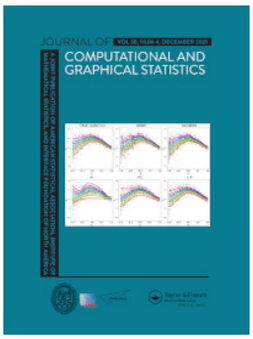

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Predictive Distribution Modeling Using Transformation Forests

Torsten Hothorn & Achim Zeileis

To cite this article: Torsten Hothorn & Achim Zeileis (2021) Predictive Distribution Modeling Using Transformation Forests, Journal of Computational and Graphical Statistics, 30:4, 1181-1196, DOI: 10.1080/10618600.2021.1872581

To link to this article: https://doi.org/10.1080/10618600.2021.1872581

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 30, NO. 4, 1181–1196
https://doi.org/10.1080/10618600.2021.1872581

Taylor & Francis
Taylor & Francis Group

Predictive Distribution Modeling Using Transformation Forests

Torsten Hothorn[a] and Achim Zeileis[b]

[a]Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich, Zürich, Switzerland; [b]Faculty of Economics and Statistics, Universität Innsbruck, Innsbruck, Austria

ABSTRACT

Regression models for supervised learning problems with a continuous response are commonly understood as models for the conditional mean of the response given predictors. This notion is simple and therefore appealing for interpretation and visualization. Information about the whole underlying conditional distribution is, however, not available from these models. A more general understanding of regression models as models for conditional distributions allows much broader inference, for example, the computation of prediction intervals or probabilistic predictions for exceeding certain thresholds. Several random forest-type algorithms aim at estimating conditional distributions, most prominently quantile regression forests. We propose a novel approach based on a parametric family of distributions characterized by their transformation function. A dedicated novel "transformation tree" algorithm able to detect distributional changes is developed. Based on these transformation trees, we introduce "transformation forests" as an adaptive local likelihood estimator of conditional distribution functions. The resulting predictive distributions are fully parametric yet very general and allow inference procedures, such as likelihood-based variable importances, to be applied in a straightforward way. Supplemental files for this article are available online.

ARTICLE HISTORY
Received October 2018
Revised June 2020

KEYWORDS
Conditional distribution;
Conditional quantiles;
Quantile regression forest;
Random forest;
Transformation model

1. Introduction

Supervised learning plays an important role in many prediction problems. Based on a learning sample consisting of N pairs of response y and predictors $\boldsymbol{x}$, one learns a rule r that predicts some unseen Y via $r(\boldsymbol{x})$ when only information about $\boldsymbol{x}$ is available. Both the statistics and machine learning communities differentiate between "classification problems," where the response Y is a class label, and "regression problems" with conceptually continuous response Y. In binary classification problems with $Y \in \{0, 1\}$ the focus is on rules r for the conditional probability of Y being 1 given $\boldsymbol{x}$, more formally $\mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = r(\boldsymbol{x})$. Such a classification rule r is probabilistic in the sense that one cannot only predict the most probable class label but also assess the corresponding probability. This additional information is extremely valuable because it allows an assessment of the rule's uncertainty about its prediction. It is much harder to obtain such an assessment of uncertainty from most contemporary regression models, because the rule or "regression function" r typically describes the conditional expectation $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = r(\boldsymbol{x})$ but not the full predictive distribution of Y given $\boldsymbol{x}$. Without making additional restrictive assumptions, for example constant variances in normal distributions, the derivation of probabilistic statements from the regression function r alone is impossible.

In many applications, the conditional mean is not, or not exclusively, of primary interest (e.g., Pinson 2013). As an example, we investigate the impact of lifestyle factors, such as smoking or physical activity, on the body mass index (BMI) in Section 6.6. Important aspects, such as under-weight or obesity, are related to lower and upper quantiles of the conditional BMI distribution given a certain lifestyle, while the conditional mean or median is only of secondary importance. The novel transformation tree presented in Figure 1 clearly shows that higher moments of the conditional BMI distributions, depicted as densities in the terminal nodes of the tree, vary between certain lifestyle configurations. While there is clearly also variation of the mean BMI across nodes, many splits mostly influence variance and/or skewness but not the mean, for example, the extent of physical activity (node 5 vs. 6) or alcohol intake (node 13 vs. 14). The development of tree and forest algorithms which are sensitive to such changes in higher moments is the aim of this article.

