Statistics and Computing
https://doi.org/10.1007/s11222-019-09870-4

Transformation boosting machines

Torsten Hothorn[1]

Received: 26 November 2018 / Accepted: 27 March 2019

Abstract
The broad class of conditional transformation models includes interpretable and simple as well as potentially very complex models for conditional distributions. This makes conditional transformation models attractive for predictive distribution modelling, especially because models featuring interpretable parameters and black-box machines can be understood as extremes in a whole cascade of models. So far, algorithms and corresponding theory was developed for special forms of conditional transformation models only: maximum likelihood inference is available for rather simple models, there exists a tailored boosting algorithm for the estimation of additive conditional transformation models, and a special form of random forests targets the estimation of interaction models. Here, I propose boosting algorithms capable of estimating conditional transformation models of arbitrary complexity, starting from simple shift transformation models featuring linear predictors to essentially unstructured conditional transformation models allowing complex nonlinear interaction functions. A generic form of the likelihood is maximized. Thus, the novel boosting algorithms for conditional transformation models are applicable to all types of univariate response variables, including randomly censored or truncated observations.

Keywords Transformation model · Distribution regression · Conditional distribution function · Conditional quantile function · Conditional hazard function · Probabilistic forecasting

1 Introduction

The future remains unknown, yet we have witnessed considerably improved predictions owing to advances in statistical and machine learning over the last two decades. Numerous procedures, such as support vector machines, random forests, and tree boosting, deliver accurate point predictions of conditional means. However, in many applications, a mean prediction is not good enough. Full predictive distributions, also known as probabilistic forecasts, are required in applications where an assessment of the associated uncertainty is essential, for example in models of future disease progression (Küffner et al. 2015), electricity demand (Cabrera and Schulz 2017), stock asset returns (Mitrodima and Griffin 2017), and counterfactual distributions (Chernozhukov et al. 2013). In these applications, the prediction "takes the form of a predictive probability distribution over future quantities or events of interest" (Gneiting and Katzfuss 2014). Here, I present a novel generic boosting approach to the estimation of full predictive distributions under mild assumptions.

Apart from completely model-free procedures (such as kernel smoothing, Li and Racine 2008), four main approaches of obtaining predictive distributions exist. (1) Flexible parametric models for conditional density functions rely on a strict parametric model of the response distribution those parameters might be linked to predictor variables in complex ways, for example, in generalized additive models for location, scale, and shape (GAMLSS, Rigby and Stasinopoulos 2005) and in heteroscedastic Bayesian additive regression tree ensembles (Pratola et al. 2017). (2) Quantile regression models for conditional quantiles of interest can be modelled in a linear or nonlinear additive form (Koenker 2005); more complex relationships can be estimated by quantile regression forests (Meinshausen 2006; Athey et al. 2019). (3) Distribution regression and transformation models potentially allow response-varying (or time-varying) effects (Foresi and Peracchi 1995; Rothe and Wied 2013; Chernozhukov et al. 2013; Wu and Tian 2013; Leorato and Peracchi 2015) in models for conditional distribution functions on the probit, logit, or complementary

✉ Torsten Hothorn
Torsten.Hothorn@uzh.ch

[1] Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich, Hirschengraben 84, 8001 Zürich, Switzerland

Published online: 16 April 2019

log–log scale. (4) Hazard regression (Kooperberg et al. 1995) aims at estimating conditional nonproportional hazard functions directly.

Boosting, and especially the statistical view on boosting (Friedman et al. 2000; Bühlmann and Hothorn 2007), have already proved helpful in these four different approaches. Mayr et al. (2012) developed boosting for GAMLSS, conditional quantile boosting was introduced by Fenske et al. (2011), and nonproportional hazard boosting was recently introduced by Lee and Chen (2018). Distribution regression is a special case of conditional transformation models (Hothorn et al. 2014). What is interesting about conditional transformation models is that very simple models, such as the linear proportional odds and hazards models, and essentially unstructured models for conditional distribution functions can be understood in a unified theoretical framework (Hothorn et al. 2018). The same level of generality is, however, lacking from an algorithmic perspective. The boosting algorithm introduced by Hothorn et al. (2014) is limited to additive models and explicitly excludes tree-based interaction models. Furthermore, the target function is approximate and applicable to responses observed without censoring or truncation only. The aim of this work is to establish a general computational framework that allows specification, estimation, evaluation, and comparison in a cascade of models starting with very simple linear models and ending with essentially unstructured models for conditional distribution functions for arbitrary response variables.

Section 2 gives a dense introduction to transformation models. An elaborate description and connections to well-established models can be found in Hothorn et al. (2014) and Hothorn et al. (2018). Sections 3 and 4 develop two boosting algorithms for complex and simple transformation models based on a generic form of the likelihood (technical details regarding the definition of the likelihood for all types of response variables, including random censoring and truncation, are discussed by Hothorn et al. 2018). Empirical evaluations are presented in Sect. 5.

2 Transformation models

Let Y denote a univariate and at least ordered response variable on a measurable space $(\mathcal{Z}, \mathfrak{C})$ and $\boldsymbol{X} \in \mathcal{X}$ a set of predictor variables with joint distribution $(Y, \boldsymbol{X}) \sim \mathbb{P}_{Y,\boldsymbol{X}}$. Based on random samples from $\mathbb{P}_{Y,\boldsymbol{X}}$, the goal is to estimate the conditional distribution $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ of a response given predictors. For each conditional cumulative distribution function $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = \mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}(\{\upsilon \in \mathcal{Z} \mid \upsilon < y\})$, a unique conditional transformation function $h : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$ exists such that $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h(y \mid \boldsymbol{x}))$, assuming $F_Z : \mathbb{R} \to [0, 1]$ is an a priori given cumulative distribution function of an absolutely continuous random variable Z with log-concave density f_Z (Hothorn et al. 2018). The conditional transformation function h is monotonic in y

$$h(\underline{y} \mid \boldsymbol{x}) \leq h(\bar{y} \mid \boldsymbol{x}) \quad \text{for all} \quad \underline{y} < \bar{y} \in \mathcal{Z}, \boldsymbol{x} \in \mathcal{X}. \tag{1}$$

Starting with Box and Cox (1964), shift transformation functions based on the decomposition $h(y \mid \boldsymbol{x}) = h_Y(y) - \beta(\boldsymbol{x})$ featuring a baseline transformation function $h_Y : \mathcal{Z} \to \mathbb{R}$ and a shift term $\beta : \mathcal{X} \to \mathbb{R}$ have been studied intensively. The proportional hazards (with $F_Z(z) = 1 - \exp(-\exp(z))$) and proportional odds (with $F_Z(z) = \text{expit}(z)$) models are the most well-known representatives of this class of shift transformation models (STM, often also referred to as linear or nonlinear transformation models, depending on the functional form of $\beta(\boldsymbol{x})$). Boosting procedures that allow flexible estimation of $\beta(\boldsymbol{x})$ have been studied for proportional hazards models under right censoring (Ridgeway 1999; Schmid and Hothorn 2008; Lu and Li 2008; Yue et al. 2017) and proportional odds models have been studied for ordered responses (Schmid et al. 2011). A comparison of prominent and less prominent members of this model class is given in Hothorn et al. (2018).

Structured additive transformation functions that allow interactions between the two arguments y and $\boldsymbol{x}$ of the form $h(y \mid \boldsymbol{x}) = \sum_{j=1}^{J} h_j(y \mid \boldsymbol{x})$ lead to conditional transformation models (CTM, Hothorn et al. 2014). The J partial transformation functions h_j allow formulation of problem-specific effects of the predictors $\boldsymbol{x}$, such as linear, nonlinear, spatio-temporal, or other model terms. Distribution regression models featuring response-varying effects are an important special case of this model class. When $\boldsymbol{x} = (x_1, \ldots, x_J) \in \mathbb{R}^J$, a distribution regression model is characterized by partial transformation functions $h_j(y \mid x_j) = \beta_j(y)x_j$ and corresponding interpretable response-varying effects $\beta_j : \mathcal{Z} \to \mathbb{R}$. The analogon of an additive model features partial transformation functions $h_j(y \mid x_j)$, *i.e.* bivariate smooth functions of both y and x_j. These bivariate terms are more complex than the one-dimensional coefficient functions $\beta_j(y)$ but can still be visualized and interpreted. If x_j is more complex, for example, if it describes a spatial location, $h_j(y \mid x_j)$ might be a spatially smooth term that captures unexplained spatial heterogeneity (Hothorn et al. 2014).

Models with transformation function $h(y \mid \boldsymbol{x}) = \sum_{j=1}^{J} h_j(y \mid x_j)$ and potential applications are discussed in Hothorn et al. (2014) and Hothorn et al. (2018). The standard estimation of maximizing the continuously ranked probability score over a discrete grid covering $\mathcal{Z}$ (Foresi and Peracchi 1995; Chernozhukov et al. 2013; Hothorn et al. 2014), potentially with inverse probability of censoring weight adjustment for right censoring (Möst and Hothorn 2015; Garcia et al. 2019), does not allow essentially unstructured transforma-

tion functions $h(y \mid \boldsymbol{x})$, including higher-order interactions, and thus relaxation of the additivity assumption on the scale of the transformation function h. Furthermore, it is computationally inefficient (because the data have to be expanded) and is unable to handle censoring or truncation directly.

