria to optimize the number of iterations, see Bühlmann and Yu (2003) and Bühlmann (2006). In the following, we will pursue this approach. Therefore, a hat-matrix is needed to estimate the degrees of freedom by its trace. For the update scheme given in (7), the hat-matrix derived by Bühlmann and Yu (2003) cannot be used. However, a similar expression can be found in our case. Let $l = 1, 2, \ldots,$ and define $\mathbf{S}_l = (\mathbf{B}^{(l)'}\mathbf{B}^{(l)} + \lambda_2\mathbf{P})^{-1}\mathbf{B}^{(l)'}$, in the lth iteration one has

$$\begin{aligned}
\boldsymbol{\mu}^{(l)} &= \mathbf{B}^{(l)}\hat{\boldsymbol{\alpha}}^{(l)} \\
&= \mathbf{B}^{(l)}[\hat{\boldsymbol{\alpha}}^{(l-1)} + \mathbf{S}_l(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)})] \\
&= \mathbf{B}^{(l)}[\hat{\boldsymbol{\alpha}}^{(l-2)} + \mathbf{S}_{l-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-2)}) + \mathbf{S}_l(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)})].
\end{aligned}$$

Thus, by recursive definition and setting $\mathbf{H}_0 = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$, a hat-matrix $\mathbf{H}_l$ satisfying $\hat{\boldsymbol{\mu}}^{(l)} = \mathbf{H}_l\mathbf{y}$ is given by

$$\mathbf{H}_l = \mathbf{B}^{(l)}[\sum_{j=1}^{l}\mathbf{S}_l(\mathbf{I} - \mathbf{H}_{l-1})]. \tag{8}$$

The trace of this hat-matrix may now be used as an estimate for the degrees of freedom in an selection criterion, i.e. $\widehat{\mathrm{df}}_l = \mathrm{tr}(\mathbf{H}_l)$. Our experiments suggest that the corrected AIC (see Hurvich, Simonoff, and Tsai (1998)), given by

$$\mathrm{AIC}_c(l) = \log(\hat{\sigma}^2) + \frac{1 + \widehat{\mathrm{df}}_l/n}{1 - (\widehat{\mathrm{df}}_l + 2)/n}$$

with $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l)})'(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l)})$ works reasonably well. The optimal number of iterations is then estimated by $l_{\mathrm{opt}} = \arg\min_l \mathrm{AIC}_c(l)$. In order to save computing time, we propose an early stopping strategy: if $\mathrm{AIC}_c(l)$ increases five times in a row, the iterations are stopped. This proceeding allows us to set the upper limit of iterations L rather high also in simulation studies, while it does not show any noticeable loss in efficiency compared to full search. We consider also an alternative optimization criterion, the g-prior minimum description length (gMDL) from Hansen and Yu (2001), given by

$$\mathrm{gMDL}(l) = \log[n\hat{\sigma}^2/\{n - \widehat{\mathrm{df}}_l\}] + \frac{\widehat{\mathrm{df}}_l}{n} \log\left[\frac{\sum_{i=1}^n y_i^2 - n\hat{\sigma}^2}{\widehat{\mathrm{df}}_l n\hat{\sigma}^2/\{n - \hat{\mathrm{df}}_l\}}\right].$$

It is a hybrid between AIC and BIC and has recently been proven to be successful in boosting techniques by Bühlmann and Yu (2006).

Another important point is the determination of the interval $[a, b]$ where the knots of the B-spline basis for the estimation of $h(.)$ must be placed. Since $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ and thus $\eta_{\min}$ and $\eta_{\max}$ are unknown, an approximation is needed. A rough guide makes use of the postulate $||\boldsymbol{\beta}|| = 1$. Then, by applying the Cauchy-Schwarz inequality it can be easily seen that $[\eta_{\min}, \eta_{\max}]$ is always in $[-u, u]$ with $u = \max_{i=1,\ldots,n}\{||\mathbf{x}_i||\}$. However, consider the case of a higher number of covariates p, where only some of them have a stronger influence on $\boldsymbol{\eta}$. The true range of η-values may then cover only a small portion of $[-u, u]$. Since it is more likely that the ηs are located near the center of $[-u, u]$ instead of somewhere near the boundary, we suggest to use a higher number of knots (say $\tilde{m} = 40$) placed at a mesh that is more dense around zero. A strategy for obtaining such a grid is to involve the quantiles of a symmetric distribution which might have somewhat heavier tails than a standard normal. We suggest to compute $\tilde{m} + 2$ quantiles equally spaced for $[0.05, 0.95]$, including the bounds of this interval. After rescaling this grid such that $-u$ corresponds to $q_{0.05}$ and u to $q_{0.95}$, respectively and removing the lowest and highest value, one obtains a set of knots $\{\tau_j\}_{j=1}^{\tilde{m}}$ with the desired properties. A t-distribution with three to ten degrees of freedom turns out to be a sensible choice for computation of the quantiles.