Contemporary random forest-type algorithms strongly rely on the notion of regression functions r describing the conditional mean $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ only (e.g., Biau, Devroye, and Lugosi 2008; Biau 2012; Scornet, Biau, and Vert 2015), although the first random forest-type algorithm for the estimation of conditional distribution functions was published more than 15 years ago ("bagging survival trees," Hothorn et al. 2004). A similar approach was later developed independently as "quantile regression forests" (Meinshausen 2006). In contrast to a mean aggregation of cumulative hazard functions (Ishwaran et al.

CONTACT Torsten Hothorn *Torsten.Hothorn@R-project.org* Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich, Hirschengraben 84, 8001 Zürich, Switzerland.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

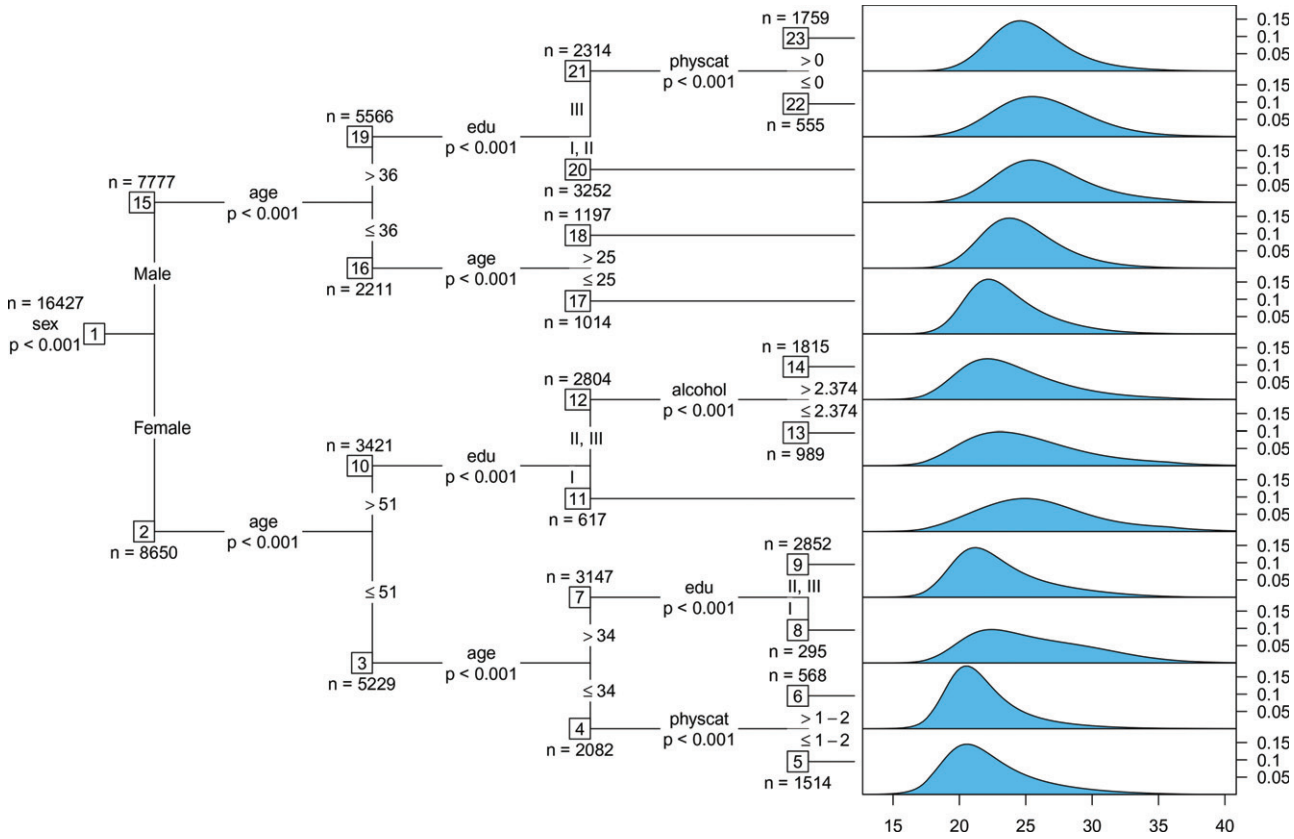

Figure 1. Body mass index (BMI). The conditional BMI distributions (depicted in terms of their densities) are given in each subgroup of the transformation tree corresponding to the terminal nodes of the tree. Variables: sex (female, male), age (in years), education (edu) at levels mandatory (I), secondary (II), and tertiary (III), physical activity (days per week), and alcohol intake (gram per day).