This paper addresses these issues by introducing computationally efficient boosted likelihood estimation for unstructured or structured additive conditional transformation functions (Sect. 3) and shift transformation functions (Sect. 4) under all forms of random censoring and truncation for at least ordered responses based on potentially correlated observations.

3 Boosting the likelihood of conditional transformation models

In the following, the conditional transformation function $h(y \mid \boldsymbol{x}) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x})$ is parameterized in terms of basis functions $\boldsymbol{a} : \mathcal{Z} \to \mathbb{R}^P$ of the response and a conditional parameter function $\boldsymbol{\vartheta} : \chi \to \mathbb{R}^P$; the latter function will be estimated.

3.1 Definition of the likelihood

The parameterisation of h implies a conditional cumulative distribution function $\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}))$ and thus a conditional density

$$\begin{aligned} f_Y(y \mid \boldsymbol{\vartheta}(\boldsymbol{x})) &= \frac{\partial F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}))}{\partial y} \\ &= f_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x})) \boldsymbol{a}'(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}) \end{aligned}$$

when $y \in \mathbb{R}$ comes from an absolutely continuous distribution ($\boldsymbol{a}'$ is the derivative of $\boldsymbol{a}$). For discrete $y \in \mathcal{Z} = \{y_1, y_2, \dots\}$, the density function is

$$f_Y(y_k \mid \boldsymbol{\vartheta}(\boldsymbol{x})) = \begin{cases} F_Z(\boldsymbol{a}(y_k)^\top \boldsymbol{\vartheta}(\boldsymbol{x})) & k = 1 \\ F_Z(\boldsymbol{a}(y_k)^\top \boldsymbol{\vartheta}(\boldsymbol{x})) - F_Z(\boldsymbol{a}(y_{k-1})^\top \boldsymbol{\vartheta}(\boldsymbol{x})) & k > 1. \end{cases}$$

There are also other forms of the density, for example, in mixed discrete-continuous distributions. The population optimizer for the conditional parameter function $\boldsymbol{\vartheta}$ is

$$\boldsymbol{\vartheta} := \operatorname*{arg\,max}_{\boldsymbol{\vartheta}: \chi \to \mathbb{R}^P} \int \log\{f_Y[y \mid \boldsymbol{\vartheta}(\boldsymbol{x})]\} \, d\mathbb{P}_{Y,\boldsymbol{X}}(y, \boldsymbol{x}) \quad \text{st. (1)}.$$

Based on N independent observations $(y_i, \boldsymbol{x}_i), i = 1, \dots, N$ from $\mathbb{P}_{Y,\boldsymbol{X}}$, empirical risk minimization with negative log-likelihood loss

$$\hat{\boldsymbol{\vartheta}}_N = \operatorname*{arg\,max}_{\boldsymbol{\vartheta}: \chi \to \mathbb{R}^P} \sum_{i=1}^N \ell_i(\boldsymbol{\vartheta}(\boldsymbol{x}_i)) \quad \text{st. (1)}$$

can be applied to estimate the conditional parameter function $\boldsymbol{\vartheta}$. The log-likelihood contribution $\ell_i : \mathbb{R}^P \to \mathbb{R}$ for the ith observation is given by

$$\ell_i(\boldsymbol{\vartheta}(\boldsymbol{x}_i)) = \begin{cases} \log\{f_Y[y_i \mid \boldsymbol{\vartheta}(\boldsymbol{x}_i)]\} & y_i \in \mathcal{Z} \\ \log\left\{\int_{y_i} f_Y[y \mid \boldsymbol{\vartheta}(\boldsymbol{x}_i)] \, d\mu(y)\right\} & y_i \in \mathcal{C} \setminus \mathcal{Z}, \end{cases} \tag{2}$$

where the first case corresponds to an observation y_i from an absolutely continuous or ordered response and the second case corresponds to the situation where a set or interval was observed (for example, for a left-, right-, or interval-censored observation y_i). Integration is with respect to the measure μ dominating $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$. Details on the likelihood function for models parameterized by $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x})$, including gradients (denoted $\boldsymbol{u}_i$ in Algorithm 1) and Hessians under censoring and truncation, are given in Hothorn et al. (2018).

3.2 Boosting the likelihood

The proposed boosting Algorithm 1 outputs a model of the form

$$\boldsymbol{\vartheta}(\boldsymbol{x}) = \boldsymbol{\vartheta}^{[0]}(\boldsymbol{x}) + \sum_{b=1}^B \boldsymbol{\Gamma}^{[b]} \boldsymbol{b}_{j(b)}(\boldsymbol{x}), \tag{3}$$

with offset term $\boldsymbol{\vartheta}^{[0]}(\boldsymbol{x})$ and $j = 1, \dots, J$ a priori defined basis functions $\boldsymbol{b}_j : \chi \to \mathbb{R}^{P_j}$ of the predictor variables. The function $j(b)$ returns the index of the basis function $\boldsymbol{b}_j$ which was selected in the bth iteration of the algorithm. Each basis may be equipped with an explicit penalty function Pen_j. The corresponding penalty parameter λ_j is chosen such that the degrees of freedom are the same for all J basis functions to facilitate unbiased model selection (Hofner et al. 2011). The number of terms B, selected basis functions $j(b)$, and corresponding coefficient matrices $\boldsymbol{\Gamma}^{[b]} \in \mathbb{R}^{P \times P_{j(b)}}$ are unknowns and are estimated from data. The basis functions $\boldsymbol{b}_j$ may feature unknown parameters. With relatively deep regression trees $\boldsymbol{b}_j$ (where the tree structure is estimated from the data in every boosting iteration and $\boldsymbol{\Gamma}$ are the parameters in each terminal node), model (3) is the sum of B trees and as such is potentially highly unstructured. Similar to GAMLSS-boosting (Mayr et al. 2012), a parameter vector $\boldsymbol{\vartheta}$ is modelled instead of a scalar predictor function. The main difference is that all dimensions of the parameter vector $\boldsymbol{\vartheta}$ are updated simultaneously whereas each dimension is assigned its own predictor function in GAMLSS-boosting.

Algorithm 1 is essentially a multivariate version of L_2 boosting (Bühlmann and Yu 2003) using the negative trans-

Algorithm 1 Boosting CTM Likelihoods

- Start with $b = 0$ and offset $\hat{\boldsymbol{\vartheta}}^{[0]}(\boldsymbol{x}_i)$. Initialize base learners $j = 1, \ldots, J$ via basis functions $\boldsymbol{b}_j$, stepsize $\nu \in (0, 1)$, and $b_{\text{stop}} > 0$
- Iterate
  1. $b \to b + 1$; stop if $b > b_{\text{stop}}$
  2. Compute multivariate negative gradient $\boldsymbol{u}_i^{[b]} \in \mathbb{R}^P$ of $-\ell_i$ evaluated at $\hat{\boldsymbol{\vartheta}}^{[m-1]}(\boldsymbol{x}_i)$

     $$\boldsymbol{u}_i^{[b]} = \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\bigg|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}^{[m-1]}(\boldsymbol{x}_i)}$$

     with components $\boldsymbol{u}_i^{[b]} = \left(u_{i,1}^{[b]}, \ldots, u_{i,P}^{[b]}\right)$ for $i = 1, \ldots, N$
  3. Fit base learners $j = 1, \ldots, J$ by (penalized) multivariate least squares

     $$\hat{\boldsymbol{\Gamma}}_j = \underset{\boldsymbol{\Gamma}^\top = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_P)}{\arg\min} \sum_{p=1}^{P} \left\{ \left[\sum_{i=1}^{N} \left(u_{ip}^{[b]} - \boldsymbol{b}_j(\boldsymbol{x}_i)^\top \boldsymbol{\gamma}_p \right)^2 \right] + \lambda_j \text{Pen}_j(\boldsymbol{\gamma}_p) \right\}$$

  4. Select base learner with minimal quadratic error

     $$j(b) = \underset{j=1,\ldots,J}{\arg\min} \sum_{i=1}^{N} \left(\boldsymbol{u}_i^{[b]} - \hat{\boldsymbol{\Gamma}}_j \boldsymbol{b}_j(\boldsymbol{x}_i) \right)^\top \left(\boldsymbol{u}_i^{[b]} - \hat{\boldsymbol{\Gamma}}_j \boldsymbol{b}_j(\boldsymbol{x}_i) \right)$$

     and $\boldsymbol{\Gamma}^{[b]} = \nu \hat{\boldsymbol{\Gamma}}_{j(b)}$
  5. Update $\hat{\boldsymbol{\vartheta}}^{[b]}(\boldsymbol{x}_i) = \hat{\boldsymbol{\vartheta}}^{[m-1]}(\boldsymbol{x}_i) + \boldsymbol{\Gamma}^{[b]} \boldsymbol{b}_{j(b)}(\boldsymbol{x}_i)$ for $i = 1, \ldots, N$
- Output $\hat{\boldsymbol{\vartheta}}_N(\boldsymbol{x}_i) = \hat{\boldsymbol{\vartheta}}^{[b_{\text{stop}}]}(\boldsymbol{x}_i), i = 1, \ldots, N$

formation log-likelihood (2) as loss function. This choice makes the algorithm agnostic with respect to the scale of the response variable and potential censoring or truncation. The default offset is the unconditional maximum-likelihood estimator $\hat{\boldsymbol{\vartheta}}^{[0]}(\boldsymbol{x}_i) \equiv \hat{\boldsymbol{\vartheta}}_{\text{ML}}$ for $i = 1, \ldots, N$ that maximizes $\sum_{i=1}^{N} \ell_i(\boldsymbol{\vartheta})$. The algorithm is also applicable to the high-dimensional setting where the number of predictor variables exceeds the number of observations N. The number of boosting iterations b_{stop} is a tuning parameter that has to be chosen by the out-of-sample log-likelihood for a validation sample $i = N + 1, \ldots, N + \tilde{N}$

$$\hat{B} = \hat{b}_{\text{stop}} = \underset{b=0,1,\ldots}{\arg\max} \sum_{i=N+1}^{N+\tilde{N}} \ell_i(\hat{\boldsymbol{\vartheta}}^{[b]}(\boldsymbol{x}_i)).$$

Model choice, for example using cross-validation, subsampling, or the bootstrap, can also be implemented conveniently by comparing this out-of-sample log-likelihood of different candidate models.