3 Numerical comparisons

We start our numerical comparisons with a simulation study similar to the one conducted by Antoniadis, Gregoire, and McKeague (2004). Therefore, a single-index model as given in (1) is considered, where

$$h(v) = v^2 \exp(v). \tag{9}$$

The covariates $\mathbf{x}_{(j)}$, $j = 1, \ldots, p$, are drawn from a $U[-1, 1]$-distribution, where $p = 4$ and 10 is investigated. The corresponding index vectors are given by $\boldsymbol{\beta} = (1, 1, 1, 2)'/\sqrt{7}$ in the first and $\boldsymbol{\beta} = (1, 1, 1, 2, 0, \ldots, 0)'/\sqrt{7}$ in the latter case. We use a sample size of $n = 100$ and generate the errors from a normal distribution with two different noise levels, $\sigma = 0.2$ and 0.5.

In the following, we apply SIBoost optimized by AIC_c as well as gMDL. A truncated power series basis of degree three is used with a sequence of $\tilde{m} = 40$. The grid of knots is determined by the quantile method described above, where quantiles of a $t(5)$-distribution are applied. The number of iterations is limited by $L = 2000$, but in most cases we found that the minimal of the selection criteria is met much earlier. We always use the early stopping strategy. Since boosting needs a weak learner, the penalization parameters λ_P and λ_R should be chosen sufficiently high. We investigate the sensitivity of the SIBoost estimates in dependence of these parameters by some preliminary simulations. Therefore, we generated 100 simulated data sets for various settings, varied λ_P while keeping λ_R fixed at 100 and varied λ_R while keeping λ_P fixed at 100. For assessing the performance of the estimates, we consider two different criteria. The accuracy of the fit is measured by the averaged squared error,

$$\text{ASE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2, \tag{10}$$

and the quality of the estimate of the index vector $\boldsymbol{\beta}$ by the angle,

$$\text{angle}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \arccos(\hat{\boldsymbol{\beta}}' \boldsymbol{\beta}). \tag{11}$$

In Figure 1, we give boxplots of log(ASE) and angle for AIC_c-optimized SIBoost estimates for certain ranges of λ_P and λ_R values in the case of $p = 10$, $\sigma = 0.5$ and $n = 100$. It is seen that SIBoost is rather robust against the choice of the penalization parameters. Other settings of p, σ and n as well as the gMDL criterion show rather similar patterns and hence are not given. In the following, $\lambda_P = \lambda_R = 100$ is used since the estimated optimal number of iterations is clearly below 2000 in most cases.

In the following the performance of SIBoost is compared to alternative estimators. In detail, we consider:

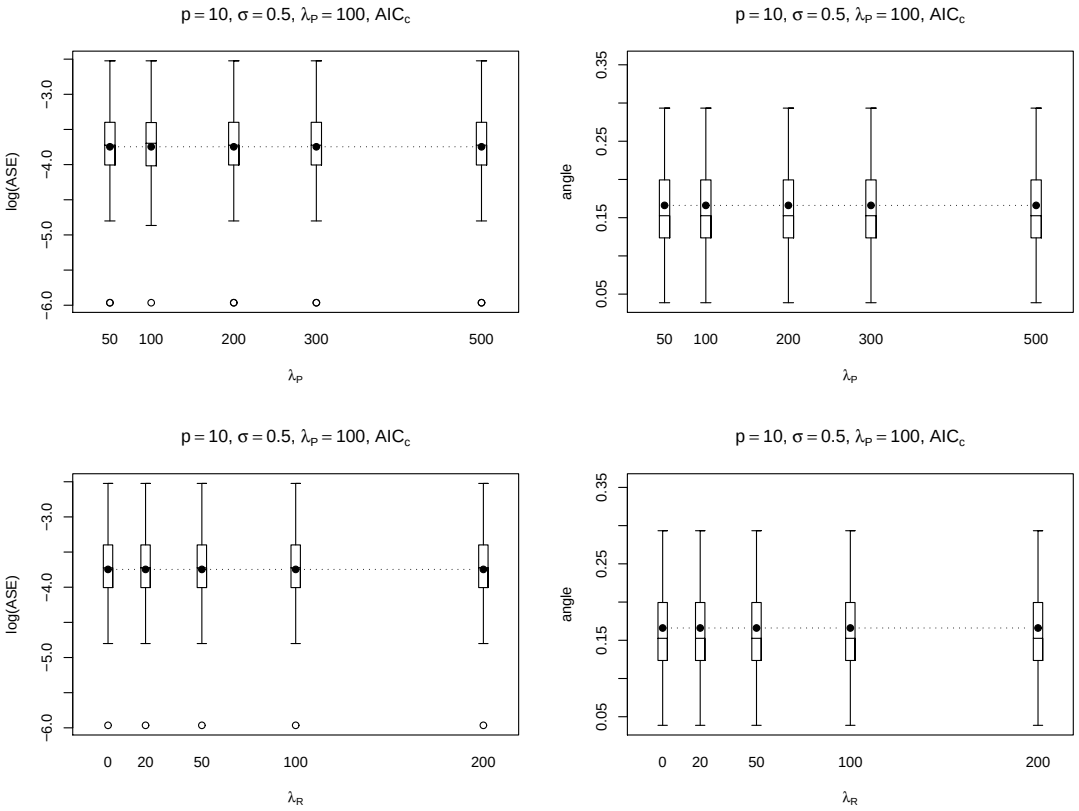

FIGURE 1: *Boxplots of* $\log(ASE)$ *and angle for* $p = 10$, $\sigma = 0.5$ *and* $n = 100$ *with varying penalization parameters over* 100 *simulated data sets. Upper panels:* $\lambda_R = 100$ *fixed, lower panels:* $\lambda_P = 100$ *fixed.*

- The oracle estimator, assuming that $h(.)$ from (9) is known. It is computed by standard nonlinear least squares techniques, using the function `nls()` from `R` and is denoted by NLS. The estimate of $\boldsymbol{\beta}$ is scaled to length one for comparison to the other methods.

- A projection pursuit regression (PPR) fit from Friedman and Stützle (1981) that stops after the first term is included in the model. This strategy can be considered as a way to fit a single-index model. We use the `R` implementation `ppr()`, which applies by default a super smoother for the estimation of the smooth function. The degree of smoothness is controlled by local cross-validation.

- The penalized spline estimation technique based on nonlinear least squares, proposed by Yu and Ruppert (2002) (YR). We apply Yan Yu's `MATLAB` code.

In order obtain comparable dimensions, 20 knots are used to construct the truncated power series basis (of degree three) for the penalized spline estimation. The smoothing parameter λ is determined by GCV performing a grid search over 30 values where $\log_{10}(\lambda) \in [-6, 7]$, as described in the original paper.

- The direct estimation method of the index coefficient introduced by Hristache, Juditsky, and Spokoiny (2001) (HJS). Therefore, we use the `R` package `EDR` provided by Jörg Polzehl. It implements the more general dimension reduction approach given in Hristache, Juditsky, Polzehl, and Spokoiny (2001), where a single-index model can be considered as a special case. A modification is used which improves the original methodology (`method="HJPS2"`, for details see Polzehl and Sperlich (2007)). For additional parameters, the default settings of the package turned out to be a sensible choice. Note that this procedure only provides an estimate for the index vector $\boldsymbol{\beta}$ and not for the unknown function $h(.)$.

We also fitted ordinary least squares without intercept but performance was so poor that the results are omitted. In Figure 2, the simulation results for the aforementioned settings and estimation methods are given. It is seen that in the lower dimensional case of $p = 4$, the SIBoost estimates come closest to NLS (where the true function $h(.)$ is known) in terms of MSE. In terms of the angle between the true and estimated parameter vector, the median over the simulations does not differ strongly over the methods. However, it can be seen that in particular YR shows rather high variability (some outliers not shown). This indicates some instability in this estimation methods, which seems not to be the case for SIBoost. For $p = 10$, in the lower noise case YR is a strong performer. However, also here this method does not seem to be very stable showing high variability. In the case of $\sigma = 0.5$, also HJS tends to produce severe outliers when estimating $\boldsymbol{\beta}$ (note that we truncated the axes of ordinates for the angle, otherwise these effects would become even more obvious). In contrast, the boosting approach yields quite reliable estimates also in higher dimensions. In order to verify that bootstrap performs better than alternative methods in higher dimensions we enlarged the number of predictors to $p = 25$ with the vector β filled up with zeros and rerun the simulation. The results are given in Figure 3. It is seen that for this case boosting approaches outperform the competitors distinctly.