2008) or densities (Criminisi, Shotton, and Konukoglu 2012), bagging survival trees and quantile regression forests are based on "nearest neighbor weights" (Lin and Jeon 2006). The core idea is to obtain a "distance" measure based on the number of times a pair of observations is assigned to the same terminal node in the different trees of the forest. Similar observations have a high probability of ending up in the same terminal node whereas this probability is low for rather dissimilar observations. Then, the prediction for a set of predictor values $\boldsymbol{x}$ (either new or observed) is simply obtained as a weighted empirical distribution function (or Kaplan–Meier estimator in the context of right-censored response values) where those observations from the learning sample similar (or dissimilar) to $\boldsymbol{x}$ in the forest receive high (or low/zero) weights, respectively. Although this aggregation procedure in the aforementioned algorithms is suitable for estimating predictive distributions, the underlying trees are not. The reason is that the ANOVA- or log-rank-type split procedures commonly applied are not able to deal with distributions in a general sense. Consequently, the splits favor the detection of changes in the mean—or have power against proportional hazards alternatives in survival trees. However, in general, they have very low power for detecting other patterns of heterogeneity (e.g., changes in variance) even if these can be explained by the predictor variables. A simple toy example illustrating this problem is given in Figure 2. Here, the response variable's conditional normal distribution has a variance split at value 0.5 of a uniform [0, 1] predictor. We fitted a quantile regression forest (Meinshausen 2006) to the 10,000 observations depicted in the figure along with ten additional independent uniformly distributed noninformative predictors (using 1000 trees without random variable selection; see online appendix). The true conditional 10% and 90% quantiles (highlighted by red vs. gray circles) are not approximated very well by the estimated 10% and 90% quantiles (solid black line) from the quantile regression forest. In particular, the split at 0.5 does not play an important role in this model. Thus, although such an abrupt change in the distribution can be represented by a binary tree, the traditional ANOVA split criterion employed here was not able to detect this split. The transformation tree (dashed blue line) exactly matching the data generating process accurately estimated a stump, and transformation forests (solid blue line) were also able to pick-up this abrupt change at 0.5.

To improve upon quantile regression forests and similar procedures in situations where changes in moments beyond the mean are important, we propose "transformation forests" for the estimation and prediction of conditional distributions for Y given predictor variables $\boldsymbol{x}$ and proceed in three steps. We first suggest to understand forests as adaptive local likelihood estimators. Second, we recap the most important features of the flexible and computationally attractive "transformation family" of distributions (Hothorn, Kneib, and Bühlmann 2014; Hothorn, Möst, and Bühlmann 2018). Finally, we adapt the core ideas of "model-based recursive partitioning" (Zeileis, Hothorn, and Hornik 2008, who also provided a review of earlier developments in this field) to this transformation family and introduce novel algorithms for "transformation trees" and "transformation

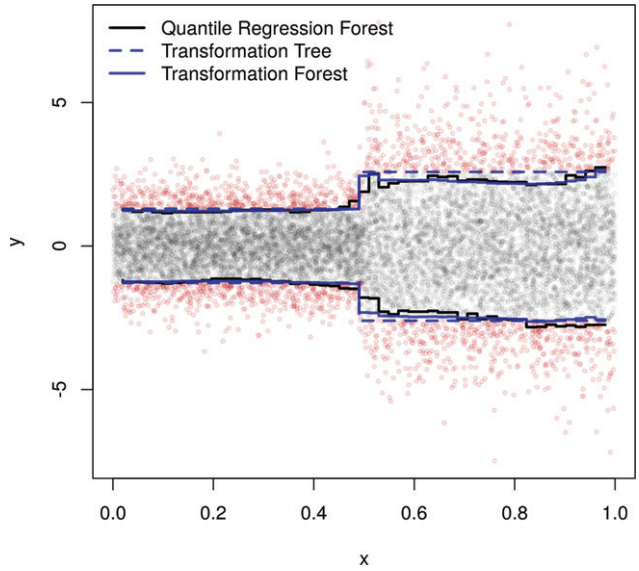

Figure 2. Empirical illustration. 10,000 observations from a normal distribution $Y \sim \mathrm{N}(0, (1 + I(x > 0.5))^2)$, that is, with an abrupt variance split at 0.5 in the uniform predictor x. Points outside the true conditional 10% and 90% quantiles are highlighted in red. Lines depict estimated conditional 10% and 90% quantiles obtained from quantile regression forests (solid black) versus transformation trees (dashed blue) and transformation forests (solid blue).