An additional advantage of this algorithm over boosted continuously ranked probability scores ("CTM-CRPS-boosting", Hothorn et al. 2014) is that computations of tensor products in $\boldsymbol{a}(y)^\top \otimes \boldsymbol{b}_j(\boldsymbol{x})^\top \text{vec}(\boldsymbol{\Gamma}) = \boldsymbol{a}(y)^\top \boldsymbol{\Gamma} \boldsymbol{b}_j(\boldsymbol{x})$ are never explicitly required because the linear array model formulation (*i.e.* the right-hand side of the equation, see Currie et al. 2006, formula 2.5) formula 2.5 is implemented by Algorithm 1. This allows estimation of potentially highly unstructured models by choosing relatively deep multivariate regression trees as basis functions $\boldsymbol{b}$. Moreover, the algorithm does not require expansion of the data set (to size sample size N^2, in the worst case).

3.3 Model interpretation

The partial transformation functions h_j can be obtained from the boosted model (3)

$$\begin{aligned} h_j(y \mid \boldsymbol{x}) &= \boldsymbol{a}(y)^\top \left(\sum_{b:j(b)=j} \boldsymbol{\Gamma}^{(b)} \boldsymbol{b}_{j(b)}(\boldsymbol{x}) \right) \\ &= \boldsymbol{a}(y)^\top \hat{\boldsymbol{\vartheta}}_j(\boldsymbol{x}). \end{aligned}$$

The choice $\boldsymbol{b}_j(\boldsymbol{x}) = x_j$ results in the distribution regression model $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h_Y(y) - \boldsymbol{x}^\top \boldsymbol{\beta}(y))$ with partial transformation functions $h_j(y \mid x_j) = x_j \beta_j(y) = x_j \boldsymbol{a}(y)^\top \hat{\boldsymbol{\vartheta}}_j$ and corresponding response-varying effects $\beta_j(y) = \boldsymbol{a}(y)^\top \hat{\boldsymbol{\vartheta}}_j$. Thus, this boosting procedure can also be used to estimate Cox models with time-varying effects under all forms of random censoring and truncation. Nonlinear effects can be implemented by a B-spline basis $\boldsymbol{b}_j(\boldsymbol{x}) = \boldsymbol{b}_j(x_j)$, and more complex bases allow specification of terms that capture spatio-temporal correlations or other forms of unexplained heterogeneity (Kneib et al. 2009; Hofner et al. 2011). A collection of commonly used basis functions $\boldsymbol{b}$, along with corresponding penalty functions and interpretable model terms, is reviewed in Mayr and Hofner (2018). Specific choices of basis functions underlying the empirical results presented in Sect. 5 are discussed in detail in Hothorn (2019).

4 Boosting the likelihood of shift transformation models

A comparison of conditional transformation models that allow interactions of y and $\boldsymbol{x}$ in the transformation function h with shift transformation models in which these terms are absent can help to identify situations where the simpler models perform as good as or even better than the more complex models. Likelihood boosting for shift transformation models of the form $h(y \mid \boldsymbol{x}) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \beta(\boldsymbol{x})$ is presented as Algorithm (2).

The procedure outputs a model

$$\beta(\boldsymbol{x}) = \beta^{[0]}(\boldsymbol{x}) + \sum_{b=1}^{B} \boldsymbol{\Gamma}^{[b]} \boldsymbol{b}_{j(b)}(\boldsymbol{x}),$$

Algorithm 2 Boosting STM Likelihoods

- Start with offset $\beta^{[0]}(\boldsymbol{x}) = 0$; other choices as in Algorithm 1
- Iterate
  1. $b \rightarrow b + 1$; stop if $b > b_{\text{stop}}$
  2. Update $\hat{\boldsymbol{\vartheta}}^{[b]} = \arg\max_{\boldsymbol{\vartheta} \in \mathbb{R}^P} \sum_{i=1}^N \ell_i(\boldsymbol{\vartheta}, \beta^{[b-1]}(\boldsymbol{x}_i))$ and compute univariate negative gradient $u_i^{[b]} \in \mathbb{R}$ evaluated at $\beta^{[b-1]}(\boldsymbol{x}_i)$
  $$u_i^{[b]} = \left. \frac{\partial \ell_i(\hat{\boldsymbol{\vartheta}}^{[b]}, \beta)}{\partial \beta} \right|_{\beta = \beta^{[b-1]}(\boldsymbol{x}_i)}$$
  3. Fit base learners $j = 1, \ldots, J$ by (penalized) multivariate least squares
  $$\hat{\boldsymbol{\gamma}}_j = \arg\min_{\boldsymbol{\gamma}} \left\{ \left[\sum_{i=1}^N \left(u_{ip}^{[b]} - \boldsymbol{b}_j(\boldsymbol{x}_i)^\top \boldsymbol{\gamma} \right)^2 \right] + \lambda_j \text{Pen}_j(\boldsymbol{\gamma}) \right\}$$
  4. Select base learner with minimal quadratic error
  $$j(b) = \arg\min_{j=1,\ldots,J} \sum_{i=1}^N \left(u_i^{[b]} - \boldsymbol{b}_j(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\gamma}}_j \right)^2$$
  and $\boldsymbol{\gamma}^{[b]} = \nu \hat{\boldsymbol{\gamma}}_{j(b)}$
  5. Update $\beta^{[b]}(\boldsymbol{x}_i) = \beta^{[b-1]}(\boldsymbol{x}_i) + \boldsymbol{b}_{j(b)}(\boldsymbol{x}_i)^\top \boldsymbol{\gamma}^{[b]}$ for $i = 1, \ldots, N$
- Output $\hat{\boldsymbol{\vartheta}}_N = \arg\max_{\boldsymbol{\vartheta} \in \mathbb{R}^P} \sum_{i=1}^N \ell_i(\boldsymbol{\vartheta}, \hat{\beta}_N(\boldsymbol{x}_i))$ and $\hat{\beta}_N(\boldsymbol{x}_i) = \beta^{[b_{\text{stop}}]}(\boldsymbol{x}_i), i = 1, \ldots, N$

with univariate shift function $\beta(\boldsymbol{x}) \in \mathbb{R}$, *i.e.* with $\boldsymbol{\gamma}^\top = \boldsymbol{\Gamma} \in \mathbb{R}^{1 \times P_j}$. In contrast to conditional transformation models, the model term $\beta(\boldsymbol{x})$ does not depend on y and thus shift transformation models are easier to interpret. L_2 boosting in this setup is performed based on log-likelihood contributions $\ell_i(\boldsymbol{\vartheta}, \beta(\boldsymbol{x}_i))$ for $\ell_i : \mathbb{R}^{P+1} \rightarrow \mathbb{R}$ from densities

$$\begin{aligned} f_Y(y \mid \boldsymbol{\vartheta}, \beta(\boldsymbol{x})) &= \frac{\partial F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \beta(\boldsymbol{x}))}{\partial y} \\ &= f_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \beta(\boldsymbol{x})) \boldsymbol{a}'(y)^\top \boldsymbol{\vartheta} \end{aligned}$$

for the absolutely continuous case and

$$\begin{aligned} &f_Y(y_k \mid \boldsymbol{\vartheta}, \beta(\boldsymbol{x})) \\ &= \begin{cases} F_Z(\boldsymbol{a}(y_k)^\top \boldsymbol{\vartheta} - \beta(\boldsymbol{x})) & k = 1 \\ F_Z(\boldsymbol{a}(y_k)^\top \boldsymbol{\vartheta} - \beta(\boldsymbol{x})) - F_Z(\boldsymbol{a}(y_{k-1})^\top \boldsymbol{\vartheta} - \beta(\boldsymbol{x})) & k > 1 \end{cases} \end{aligned}$$

for the discrete case. The core idea is to update the nuisance parameter $\boldsymbol{\vartheta}$ before computing the gradients in every boosting iteration (following Schmid and Hothorn 2008). For discrete proportional odds models, Algorithm (2) is equivalent to the procedure proposed by Schmid et al. (2011). The ability to handle censoring and truncation in the likelihood allows proportional hazards models with potentially very flexible log-hazard ratios $\beta(\boldsymbol{x})$ to be fitted to randomly censored (including left- and interval-censoring) or truncated responses.

5 Empirical evaluation

The rationale for the empirical evaluation of Algorithms 1 and 2 was to demonstrate that interpretable transformation models can be estimated for applications where information about predictive distribution matters (Sect. 5.1) and to investigate the robustness of boosted transformation models under model misspecification (Sect. 5.2).