4 An Application

In the following the method is applied to the body fat data set that was originally used by Penrose et al. (1985). The study aims at the estimation of the percentage of body fat by various body measurements for 252 men. The thirteen

Model	SIB(aicc) (100, 100)	SIB(aicc) (10000000, 1)	Normal Regression	Log Normal Regression
β_1	0.029	0.032	0.045(0.052)	0.029(0.013)
β_2	0.010	0.009	0.014(0.581)	-0.018(0.268)
β_3	-0.123	-0.137	-0.018(0.016)	-0.059(0.675)
β_4	-0.353	-0.184	-0.333(0.007)	-0.247(0.008)
β_5	-0.063	-0.056	-0.086(0.167)	-0.036(0.089)
β_6	0.487	0.377	0.403(0.0)	0.449(0.0)
β_7	-0.154	-0.056	-0.261(0.001)	-0.135(0.0)
β_8	0.061	0.043	0.197(0.182)	0.097(0.016)
β_9	-0.070	-0.057	-0.001(0.674)	-0.017(0.992)
β_{10}	-0.011	0.060	-0.021(0.703)	0.060(0.864)
β_{11}	0.108	0.067	0.142(0.315)	0.086(0.141)
β_{12}	0.170	0.089	0.161(0.044)	0.198(0.151)
β_{13}	-0.737	-0.880	-0.749(0.000)	-0.807(0.012)

TABLE 1: *Parameter estimates for body fat data*

regressors are age (1), weight (lbs) (2), height (inches) (3), neck circumference (4), chest circumference (5), abdomen 2 circumference (6), hip circumference (7), thigh circumference (8), knee circumference (9), ankle circumference (10), biceps (extended) circumference (11), forearm circumference (12), and wrist circumference (13). All circumferences are measured in cm. The percent body fat has been calculated by using the body density determined by underwater weighting.

Figure 4 shows the link functions for the standard choice $\lambda_P = \lambda_R = 100$ and, for illustration, for the extreme choice $\lambda_P = 10^7, \lambda_R = 1$. The latter yields a very smooth function whereas the former is somewhat wiggly and rather close to the data at the boundary. The (standardized) estimated coefficients for centered response data are given in Table 1. In addition a normal regression model was fit and a model with logarithmic response. It is seen that for the normal model and the log-normal model different variables seem to be relevant. In some cases only coefficients for one of the two models are significant, (see variable 3, 8 and 12). The values of the multiple R-squared suggest that the linear model has a slightly higher explanatory power than the logarithmic model (0.740 for the linear model and 0.612 for the logarithmic model). The more flexible approach lets the link function be determined by the data. As far as estimates are concerned it is seen that for the model with the more wiggly link that is closer to the data ($\lambda_P = \lambda_P = 100$) most of the estimates are closer to those of the log-link model.

Acknowledgement

We thank Sebastian Petry for his help with computational work.

References

Antoniadis, A., G. Gregoire, and I. W. McKeague (2004). Bayesian estimation in single-index models. *Statistica Sinica 14*, 1147–1164.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics 34*, 559 –583.

Bühlmann, P. and T. Hothorn (2008). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science (to appear)*.

Bühlmann, P. and B. Yu (2003). Boosting with the L_2-loss: regression and classification. *Journal of the American Statistical Association 98*, 324–339.

Bühlmann, P. and B. Yu (2006). Sparse boosting. *Journal of Machine Learning Research 7*, 1001–1024.

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.

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

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics 29*, 1189–1232.

Friedman, J. H. and W. Stützle (1981). Projection pusuit regression. *J. Amer. Statist. Assoc. 76*, 817–823.

Hansen, M. and B. Yu (2001). Model selection and minimum description length principle. *Journal of the American Statistical Association 96*, 746–774.