forests" for the estimation of conditional distribution functions which potentially vary in the mean and also in higher moments as a function of predictor variables $\boldsymbol{x}$. In the toy example from Figure 2, these novel transformation trees (dashed blue line) and forests (solid blue line) are able to recover the true conditional quantiles more precisely than quantile regression forests. Some remarks on asymptotic properties are given in Section 5 and the empirical performance of transformation trees and forests is evaluated on four artificial data-generating processes as well as the previously mentioned BMI survey data from Switzerland in Section 6.

2. Adaptive Local Likelihood Trees and Forests

We first deal with the unconditional distribution $\mathbb{P}_Y$ of a response random variable $Y \in \mathcal{Y}$ and we restrict our attention to a specific probability model defined by the parametric family of distributions $\mathbb{P}_{Y,\Theta} = \{\mathbb{P}_{Y,\boldsymbol{\vartheta}} \mid \boldsymbol{\vartheta} \in \Theta\}$ with parameters $\boldsymbol{\vartheta}$ and parameter space $\Theta \subseteq \mathbb{R}^p$. With predictors $\boldsymbol{X} \in \mathcal{X}$ from some predictor sample space $\mathcal{X}$, our main interest is in the conditional distribution $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ and we assume that this conditional distribution is a member of the family of distributions introduced above, that is, we assume that a parameter $\boldsymbol{\vartheta}(\boldsymbol{x}) \in \Theta$ exists such that $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}} = \mathbb{P}_{Y,\boldsymbol{\vartheta}(\boldsymbol{x})}$. We call $\boldsymbol{\vartheta} : \mathcal{X} \to \Theta$ the "conditional parameter function" and the task of estimating the conditional distributions $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ for all $\boldsymbol{x}$ reduces to the problem of estimating this conditional parameter function.

From the probability model $\mathbb{P}_{Y,\Theta}$ we can derive the log-likelihood contribution $\ell_i : \Theta \to \mathbb{R}$ for each of N independent observations $(y_i, \boldsymbol{x}_i)$ from the learning sample for $i = 1, \ldots, N$. We propose and study a novel random forest-type estimator $\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}$ of the conditional parameter function $\boldsymbol{\vartheta}$ in the class of adaptive local likelihood estimators of the form

$$\hat{\boldsymbol{\vartheta}}^N(\boldsymbol{x}) := \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^N w_i^N(\boldsymbol{x}) \ell_i(\boldsymbol{\vartheta}); \quad \boldsymbol{x} \in \mathcal{X}, \tag{1}$$

where $w_i^N : \mathcal{X} \to \mathbb{R}^+$ is the "conditional weight function" for observation i given a specific configuration $\boldsymbol{x}$ of the predictor variables (which may correspond to an observation from the learning sample or to new data). This weight measures the similarity of the two distributions $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}_i}$ and $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ under the probability model $\mathbb{P}_{Y,\Theta}$. The main idea is to obtain a large weight for observations i which are "close" to $\boldsymbol{x}$ in light of the model and essentially zero in the opposite case. The superscript N indicates that the weight function may depend on the learning sample, and in fact the choice of the weight function w_i^N is crucial in what follows.

Local likelihood estimation was founded by Brillinger (1977) and Tibshirani and Hastie (1987). Early regression models in this class were based on the idea of fitting polynomial models locally within a fixed smoothing window. Adaptivity of the weights refers to an $\boldsymbol{x}$-dependent, nonconstant smoothing window, that is, different weighting schemes are applied in different parts of the predictor sample space $\mathcal{X}$. Subsequently, we illustrate how classical maximum likelihood estimators, trees, and forests can be embedded in this general framework by choosing suitable conditional weight functions and plugging these into (1).