5.1 Applications

Eight life science applications in which estimation of a predictive distribution is of special interest are listed in Table 1. Four applications are described by a continuous response, two feature an ordered categorical response, and two feature a right-censored response. Except for the Beetle Extinction Risk application, which requires a discrete basis $\boldsymbol{a}$, a Bernstein basis $\boldsymbol{a}$ of order $M = 6$ (for technical details see Hothorn et al. 2018) was used to parameterize the transformation functions. Conditional transformation models (Algorithm 1) with nonlinear (N, using B-splines), linear (L), and tree-based (T, of depth two and thus allowing only two-way interactions) basis functions $\boldsymbol{b}$ as well as shift transformation models (Algorithm 2) using the same bases were evaluated. The performance of these boosted transformation models was compared to the performance of transformation trees and transformation forests (Hothorn and Zeileis 2017). The latter two procedures estimate conditional transformation models of the form $F_Z(\boldsymbol{a}(y)^\top \hat{\boldsymbol{\vartheta}}(\boldsymbol{x}))$, where $\hat{\boldsymbol{\vartheta}}(\boldsymbol{x})$ is obtained either from a single tree (transformation trees) or from a nonlinear interaction function (transformation forest). I hypothesized a priori that transformation trees should perform worst across all applications because this method corresponds to the most simple (but easily interpretable) model. Also, I expected transformation forests to outperform transformation trees and to perform only slightly worse than the best performing boosting procedure because of the high adaptivity of the underlying random forest procedure. My motivation for this experiment was the hope that I would be able to find a simple and interpretable transformation model that outperforms the most complex transformation forests by means of either Algorithm 1 or 2.

Subsampling (with $n = 3/4N$ observations in the learning and $\tilde{n} = 1/4N$ observations in the validation sample) was performed 100 times. Performance of the competitors was assessed by the out-of-sample log-likelihood centered by the out-of-sample log-likelihood of the unconditional transformation model $F_Z(\boldsymbol{a}(y)^\top \hat{\boldsymbol{\vartheta}}_{\text{ML}})$. For a learning sample of size

Table 1 Applications. Eight prediction problems with continuous, ordered categorical (number of categories in parentheses), or right-censored response (per cent censored in parentheses)

Application	Source	N	Y	$\boldsymbol{X}$	F_Z
Beetle Extinction Risk	Seibold et al. (2015)	1025	Ordered (6)	(7, 3)	expit
Birth Weight Prediction	Schild et al. (2008)	150	Continuous	(5, 0)	Φ
Body Fat Mass	Garcia et al. (2005)	71	Continuous	(9, 0)	Φ
CAO/ARO/AIO-04 DFS	Rödel et al. (2015)	1153	Survival (71%)	(3, 15)	MEV
Childhood Malnutrition	Fenske et al. (2011)	24166	Continuous	(6, 14)	Φ
Head Circumference	Fredriks et al. (2000)	7040	Continuous	(1, 0)	Φ
PRO-ACT ALSFRS	Küffner et al. (2015)	1013	Ordered (50)	(43, 16)	expit
PRO-ACT OS	Seibold et al. (2017)	2711	Survival (69%)	(3, 16)	MEV

Number of complete observations N and number of (numeric, categorical) predictor variables $\boldsymbol{X}$. F_Z is the cumulative distribution function of the standard normal (Φ), standard logistic (expit), or standard minimum extreme value distribution (MEV, the inverse complementary log–log)

n and a validation sample $i = n + 1, \ldots, \tilde{n}$, the centered out-of-sample log-likelihood is given by

$$\tilde{n}^{-1} \sum_{i=n+1}^{n+\tilde{n}} \ell_i \left(\hat{\boldsymbol{\vartheta}}^{[\hat{b}_{\text{stop}}]}(\boldsymbol{x}_i) \right) - \tilde{n}^{-1} \sum_{i=n+1}^{n+\tilde{n}} \ell_i(\hat{\boldsymbol{\vartheta}}_{\text{ML}}).$$

Values close to zero indicate that the conditional model did not outperform the unconditional model.

The results presented in Table 2 demonstrate that the best-performing method was always a boosted transformation model. Transformation forests performed only slightly worse than the top model for the Beetle Extinction Risk, Birth Weight Prediction, Body Fat Mass, and Childhood Malnutrition applications. In the remaining four applications, the best boosting procedure outperformed transformation forests substantially. Nonlinear conditional transformation models (N $\boldsymbol{\vartheta}(\boldsymbol{x})$) performed best twice, as did tree-based shift transformation models (T $\beta(\boldsymbol{x})$). Each of the remaining models ranked at the top once. Transformation trees outperformed transformation forests for two applications (Head Cirumference and PRO-ACT ALSFRS) but never performed better than any of the boosted transformation models.

Graphical representations of the distributions of out-of-sample log-likelihoods along with the exact model and algorithm specification and corresponding software implementation are presented for all eight applications in Hothorn (2019).

5.2 Artificial data-generating processes

The response Y was generated conditionally on two groups and one numeric predictor variable $x \in [0, 1]$ following a transformation model of the form

$$\mathbb{P}(Y \leq y \mid \text{Group}, x) = \Phi(h(y \mid \text{Group}, x)),$$

where the conditional transformation function $h(y \mid \text{Group}, x) = \Phi^{-1}(\mathbb{P}(Y \leq y \mid \text{Group}, x))$ for four data-generating processes (DGPs) is given in Table 3. The model labelled "Linear $\beta(\boldsymbol{x})$" is a shift transformation model with a main effect of group, a linear main effect of x, and a corresponding linear interaction effect. The linear main and interaction effects of x are replaced by nonlinear effects (a scaled sin function) of x in the shift transformation model "Nonlinear $\beta(\boldsymbol{x})$". The extension to response-varying main and interaction effects defines the distribution regression model "Linear $\boldsymbol{\vartheta}(\boldsymbol{x})$" and the conditional transformation model "Nonlinear $\boldsymbol{\vartheta}(\boldsymbol{x})$". The coefficients of the terms introduced in Table 3 are given in Table 4. Details of the implementation of these DGPs are explained in Hothorn (2019). The conditional densities associated with the four DGPs are shown in Fig. 1.

Models were evaluated by out-of-sample log-likelihoods centered by the out-of-sample log-likelihood of the true DGP (for test samples of size $\tilde{N} = 2000$)

$$\tilde{N}^{-1} \sum_{i=N+1}^{N+\tilde{N}} \ell_i \left(\hat{\boldsymbol{\vartheta}}^{[\hat{b}_{\text{stop}}]}(\boldsymbol{x}_i) \right) - \tilde{N}^{-1} \sum_{i=N+1}^{N+\tilde{N}} \ell_i(\boldsymbol{\vartheta}_{\text{True}}),$$

where $\boldsymbol{\vartheta}_{\text{True}}$ is as given in Tables 3 and 4.

In Part A of this simulation, nonlinear (N, using B-splines), linear (L), and tree-based (T, of depth six, which allows higher-order interactions) basis functions $\boldsymbol{b}$ for shift transformation models, *i.e.* models for $\beta(\boldsymbol{x})$, and for conditional transformation models, *i.e.* models for $\boldsymbol{\vartheta}(\boldsymbol{x})$, were evaluated for sample sizes $N = 75, 150, 300$ under correctly specified models; this means that models were fitted using the correct distribution function $F_Z = \Phi$, the correct order $M = 6$ of $\boldsymbol{a}$, no uninformative predictor variables, and the correct basis functions. Both the linear and nonlinear models were fitted with basis functions representing a main effect of group, a main effect of x, and a corresponding interaction effect, whereas trees had to learn this structure from the data. In Part B, these models were evaluated under model misspecification, *i.e.* using the incorrect distribution function

Table 2 Applications. Median out-of-sample log-likelihoods (centered by out-of-sample log-likelihoods of the corresponding unconditional model) from 100 subsampled divisions into learning and test samples