Härdle, W., P. Hall, and H. Ichimura (1993). Optimal smoothing in single-index models. *Annals of Statistics 21*, 157–178.

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

Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny (2001). Structure adaptive approach for dimension reduction. *Annals of Statistics 29*, 1537–1566.

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.

Hurvich, C. M., J. S. Simonoff, and C. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B 60*, 271–293.

Klein, R. L. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*, 387–421.

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

Penrose, K. W., A. G. Nelson, and A. G. Fisher (1985). Generalized body composition prediction equation for men using simple measurement techniques. *Medicine and Science in Sports and Exercise 17*, 189.

Polzehl, J. and S. Sperlich (2007). Structural adaptive dimension reduction. WIAS report 1227, Weierstrass Institute for Applied Analysis and Stochastics, Berlin.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*, 1403–1430.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*, 735–757.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*, 197–227.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica 54*, 1461–1481.

Tutz, G. and F. Leitenstorfer (2007). Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics 16*, 165–188.

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.

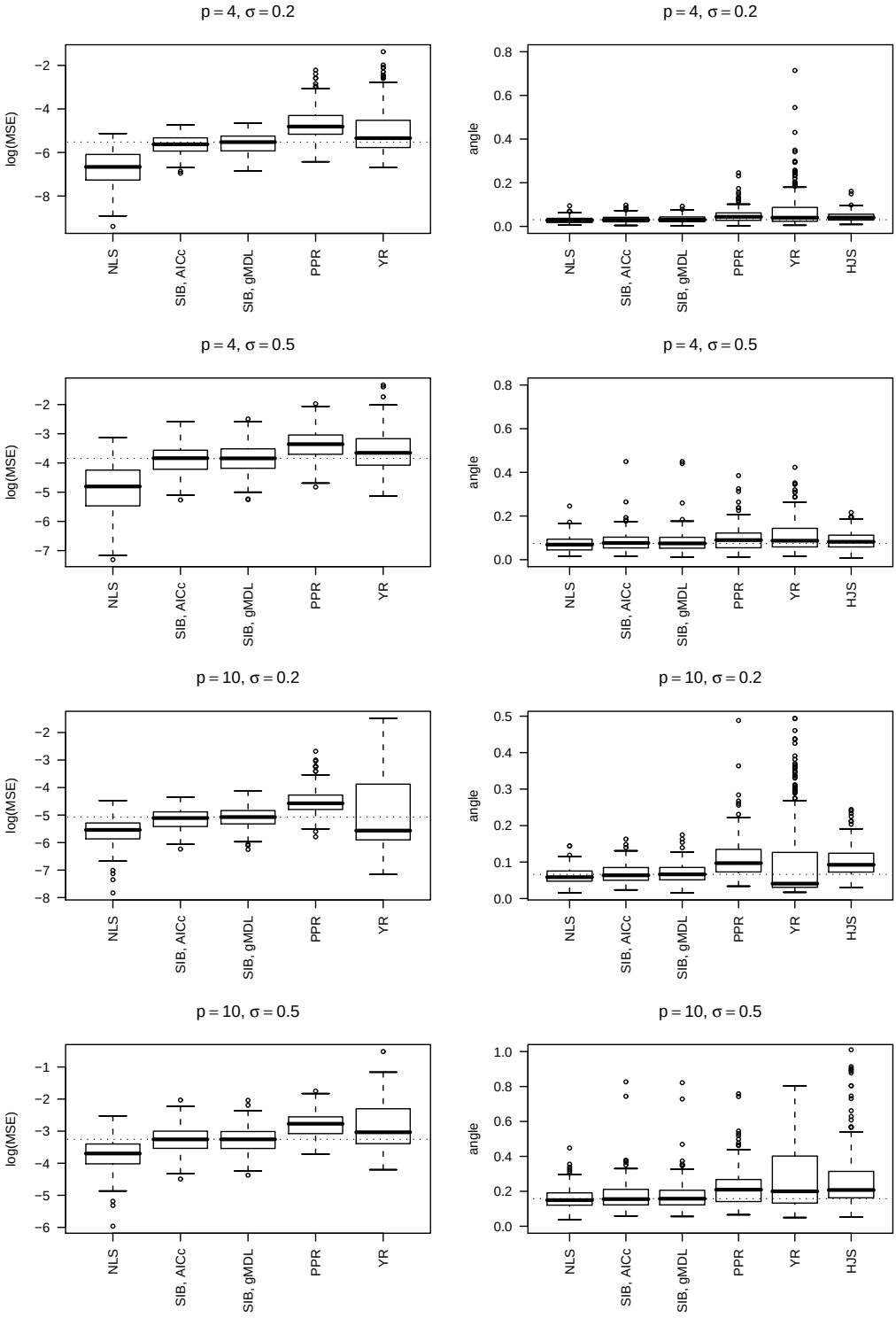

FIGURE 2: *Boxplots of* $\log(ASE)$ *and angle for several simulation settings and estimation methods*