The unconditional maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^N_{\text{ML}}$ is based on unit weights $w^N_{\text{ML},i} :\equiv 1$ not depending on $\boldsymbol{x}$, that is, all observations in the learning sample are considered to be equally "close." In contrast, trees can adapt to the learning sample by employing rectangular splits to define a partition $\mathcal{X} = \dot{\bigcup}_{b=1,\ldots,B} \mathcal{B}_b$ of the predictor sample space. Each of the B cells then contains a different local unconditional model. More precisely, the conditional weight function $w^N_{\text{Tree},i} : \mathcal{X} \to \{0, 1\}$ is simply an indicator for $\boldsymbol{x}_i$ and $\boldsymbol{x}$ being elements of the same terminal node so that only observations in the same terminal node are considered to be "close." The weight and parameter functions defining a "model-based tree" are

$$\begin{aligned} w^N_{\text{Tree},i}(\boldsymbol{x}) &:= \sum_{b=1}^B I(\boldsymbol{x} \in \mathcal{B}_b \wedge \boldsymbol{x}_i \in \mathcal{B}_b), \\ \hat{\boldsymbol{\vartheta}}^N_{\text{Tree}}(\boldsymbol{x}) &:= \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^N w^N_{\text{Tree},i}(\boldsymbol{x}) \ell_i(\boldsymbol{\vartheta}). \end{aligned} \tag{2}$$

Thus, if $\boldsymbol{x}$ falls into the cell $\mathcal{B}_b$, this essentially just picks the corresponding parameter estimate from the bth terminal node

$$\hat{\boldsymbol{\vartheta}}^N_b = \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^N I(\boldsymbol{x}_i \in \mathcal{B}_b) \ell_i(\boldsymbol{\vartheta}),$$

along with the associated conditional distribution $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_b}$. Model-based recursive partitioning (MOB, Zeileis, Hothorn, and Hornik 2008) is one representative of such a tree-structured approach.

A forest of T trees is associated with partitions $\mathcal{X} = \dot{\bigcup_{b=1,\ldots,B_t}} \mathcal{B}_{t,b}$ for $t = 1, \ldots, T$. The bth terminal node of the tth tree contains the parameter estimate $\hat{\boldsymbol{\vartheta}}^N_{t,b}$ and the tth tree defines the conditional parameter function $\hat{\boldsymbol{\vartheta}}^N_{\text{Tree},t}(\boldsymbol{x})$. We define the forest conditional parameter function via "nearest neighbor" forest weights (with $|\mathcal{B}_{t,b}|$ denoting the number of observations in the bth cell of the tth tree)

$$
\begin{aligned}
w^N_{\text{Forest},i}(\boldsymbol{x}) &:= \sum_{t=1}^{T} \sum_{b=1}^{B_t} \frac{I(\boldsymbol{x} \in \mathcal{B}_{t,b} \wedge \boldsymbol{x}_i \in \mathcal{B}_{t,b})}{|\mathcal{B}_{t,b}|}, \\
\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x}) &:= \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w^N_{\text{Forest},i}(\boldsymbol{x}) \ell_i(\boldsymbol{\vartheta}).
\end{aligned} \tag{3}
$$

The conditional weight function $w^N_{\text{Forest},i} : \mathcal{X} \to \{0, \ldots, T\}$ computes how often $\boldsymbol{x}_i$ and $\boldsymbol{x}$ are element of the same terminal node, that is, captures how "close" the observations are across the trees in the forest. These weights were first suggested for the aggregation of T survival trees (Hothorn et al. 2004) and have later been used for estimating conditional means (Lin and Jeon 2006), for estimating conditional quantiles (Meinshausen 2006; Athey, Tibshirani, and Wager 2019), for estimating local linear models (Bloniarz et al. 2016), and for the estimation of heterogeneous treatment effects (Seibold, Zeileis, and Hothorn 2018; Wager and Athey 2018). An "out-of-bag" version only counts the contribution of the tth tree for observation i when i was not used for fitting the tth tree.