	Parm. $\boldsymbol{b}(\boldsymbol{x})$	Conditional Transformation Model $\vartheta(\boldsymbol{x})$			Shift Transformation Model $\beta(\boldsymbol{x})$			Transformation	
		N	L	T	N	L	T	Tree	Forest
Beetles Exctinction Risk		0.27	0.26	0.18	0.30	0.29	**0.33**	0.03	0.30
		(0.15, 0.28)	(0.16, 0.28)	(0.11, 0.29)	(0.29, 0.32)	(0.27, 0.31)	(0.30, 0.35)	(−0.13, 0.12)	(0.28, 0.33)
Birth Weight Prediction		1.47	1.43	1.25	**1.50**	1.50	1.37	0.39	1.32
		(1.31, 1.55)	(1.19, 1.51)	(1.13, 1.30)	(1.32, 1.57)	(1.31, 1.56)	(1.28, 1.46)	(−0.17, 0.69)	(1.21, 1.37)
Body Fat Mass		**1.17**	1.08	0.76	0.12	0.12	0.24	0.08	1.03
		(1.05, 1.27)	(0.96, 1.20)	(0.58, 0.87)	(0.11, 0.14)	(0.11, 0.14)	(0.22, 0.27)	(−0.61, 0.49)	(0.95, 1.09)
CAO/ARO/AIO-04 DFS		0.00	0.01	0.00	0.02	**0.02**	0.01	−0.00	−0.01
		(−0.00, 0.01)	(−0.00, 0.01)	(−0.00, 0.00)	(0.01, 0.03)	(0.01, 0.03)	(0.01,0.02)	(−0.01, 0.00)	(−0.02, −0.00)
Childhood Malnutrition		0.05	0.05	**0.15**	0.13	0.12	0.14	0.12	0.15
		(0.05, 0.05)	(0.04, 0.05)	(0.14, 0.15)	(0.13, 0.14)	(0.11, 0.12)	(0.13, 0.14)	(0.12, 0.13)	(0.14, 0.15)
Head Circumference		**1.09**	0.96	1.09	1.01	0.96	1.08	1.06	0.99
		(1.08, 1.11)	(0.94, 0.97)	(1.07, 1.10)	(1.00, 1.02)	(0.95, 0.97)	(1.07, 1.10)	(1.04, 1.08)	(0.92, 1.02)
PRO-ACT ALSFRS		0.52	**0.52**	0.48	0.51	0.49	0.49	0.38	0.32
		(0.50, 0.55)	(0.49, 0.55)	(0.46, 0.50)	(0.47, 0.54)	(0.46, 0.52)	(0.46, 0.53)	(0.33, 0.43)	(0.30, 0.33)
PRO-ACT OS		0.04	0.04	0.04	0.05	0.05	**0.06**	0.02	0.04
		(0.03, 0.05)	(0.03, 0.05)	(0.02, 0.05)	(0.04, 0.06)	(0.04, 0.06)	(0.05, 0.06)	(−0.02, 0.04)	(−0.01, 0.05)

Positive values indicate a superior performance of the conditional model, taking predictor variables into account, over the unconditional model. Interquartile ranges are given in parentheses. Boosting CTM Likelihoods (parameter $\vartheta(\boldsymbol{x})$) with nonlinear (N), linear (L), and tree-based (T, depth two) basis functions $\boldsymbol{b}$ as well as Boosting STM Likelihoods (parameter $\beta(\boldsymbol{x})$) with the same basis functions were compared to transformation trees and transformation forests. The result of the best-performing model is printed in boldface

Table 3 Artificial data-generating processes (DGPs). Description of four simulation models

DGP	$\Phi^{-1}(\mathbb{P}(Y \leq y \mid \text{Group } 1, x))$	$\Phi^{-1}(\mathbb{P}(Y \leq y \mid \text{Group } 2, x))$
Linear $\beta(\boldsymbol{x})$	$h_Y(y) - 2x$	$h_Y(y) + 2 - x$
Nonlinear $\beta(\boldsymbol{x})$	$h_Y(y) - 2g(x)$	$h_Y(y) + 2 - g(x)$
Linear $\boldsymbol{\vartheta}(\boldsymbol{x})$	$h_Y(y) - \beta_1(y) - \beta_2(y)x$	$h_Y(y) + \beta_1(y) - (\beta_2(y) + \beta_3(y))x$
Nonlinear $\boldsymbol{\vartheta}(\boldsymbol{x})$	$h_Y(y) - \beta_1(y) - \beta_2(y)g(x)$	$h_Y(y) + \beta_1(y) - (\beta_2(y) + \beta_3(y))g(x)$

$g(x) = \sin(2\pi x)(1 + x)$, h_Y, β_1, β_2, β_3 are Bernstein polynomials of order $M = 6$ on the interval $(-4, 6)$ (see Hothorn et al. 2018)

Table 4 Artificial data-generating Processes (DGPs). Coefficients of baseline transformation h_Y and response-varying effects β_1, β_2, and β_3

	ϑ_1	ϑ_2	ϑ_3	ϑ_4	ϑ_5	ϑ_6	ϑ_7
$h_Y(y)$	− 4.000	− 0.601	1.065	1.000	2.667	4.333	6.000
$\beta_1(y)$	0.000	− 0.518	− 1.000	− 1.414	− 1.732	− 1.932	− 2.000
$\beta_2(y)$	0.000	0.816	1.155	1.414	1.633	1.826	2.000
$\beta_3(y)$	0.000	− 0.259	− 0.500	− 0.707	− 0.866	− 0.966	− 1.000

All functions are parameterized as a Bernstein polynomial $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ of order $M = 6$ on the interval $(-4, 6)$ (Hothorn et al. 2018), whose coefficients $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_7)$ are given in this table

F_Z = expit (standard logistic distribution) or F_Z = MEV (standard minimum extreme value distribution), a too large dimension of the Bernstein basis $\boldsymbol{a}$ ($M = 12$), or $J^+ = 5, 25$ additional uninformative uniform predictor variables. The same "correct" basis functions as in Part A were used in Part B. I hypothesized a priori that models exactly matching the DGP would perform best and that tree-based boosting would outperform boosting with linear basis functions in nonlinear problems. Under misspecification, I expected the performance of all models to decrease, but this general ranking to persist.

The results for Part A presented in the top three rows of Table 5 show that the model corresponding to the underlying DGP was associated with the largest median out-of-sample log-likelihood. For linear DGPs, the performance of boosted models with nonlinear basis functions was only slightly inferior to the performance of boosted models with linear basis functions, while tree-based boosting performed substantially worse in this situation. By contrast, the signal in nonlinear DGPs was captured relatively well by tree-based boosting, whereas linear basis functions were not able to recover this signal. This shows that tree-based boosting was able to adapt to the underlying nonlinear interaction signal in the two nonlinear simulation models "Nonlinear $\beta(\boldsymbol{x})$" and "Nonlinear $\boldsymbol{\vartheta}(\boldsymbol{x})$".

The out-of-sample log-likelihoods for misspecified models presented in Table 5, Part B for $F_Z = \Phi$, follow this general pattern in that the model corresponding to the DGP performed best and tree-based boosting outperformed boosting with linear basis functions on nonlinear problems. In only two cases, which were characterized by small samples, did a linear model for $\boldsymbol{\vartheta}(\boldsymbol{x})$ outperform a true linear model for $\beta(\boldsymbol{x})$ or vice versa. More frequently, the too complex nonlinear model for $\boldsymbol{\vartheta}(\boldsymbol{x})$ outperformed the nonlinear model for $\beta(\boldsymbol{x})$ slightly. Overall, Algorithms 1 and 2 seemed to be robust against overly complex basis functions $\boldsymbol{a}$ and additional noninformative predictor variables.

This was also true under a misspecified distribution function F_Z = expit for linear shift transformation models "Linear $\beta(\boldsymbol{x})$". More severe deviations occurred when an incorrect F_Z = expit was used for model specification in Algorithms 1 and 2 under the nonlinear shift transformation model "Nonlinear $\beta(\boldsymbol{x})$", distribution regression model "Linear $\boldsymbol{\vartheta}(\boldsymbol{x})$", and conditional transformation model "Nonlinear $\boldsymbol{\vartheta}(\boldsymbol{x})$". The absolute differences in the corresponding out-of-sample log-likelihoods were, however, marginal in most of these cases.

When the asymmetric standard minimum value distribution was used (F_Z = MEV), the distortions were more pronounced. The general pattern observed for F_Z = expit was the same, but the centered out-of-sample log-likelihoods seemed in general smaller in this setup. Visualizations of the distributions underlying the figures in Table 5 are presented in Hothorn (2019).

6 Discussion

Models defined in terms of simple linear transformation functions up to models featuring unstructured complex transformation functions can be specified, estimated, evaluated, and compared in the unified computational framework of Algorithms 1 and 2. Data analysts are no longer limited in their freedom to define and estimate transformation models, because the strong ties between models of a certain complexity and a tailored estimation procedure (such as CTM-CPRS-boosting for additive or transformation forests