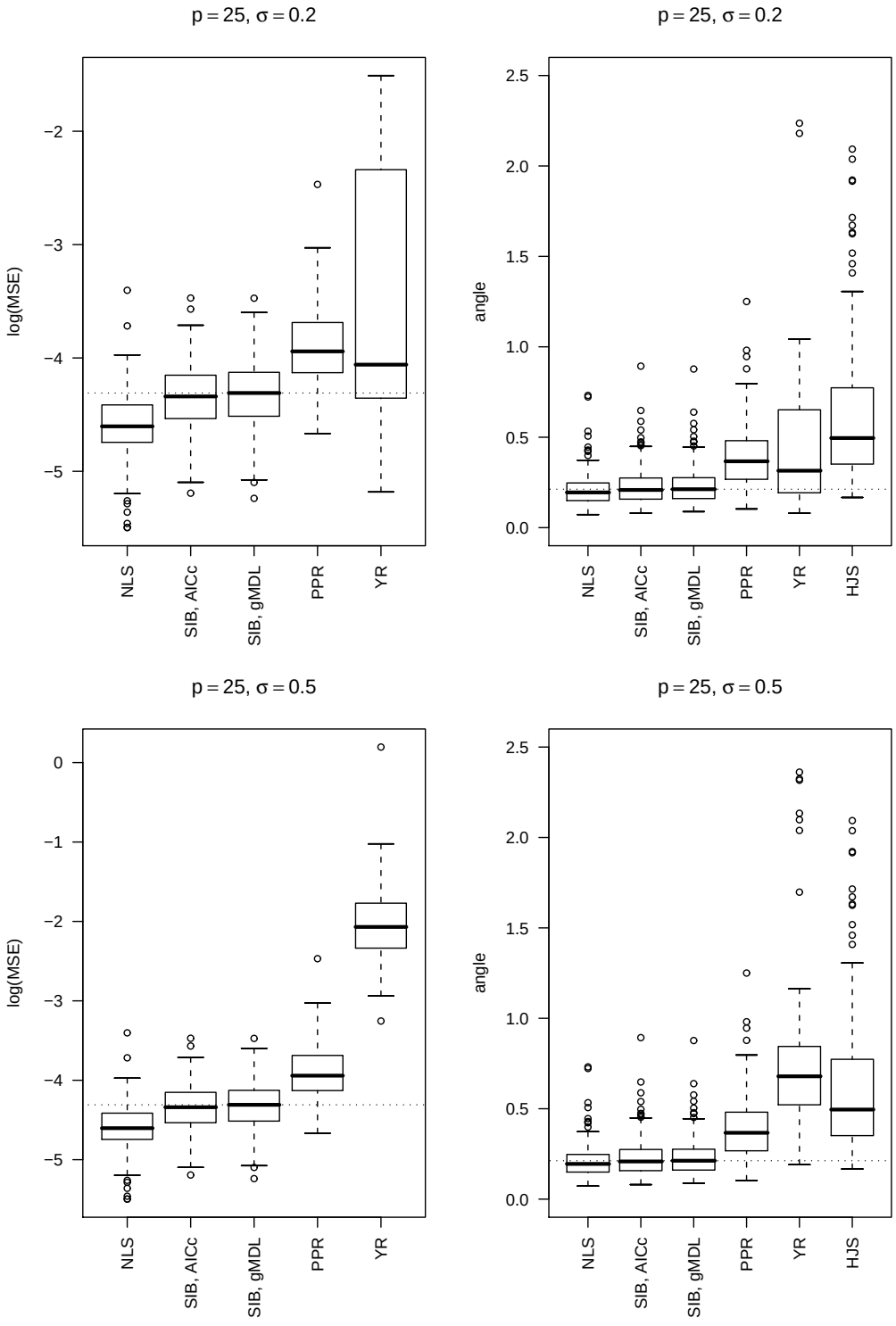

FIGURE 3: *Boxplots of* $\log(ASE)$ *and angle for high dimensional case*

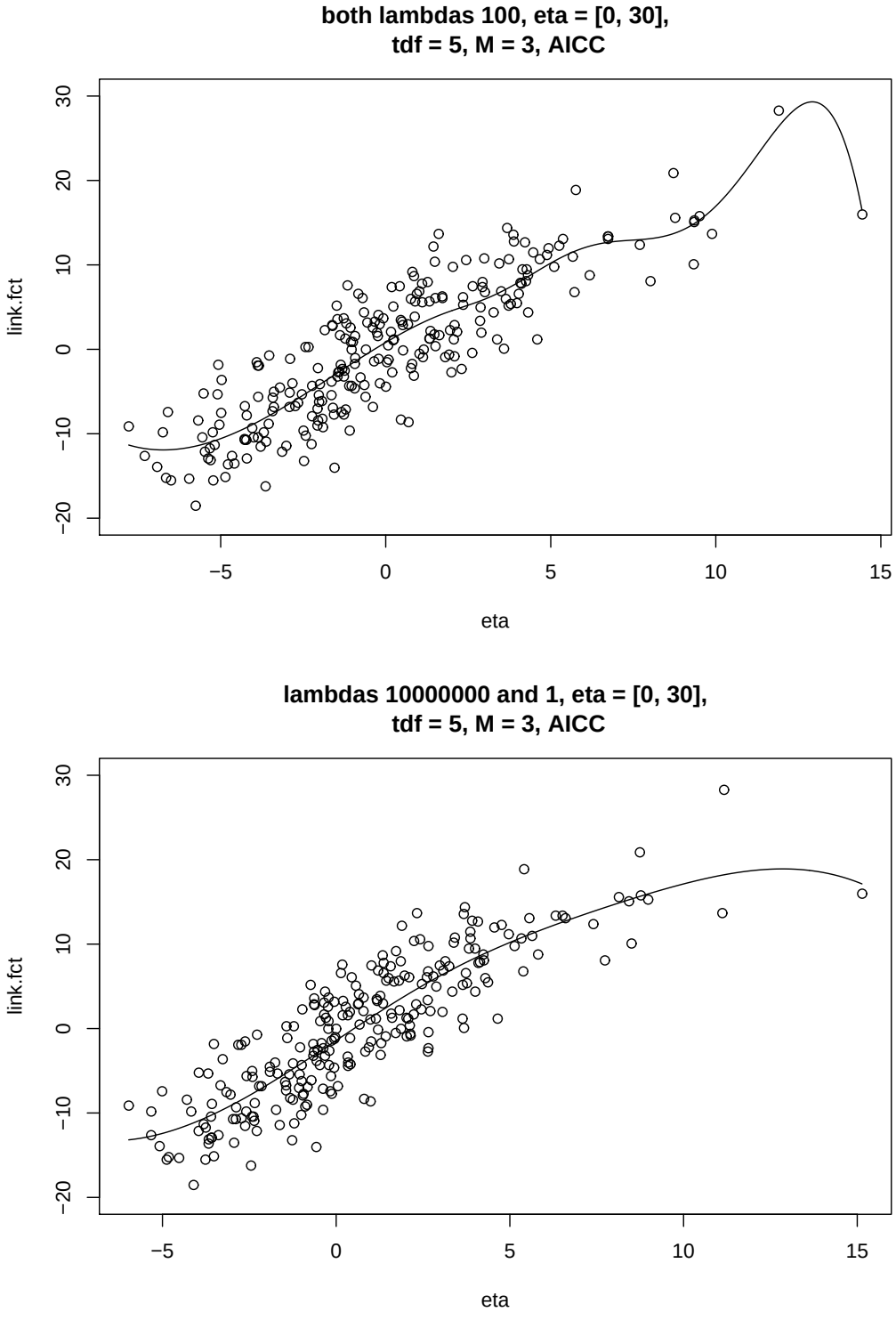

FIGURE 4: *Estimated link functions for the standard choice* $\lambda_R = \lambda_P = 100$ *(left) and* $\lambda_P = 10^7, \lambda_R = 1$ *(right)*