Forests relying on the aggregation scheme (3) model the conditional distribution $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ for some configuration $\boldsymbol{x}$ of the predictors as $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x})} \in \mathbb{P}_{Y,\Theta}$. In this sense, such a forest is a fully specified parametric model with (in-bag or out-of-bag) log-likelihood

$$
\sum_{i=1}^{N} \ell_i \left(\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x}_i) \right).
$$

This core idea of tree-based and forest-based adaptive local likelihood estimation is very flexible and seems to be straightforward to implement. However, simply picking our favorite (a) tree algorithm and coupling it with (b) some parametric model will not necessarily lead to a good predictive distribution model. In terms of (a), the variable and split selection procedures of most standard tree algorithms are not sensitive to distributional changes that are not linked to changes in the mean—as illustrated by the simple toy example from Figure 2. Therefore, a tailored tree algorithm inspired by model-based recursive partitioning (Zeileis, Hothorn, and Hornik 2008) is introduced in Section 4, to detect changes in higher moments with high power. In terms of (b), in principle, all classical probability models with families $\mathbb{P}_{Y,\Theta}$ are suitable in this general framework. However, different parameterizations render unified presentation and especially implementation burdensome. We address problem (b) by restricting our implementation to a novel transformation family of distributions. Theoretically, this family contains all univariate probability distributions $\mathbb{P}_Y$ and practically close approximations thereof (see Hothorn, Kneib, and Bühlmann 2014; Hothorn, Möst, and Bühlmann 2018, for technical details). We highlight important aspects of this family and the corresponding likelihood function in the next section.

3. Transformation Models

A transformation model $\mathbb{P}(Y \leq y) = F_Y(y) = F_Z(h(y))$ describes the distribution function of Y by an unknown monotone increasing transformation function h and some a priori chosen continuous distribution function F_Z. We use this framework because simple, that is, linear or log-linear, transformation functions implement many of the classical parametric models whereas more complex transformation functions provide similar flexibility as nonparametric models. In addition, discrete and continuous response variables, also under all forms of random censoring and truncation, can be handled in a unified way (Hothorn, Möst, and Bühlmann 2018). As a consequence, our corresponding "transformation forests" will be applicable to a wide range of response variables (discrete, continuous with or without censoring and truncation, counts, survival times) with the option to gradually move from simple to very flexible models for the conditional distribution functions $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x})}$.

In more detail, let $Z \sim \mathbb{P}_Z$ denote an absolutely continuous random variable with probability density function f_Z, cumulative distribution function F_Z, and quantile function F_Z^{-1}, respectively. We furthermore assume $0 < f_Z(z) < \infty \, \forall z \in \mathbb{R}$ for a log-concave density f_Z as well as the existence of the first two derivatives of the density $f_Z(z)$ with respect to z, both derivatives shall be bounded. We do not allow any unknown parameters for this distribution. Possible choices include the standard normal distribution with $F_Z(z) = \Phi(z)$, the standard logistic distribution with $F_Z(z) = (1 + \exp(-z))^{-1}$, the standard minimum extreme value distribution with $F_Z(z) = 1 - \exp(-\exp(z))$, or the standard maximum extreme value distribution with $F_Z(z) = \exp(-\exp(-z))$, respectively.

Let $\mathcal{H} = \{h : \mathcal{Y} \to \mathbb{R} \mid h(y_1) < h(y_2) \, \forall y_1 < y_2 \in \mathcal{Y}\}$ denote the space of all strictly monotone transformation functions. With the transformation function h we can write F_Y as $F_Y(y \mid h) = F_Z(h(y)) \, \forall y \in \mathcal{Y}$ with density $f_Y(y \mid h)$ and there exists a unique transformation function $h = F_Z^{-1} \circ F_Y$ for all distribution functions F_Y (Hothorn, Möst, and Bühlmann 2018). A convenient feature of characterizing the distribution of Y by means of the transformation function h is that the likelihood for arbitrary measurement scales can be written and implemented in an extremely compact form (Hothorn, Möst, and Bühlmann 2018).

For a given transformation function h, the likelihood contribution of an observation $y \in \mathbb{R}$ is given by the corresponding density $\mathcal{L}(h \mid Y = y) = f_Z(h(y))h'(y)$. The likelihood contribution from intervals $(\underline{y}, \bar{y}] \subset \mathcal{Y}$ is, unlike in the above "exact continuous" case, defined in terms of the distribution function (Lindsey 1996), where one can differentiate between three special cases:

$$
\begin{aligned}
&\mathcal{L}(h \mid Y \in (\underline{y}, \bar{y}]) \\
&= \begin{cases}
F_Z(h(\bar{y})) - F_Z(h(\underline{y})) & y \in (\underline{y}, \bar{y}] \quad \text{"interval-censored,"} \\
1 - F_Z(h(\underline{y})) & y \in (\underline{y}, \infty) \quad \text{"right-censored,"} \\
F_Z(h(\bar{y})) & y \in (-\infty, \bar{y}] \quad \text{"left-censored."}
\end{cases}
\end{aligned}
$$