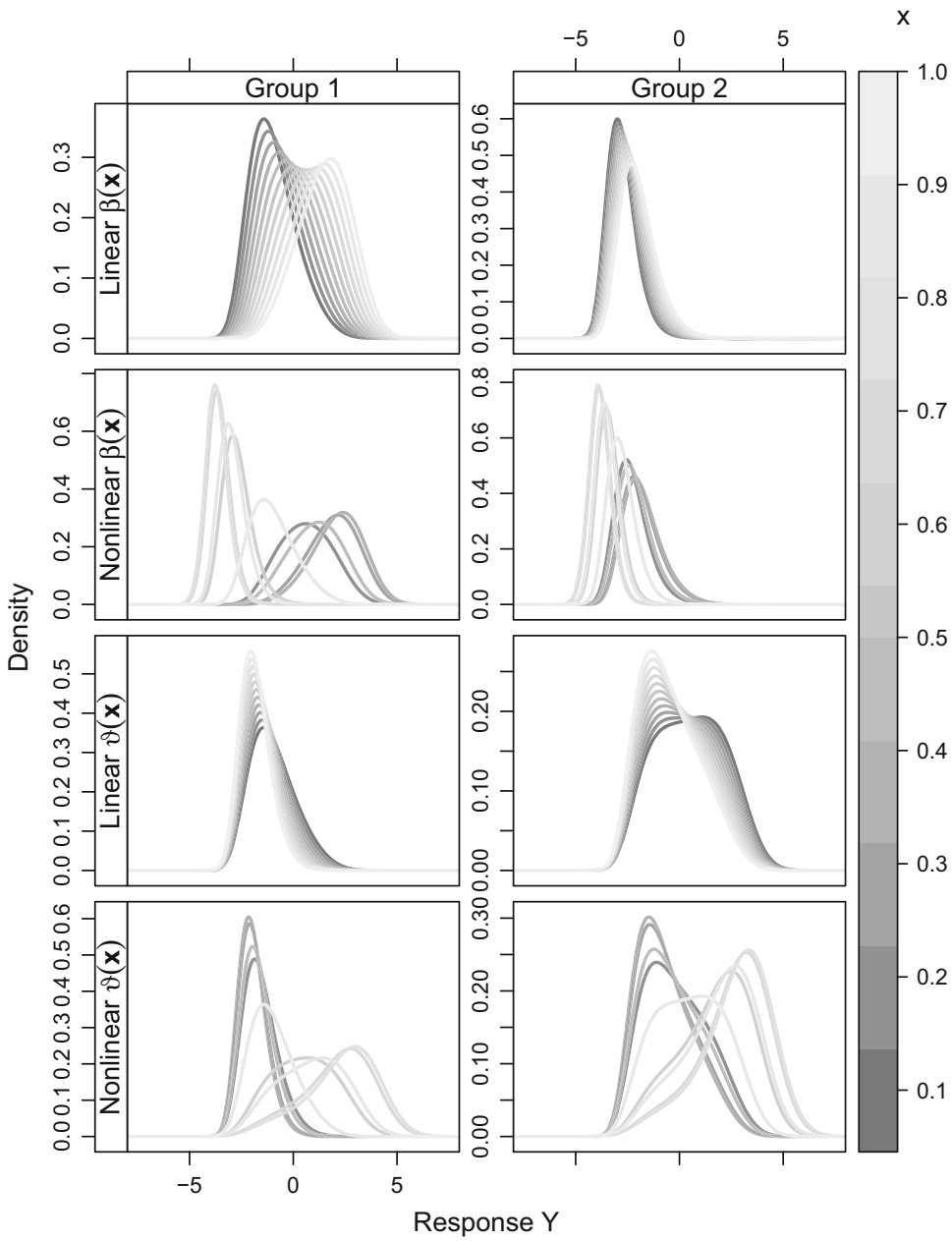

Fig. 1 Artificial Data-generating Processes (DGPs). Conditional densities given two groups (left and right panel) and $x \in [0, 1]$ (gray color coding) for four different data-generating processes

for interaction models) can be cut with the boosting algorithms presented here.

For model specification, the choice of F_Z is important in simple shift transformation models because it affects the interpretation of model parameters (log-odds ratios vs. log-hazard ratios, for example). In more complex models, a direct interpretation of parameters is hardly possible, and the estimated conditional distribution functions are insensitive to the choice of F_Z (Hothorn et al. 2018). However, one could use Algorithm 1 to estimate an unstructured log-hazard function $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}) + \log(\boldsymbol{a}'(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}))$ in the model $\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = 1 - \exp(-\exp(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}))$, with $\boldsymbol{\vartheta}(\boldsymbol{x})$ being, for example, the sum of B deep trees. The log-likelihood risk function employed here, which is also able to handle time-varying covariates through appropriate truncation, avoids the technical obstacles reported by Lee and Chen (2018) when defining an appropriate nonparametric risk function for boosting in a class of models for conditional log-hazard functions.

In contrast to quantile regression, where separate models for each quantile are fitted, likelihood boosting for transformation models estimates conditional distribution functions simultaneously for all quantiles. It is interesting to note that a recently suggested Bayesian approach to simultaneous linear quantile regression (Yang and Tokdar 2017) maximizes a log-likelihood obtained from a numerical inver-

Table 5 Artificial data-generating processes (DGPs). Median out-of-sample log-likelihoods (centered by out-of-sample log-likelihoods of the true model) based on 100 simulation runs of varying sample sizes $N = 75, 150, 300$, number of additional noninformative predictor variables $J^+ = 0, 5, 25$, choices of F_Z (Φ = standard normal, expit = standard logistic, and MEV = standard minimum extreme value), and choices of the Bernstein polynomial order $M = 6, 12$ for four different DGPs and Boosting CTM Likelihoods (parameter $\boldsymbol{\vartheta}(\boldsymbol{x})$) with nonlinear (N), linear (L), and tree-based (T, depth two) basis functions $\boldsymbol{b}$ as well as Boosting STM Likelihoods (parameter $\beta(\boldsymbol{x})$) with the same basis functions

				DGP	Linear $\beta(\boldsymbol{x})$						Nonlinear $\beta(\boldsymbol{x})$						Linear $\boldsymbol{\vartheta}(\boldsymbol{x})$						Nonlinear $\boldsymbol{\vartheta}(\boldsymbol{x})$					
				Parm.	$\boldsymbol{\vartheta}(\boldsymbol{x})$			$\beta(\boldsymbol{x})$			$\boldsymbol{\vartheta}(\boldsymbol{x})$			$\beta(\boldsymbol{x})$			$\boldsymbol{\vartheta}(\boldsymbol{x})$			$\beta(\boldsymbol{x})$			$\boldsymbol{\vartheta}(\boldsymbol{x})$			$\beta(\boldsymbol{x})$		
F_Z	M	J^+	N	$\boldsymbol{b}(\boldsymbol{x})$	N	L	T	N	L	T	N	L	T	N	L	T	N	L	T	N	L	T	N	L	T	N	L	T
Φ	6	0	75	Part A	-135	-90	-373	-86	**-67**	-389	**-189**	-725	-414	-202	-799	-504	-125	**-87**	-217	-130	-106	-284	**-143**	-383	-348	-183	-422	-423
			150		-86	-51	-301	-47	**-34**	-280	-116	-694	-322	**-87**	-764	-364	-68	**-51**	-171	-91	-86	-228	**-93**	-356	-288	-126	-386	-329
			300		-55	-32	-230	-22	**-17**	-194	-74	-673	-250	**-36**	-746	-252	-51	**-32**	-132	-71	-68	-172	**-59**	-336	-224	-102	-363	-244
Φ		5	75		-216	-149	-409	-128	**-95**	-491	**-266**	-768	-624	-295	-824	-746	-163	**-115**	-224	-138	-133	-293	**-212**	-418	-414	-250	-433	-539
			150		-142	-84	-315	-69	**-53**	-367	-185	-717	-466	**-157**	-782	-534	-94	**-67**	-176	-101	-93	-267	**-134**	-368	-333	-163	-390	-422
			300		-86	-56	-254	-40	**-27**	-259	-125	-686	-370	**-65**	-752	-358	-63	**-46**	-139	-74	-73	-212	**-90**	-346	-266	-117	-373	-325
		25	75		-258	-203	-405	-138	**-120**	-523	**-352**	-816	-688	-406	-862	-897	-222	-162	-219	-167	**-148**	-333	**-287**	-477	-443	-322	-442	-609
			150		-175	-129	-344	-83	**-70**	-419	**-235**	-749	-566	-249	-803	-705	-130	**-97**	-179	-107	-101	-291	**-170**	-400	-355	-188	-404	-511
			300		-114	-81	-288	-53	**-42**	-307	-156	-701	-441	**-144**	-769	-481	-81	**-63**	-142	-81	-78	-233	**-117**	-360	-304	-142	-380	-384
	12	0	75		-170	-131	-394	-107	**-95**	-482	**-228**	-743	-468	-242	-840	-567	-123	**-99**	-240	-172	-150	-338	**-175**	-421	-364	-233	-459	-471
			150		-103	-70	-312	-63	**-45**	-315	-146	-706	-381	**-143**	-772	-384	-72	**-51**	-175	-98	-96	-250	**-104**	-367	-281	-143	-397	-363
			300		-65	-40	-245	-29	**-22**	-199	-84	-678	-306	**-46**	-751	-259	-44	**-30**	-129	-78	-71	-184	**-62**	-343	-230	-109	-376	-260
		5	75		-247	-186	-412	-154	**-126**	-601	**-315**	-798	-704	-337	-879	-844	-197	**-147**	-243	-192	-170	-334	**-250**	-457	-439	-284	-475	-646
			150		-153	-108	-345	-76	**-67**	-402	-213	-728	-536	**-195**	-803	-590	-99	**-70**	-177	-109	-101	-273	**-137**	-377	-358	-178	-412	-450
			300		-97	-68	-260	-47	**-32**	-288	-141	-691	-403	**-101**	-760	-371	-59	**-43**	-133	-82	-76	-224	**-86**	-349	-271	-121	-379	-336
		25	75		-309	-241	-433	-176	**-147**	-630	**-387**	-846	-765	-496	-918	-990	-231	**-186**	-256	-216	-188	-356	**-318**	-533	-495	-385	-503	-666
			150		-186	-145	-362	-98	**-82**	-456	**-252**	-753	-612	-282	-827	-780	-129	**-94**	-188	-127	-117	-301	**-181**	-410	-374	-218	-414	-534
			300		-123	-94	-285	-57	**-46**	-331	-175	-702	-450	**-154**	-768	-516	-80	**-61**	-142	-87	-80	-245	**-112**	-364	-294	-144	-385	-400
expit	6	0	75		-165	-107	-415	-98	**-73**	-340	-198	-746	-461	**-197**	-813	-454	-132	**-89**	-214	-122	-101	-262	-168	-386	-365	**-151**	-394	-355
			150		-102	-67	-347	-57	**-44**	-264	**-125**	-724	-358	-131	-776	-346	-79	**-57**	-169	-83	-83	-207	-104	-357	-289	**-103**	-367	-294
			300		-72	-50	-255	-35	**-31**	-186	**-86**	-698	-274	-95	-762	-256	-64	**-40**	-134	-67	-64	-161	**-66**	-342	-235	-85	-357	-217
		5	75		-270	-164	-431	-132	**-104**	-379	**-300**	-803	-639	-304	-846	-646	-185	-130	-217	-133	**-115**	-285	-240	-425	-406	**-204**	-416	-466
			150		-168	-107	-359	-73	**-60**	-299	-200	-742	-496	**-173**	-802	-469	-105	**-74**	-177	-95	-90	-235	-155	-377	-344	**-137**	-387	-369
			300		-109	-71	-300	-50	**-40**	-221	-139	-713	-386	**-105**	-764	-329	-80	**-56**	-141	-71	-68	-186	-107	-353	-272	**-98**	-365	-264
		25	75		-321	-245	-447	-158	**-135**	-416	**-390**	-858	-699	-421	-868	-781	-229	-172	-239	-154	**-133**	-305	-328	-477	-451	**-316**	-437	-518
			150		-213	-142	-386	-96	**-82**	-319	-261	-780	-574	**-251**	-827	-595	-136	-103	-186	-103	**-98**	-260	-210	-405	-357	**-174**	-388	-431
			300		-141	-96	-329	-67	**-54**	-248	-179	-730	-465	**-171**	-784	-402	-96	**-72**	-148	-77	-74	-212	-137	-369	-287	**-120**	-369	-320
	12	0	75	Part B	-198	-131	-444	-110	**-92**	-387	**-233**	-766	-490	-251	-839	-495	-138	**-108**	-231	-138	-121	-276	**-186**	-417	-368	-206	-425	-388
			150		-121	-78	-341	-71	**-58**	-268	**-151**	-720	-419	-162	-785	-365	-83	**-58**	-174	-98	-87	-214	**-122**	-373	-292	-126	-383	-297
			300		-80	-48	-277	-43	**-37**	-190	**-94**	-697	-325	-104	-764	-267	-57	**-39**	-133	-71	-66	-172	**-78**	-349	-234	-97	-368	-222
		5	75		-284	-194	-455	-147	**-124**	-417	-372	-814	-720	**-337**	-861	-690	-200	**-137**	-230	-156	-139	-296	-258	-457	-428	**-240**	-434	-478
			150		-185	-127	-388	-83	**-69**	-306	-240	-751	-559	**-201**	-804	-478	-110	**-79**	-180	-104	-94	-238	-157	-380	-348	**-153**	-389	-371
			300		-119	-82	-300	-55	**-45**	-235	-160	-717	-414	**-120**	-770	-335	-72	**-53**	-142	-75	-74	-183	**-109**	-356	-267	-109	-372	-268
		25	75		-322	-257	-460	-180	**-142**	-437	**-444**	-860	-763	-462	-906	-840	-230	-172	-241	-184	**-158**	-325	**-324**	-515	-480	-375	-458	-573
			150		-228	-167	-402	-110	**-88**	-339	**-276**	-776	-630	-281	-834	-636	-133	-105	-189	-121	**-99**	-259	-201	-409	-373	**-193**	-402	-435
			300		-150	-113	-323	-67	**-62**	-259	-202	-723	-470	**-177**	-791	-420	-88	**-69**	-146	-81	-75	-207	-134	-365	-293	**-126**	-377	-328
MEV	6	0	75		-323	-232	-476	-208	**-191**	-723	-391	-833	-602	**-359**	-868	-828	-126	-87	-251	-88	**-78**	-330	**-184**	-423	-441	-190	-426	-647
			150		-246	-164	-418	-140	**-134**	-520	-297	-772	-533	**-250**	-826	-600	-57	-41	-179	-48	**-38**	-273	-117	-377	-340	**-109**	-379	-516
			300		-195	**-61**	-347	-123	-116	-365	-237	-737	-446	**-164**	-801	-436	-40	-30	-138	-26	**-24**	-184	-74	-359	-275	**-65**	-360	-346
		5	75		-412	-342	-506	-234	**-212**	-870	-481	-900	-783	**-468**	-914	-1142	-182	-124	-241	-144	**-94**	-360	-285	-463	-513	**-281**	-458	-724
			150		-305	-245	-422	-169	**-151**	-699	-361	-813	-656	**-309**	-847	-857	-84	-61	-179	-53	**-44**	-324	-174	-406	-410	**-154**	-404	-617
			300		-244	-165	-363	-134	**-122**	-547	-294	-774	-539	**-210**	-816	-586	-49	-38	-135	-32	**-29**	-261	-112	-375	-326	**-94**	-376	-456
		25	75		-460	-414	-505	-252	**-226**	-997	-532	-982	-874	**-509**	-908	-1299	-273	-191	-242	-162	**-132**	-367	-417	-566	-523	**-401**	-510	-764
			150		-339	-302	-433	-179	**-158**	-772	-435	-868	-736	**-368**	-858	-1038	-134	-103	-194	-75	**-65**	-331	-226	-447	-445	**-205**	-425	-682
			300		-285	-248	-381	-140	**-133**	-623	-339	-801	-618	**-284**	-823	-788	-67	-51	-148	-36	**-31**	-285	-149	-392	-352	**-139**	-389	-569
	12	0	75		-368	-326	-531	-290	**-214**	-980	**-447**	-843	-684	-681	-932	-1060	-153	**-111**	-274	-212	-168	-397	**-226**	-472	-470	-317	-469	-747
			150		-255	-168	-431	-159	**-150**	-665	-319	-767	-550	**-300**	-835	-633	-65	**-46**	-184	-61	-60	-314	**-131**	-392	-368	-143	-402	-505
			300		-186	**-59**	-348	-124	-116	-419	-238	-724	-460	**-194**	-794	-437	-40	-28	-127	-26	**-21**	-208	**-78**	-355	-277	-81	-365	-346
		5	75		-445	-397	-555	**-289**	-372	-1034	**-539**	-931	-896	-614	-934	-1498	-227	-175	-250	**-156**	-163	-380	**-370**	-530	-537	-377	-521	-776
			150		-313	-262	-464	-180	**-172**	-841	-398	-810	-705	**-328**	-864	-1046	-91	-63	-190	-72	**-62**	-335	-177	-418	-419	**-172**	-412	-666
			300		-242	-184	-365	-140	**-131**	-609	-318	-763	-542	**-226**	-801	-617	-58	-43	-132	-33	**-29**	-253	**-110**	-374	-325	-111	-369	-482
		25	75		-508	-474	-599	-389	**-368**	-1023	**-621**	-984	-972	-660	-974	-1510	-280	-248	-267	-223	**-197**	-397	-489	-626	-581	**-444**	-571	-787
			150		-365	-325	-471	-186	**-181**	-902	**-434**	-848	-794	-454	-856	-1221	-149	-106	-198	**-87**	-89	-338	-254	-468	-443	**-237**	-429	-694
			300		-286	-256	-387	-150	**-139**	-712	-360	-786	-600	**-301**	-822	-877	-82	-60	-143	-43	**-37**	-294	-161	-400	-352	**-141**	-388	-610

Part A compares correctly specified models ($F_Z = \Phi$, $J^+ = 0$, $M = 6$) for three different sample sizes. Part B compares the performance under various forms of misspecification. The result of the best-performing model is printed in boldface

sion of the quantile function instead of using the traditional check risk minimization. In light of this approach, it seems computationally attractive to model the distribution function in the distribution regression model $F_{Y|X=x}(y \mid X = x) = F_Z(h_Y(y) - x^\top \beta(y))$ rather than the quantile function in a quantile regression model $Q_{Y|X=x}(\tau \mid X = x) = \alpha(\tau) + x^\top \delta(\tau)$ of the same complexity ($\tau \in [0, 1]$; α and δ being the probability-varying intercept and coefficient functions, respectively). Bayesian inference for the corresponding model parameters in conditional transformation models is, however, still under development (Mitrodima and Griffin 2017).

Computational details

A reference implementation of transformation boosting machines (Algorithms 1 and 2) is available in the **tbm** package (Hothorn 2019). Analyses of all applications and simulation results can be reproduced in the dynamic document Hothorn (2019). All computations were performed using R version 3.5.2 (R Core Team 2018).

References

Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. Ann. Stat. **47**(2), 1148–1178 (2019). https://doi.org/10.1214/18-AOS1709

Box, G.E.P., Cox, D.R.: An analysis of transformations. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **26**(2), 211–252 (1964)

Bühlmann, P., Hothorn, T.: Boosting algorithms: regularization, prediction and model fitting. Stat. Sci. **22**(4), 477–505 (2007). https://doi.org/10.1214/07-STS242. with discussion

Bühlmann, P., Yu, B.: Boosting with the L_2 loss: regression and classification. J. Am. Stat. Assoc. **98**(462), 324–339 (2003). https://doi.org/10.1198/016214503000125

Cabrera, B.L., Schulz, F.: Forecasting generalized quantiles of electricity demand: a functional data approach. J. Am. Stat. Assoc. **112**(517), 127–136 (2017). https://doi.org/10.1080/01621459.2016.1219259

Chernozhukov, V., Fernández-Val, I., Melly, B.: Inference on counterfactual distributions. Econometrica **81**(6), 2205–2268 (2013). https://doi.org/10.3982/ECTA10582

Currie, I.D., Durban, M., Eilers, P.H.C.: Generalized linear array models with applications to multidimensional smoothing. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **68**(2), 259–280 (2006). https://doi.org/10.1111/j.1467-9868.2006.00543.x

Fenske, N., Kneib, T., Hothorn, T.: Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. J. Am. Stat. Assoc. **106**(494), 494–510 (2011). https://doi.org/10.1198/jasa.2011.ap09272

Foresi, S., Peracchi, F.: The conditional distribution of excess returns: an empirical analysis. J. Am. Stat. Assoc. **90**(430), 451–466 (1995). https://doi.org/10.1080/01621459.1995.10476537

Fredriks, A.M., van Buuren, S., Burgmeijer, R.J.F., Meulmeester, J.F., Beuker, R.J., Brugman, E., Roede, M.J., Verloove-Vanhorick, S.P., Wit, J.: Continuing positive secular growth change in the Netherlands 1955–1997. Pediatr. Res. **47**(3), 316–323 (2000)

Friedman, J.H., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (with discussion). Ann. Stat. **28**, 337–407 (2000). https://doi.org/10.1214/aos/1016218223

Garcia, A.L., Wagner, K., Hothorn, T., Koebnick, C., Zunft, H.J.F., Trippo, U.: Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths. Obesity **13**(3), 626–634 (2005). https://doi.org/10.1038/oby.2005.67

Garcia, T.P., Marder, K., Wang, Y.: Time-varying proportional odds model for mega-analysis of clustered event times. Biostatistics **20**(1), 129–146 (2019). https://doi.org/10.1093/biostatistics/kxx065

Gneiting, T., Katzfuss, M.: Probabilistic forecasting. Annu. Rev. Stat. Its Appl. **1**(1), 125–151 (2014). https://doi.org/10.1146/annurev-statistics-062713-085831

Hofner, B., Hothorn, T., Kneib, T., Schmid, M.: A framework for unbiased model selection based on boosting. J. Comput. Graph. Stat. **20**(4), 956–971 (2011). https://doi.org/10.1198/jcgs.2011.09220

Hothorn, T.: tbm: Transformation Boosting Machines. R package and vignette version 0.3-0 (2019). http://CRAN.R-project.org/package=tbm

Hothorn, T., Zeileis, A.: Transformation forests. Tech. rep. v2, https://arxiv.org/abs/1701.02110 (2017)

Hothorn, T., Kneib, T., Bühlmann, P.: Conditional transformation models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **76**(1), 3–27 (2014). https://doi.org/10.1111/rssb.12017

Hothorn, T., Möst, L., Bühlmann, P.: Most likely transformations. Scand. J. Stat. **45**(1), 110–134 (2018). https://doi.org/10.1111/sjos.12291

Kneib, T., Hothorn, T., Tutz, G.: Variable selection and model choice in geoadditive regression models. Biometrics **65**(2), 626–634 (2009). https://doi.org/10.1111/j.1541-0420.2008.01112.x

Koenker, R.: Quantile Regression. Economic Society Monographs. Cambridge University Press, New York (2005)

Kooperberg, C., Stone, C.J., Truong, Y.K.: Hazard regression. J. Am. Stat. Assoc. **90**(429), 78–94 (1995). https://doi.org/10.1080/01621459.1995.10476491

Küffner, R., Zach, N., Norel, R., Hawe, J., Schoenfeld, D., Wang, L., Li, G., Fang, L., Mackey, L., Hardiman, O., Cudkowicz, M., Sherman, A., Ertaylan, G., Grosse-Wentrup, M., Hothorn, T., van Ligtenberg, J., Macke, J.H., Meyer, T., Schölkopf, B., Tran, L., Vaughan, R., Stolovitzky, G., Leitner, M.L.: Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nat. Biotechnol. **33**, 51–57 (2015). https://doi.org/10.1038/nbt.3051

Lee, D.K.K., Chen, N.: Boosting hazard regression with time-varying covariates. Tech. rep. v3, https://arxiv.org/abs/1701.07926 (2018)

Leorato, S., Peracchi, F.: Comparing distribution and quantile regression. Tech. Rep. 1511, Einaudi Institute for Economics and Finance, Rome, Italy (2015). https://ideas.repec.org/p/eie/wpaper/1511.html. Accessed 24 Nov 2018

Li, Q., Racine, J.S.: Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. J. Bus. Econ. Stat. **26**(4), 423–434 (2008). https://doi.org/10.1198/073500107000000250

Lu, W., Li, L.: Boosting method for nonlinear transformation models with censored survival data. Biostatistics **9**(4), 658–667 (2008). https://doi.org/10.1093/biostatistics/kxn005

Mayr, A., Hofner, B.: Boosting for statistical modelling—a non-technical introduction. Stat. Model. **18**(3–4), 365–384 (2018). https://doi.org/10.1177/1471082X17748086

Mayr, A., Fenske, N., Hofner, B., Kneib, T., Schmid, M.: GAMLSS for high-dimensional data—a flexible approach based on boosting. J. R. Stat. Soc. Ser. C (Appl. Stat.) **61**(3), 403–427 (2012). https://doi.org/10.1111/j.1467-9876.2011.01033.x

Meinshausen, N.: Quantile regression forests. J. Mach. Learn. Res. **7**, 983–999 (2006). http://jmlr.org/papers/v7/meinshausen06a.html

Mitrodima, G., Griffin, J.E.: A Bayesian quantile time series model for asset returns. Tech. rep., SSRN, https://doi.org/10.2139/ssrn.3050989 (2017)

Möst, L., Hothorn, T.: Conditional transformation models for survivor function estimation. Int. J. Biostat. **11**(1), 23–50 (2015). https://doi.org/10.1515/ijb-2014-0006

Pratola, M., Chipman, H., George, E.I., McCulloch, R.: Heteroscedastic bart using multiplicative regression trees. Tech. rep. v1, http://arxiv.org/abs/1709.07542 (2017)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). http://www.R-project.org/

Ridgeway, G.: The state of boosting. Comput. Sci. Stat. **31**, 172–181 (1999)

Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. J. R. Stat. Soc. Ser. C (Appl. Stat.) **54**(3), 507–554 (2005). https://doi.org/10.1111/j.1467-9876.2005.00510.x

Rödel, C., Graeven, U., Fietkau, R., Hohenberger, W., Hothorn, T., Arnold, D., Hofheinz, R.D., Ghadimi, M., Wolff, H.A., Lang-Welzenbach, M., Raab, H.R., Wittekind, C., Ströbel, P., Staib, L., Wilhelm, M., Grabenbauer, G.G., Hoffmanns, H., Lindemann, F., Schlenska-Lange, A., Folprecht, G., Sauer, R.: Torsten Liersch on behalf of the German Rectal Cancer Study Group: Oxaliplatin added to fluorouracil-based preoperative chemoradiotherapy and postoperative chemotherapy of locally advanced rectal cancer (the German CAO/ARO/AIO-04 study): final results of the multicentre, open-label, randomised, phase 3 trial. Lancet Oncol. **16**(8), 979–989 (2015). https://doi.org/10.1016/S1470-2045(15)00159-X

Rothe, C., Wied, D.: Misspecification testing in a class of conditional distributional models. J. Am. Stat. Assoc. **108**(501), 314–324 (2013). https://doi.org/10.1080/01621459.2012.736903

Schild, R.L., Maringa, M., Siemer, J., Meurer, B., Hart, N., Goecke, T.W., Schmid, M., Hothorn, T., Hansmann, M.E.: Weight estimation by three-dimensional ultrasound in the small fetus. Ultrasound Obstetr. Gynecol. **32**(2), 168–175 (2008). https://doi.org/10.1002/uog.6111

Schmid, M., Hothorn, T.: Flexible boosting of accelerated failure time models. BMC Bioinform. **9**, 269 (2008). https://doi.org/10.1186/1471-2105-9-269

Schmid, M., Hothorn, T., Maloney, K.O., Weller, D.E., Potapov, S.: Geoadditive regression modeling of stream biological condition. Environ. Ecol. Stat. **18**(4), 709–733 (2011). https://doi.org/10.1007/s10651-010-0158-4

Seibold, S., Brandl, R., Schmidl, J., Busse, J., Thorn, S., Hothorn, T., Müller, J.: Extinction risk status of saproxylic beetles reflects the ecological degradation of forests in Europe. Conserv. Biol. **29**(2), 382–390 (2015). https://doi.org/10.1111/cobi.12427

Seibold, H., Zeileis, A., Hothorn, T.: Individual treatment effect prediction for ALS patients. Stat. Methods Med. Res. (2017). https://doi.org/10.1177/0962280217693034

Wu, C.O., Tian, X.: Nonparametric estimation of conditional distributions and rank-tracking probabilities with time-varying transformation models in longitudinal studies. J. Am. Stat. Assoc. **108**(503), 971–982 (2013). https://doi.org/10.1080/01621459.2013.808949

Yang, Y., Tokdar, S.T.: Joint estimation of quantile planes over arbitrary predictor spaces. J. Am. Stat. Assoc. **112**(519), 1107–1120 (2017). https://doi.org/10.1080/01621459.2016.1192545

Yue, M., Li, J., Ma, S.: Sparse boosting for high-dimensional survival data with varying coefficients. Stat. Med. **37**(5), 789–800 (2017). https://doi.org/10.1002/sim.7544