For truncated observations in the interval $(y_l, y_r] \subset \mathcal{Y}$, the above likelihood contribution has to be multiplied by the factor $(F_Z\left(h(y_r)\right) - F_Z\left(h(y_l)\right))^{-1}$ when $y_l < \underline{y} < \bar{y} \leq y_r$. A more detailed discussion of likelihood contributions in transformation models can be found in Hothorn, Möst, and Bühlmann (2018).

We parameterize the unknown transformation function $h(y)$ as a linear function of its basis-transformed argument y using a basis function $\boldsymbol{a} : \mathcal{Y} \to \mathbb{R}^{\mathbb{P}}$ such that $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}, \boldsymbol{\vartheta} \in \mathbb{R}^{\mathbb{P}}$. In the following, we will write $h = \boldsymbol{a}^\top \boldsymbol{\vartheta}$ and assume that the true unknown transformation function is of this form. For continuous response variables Y the parameterization $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ needs to be smooth in y, so any polynomial or spline basis is a suitable choice for $\boldsymbol{a}$. For the empirical experiments in Section 6 we employed Bernstein polynomials (for an overview see Farouki 2012) of order M ($\mathbb{P} = M + 1$) defined on the interval $[\underline{\iota}, \bar{\iota}]$ with

$$\begin{aligned}
\boldsymbol{a}_{\text{Bs},M}(y) &= (M+1)^{-1}(f_{\text{Be}(1,M+1)}(\tilde{y}), \ldots, f_{\text{Be}(M+1,1)}(\tilde{y}))^\top \in \mathbb{R}^{M+1}, \\
h(y) &= \boldsymbol{a}_{\text{Bs},M}(y)^\top \boldsymbol{\vartheta} = \sum_{m=0}^{M} \vartheta_m f_{\text{Be}(m+1,M-m+1)}(\tilde{y})/(M+1), \\
h'(y) &= \boldsymbol{a}'_{\text{Bs},M}(y)^\top \boldsymbol{\vartheta} = \sum_{m=0}^{M-1} \frac{(\vartheta_{m+1} - \vartheta_m) f_{\text{Be}(m+1,M-m)}(\tilde{y}) M}{(M+1)(\bar{\iota} - \underline{\iota})},
\end{aligned}$$

where $\tilde{y} = (y - \underline{\iota})/(\bar{\iota} - \underline{\iota}) \in [0, 1]$ and $f_{\text{Be}(m,M)}$ is the density of the Beta distribution with parameters m and M. Outside $[\underline{\iota}, \bar{\iota}]$, h is linearly extrapolated. This parameterization is computationally attractive because strict monotonicity can be formulated by M linear constraints on the parameters $\vartheta_m < \vartheta_{m+1}$ for all $m = 0, \ldots, M$ (Curtis and Ghosh 2011).

The distribution family $\mathbb{P}_{Y,\Theta} = \{F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta} \mid \boldsymbol{\vartheta} \in \Theta\}$ transformation forests are based upon is called transformation family of distributions with parameter space $\Theta = \{\boldsymbol{\vartheta} \in \mathbb{R}^{\mathbb{P}} \mid \boldsymbol{a}^\top \boldsymbol{\vartheta} \in \mathcal{H}\}$ and transformation functions $\boldsymbol{a}^\top \boldsymbol{\vartheta} \in \mathcal{H}$. This family encompasses a wide variety of densities capturing different locations and shapes (including scale and skewness), see Figure 1 for an illustration of different body mass index distributions. The log-likelihood contribution for an observation $y_i \in \mathbb{R}$ is now the log-density of the transformation model $\ell_i(\boldsymbol{\vartheta}) = \log(f_Z(\boldsymbol{a}(y_i)^\top \boldsymbol{\vartheta})) + \log(\boldsymbol{a}'(y_i)^\top \boldsymbol{\vartheta})$.

4. Transformation Trees and Forests

Based on the parameterization of unconditional transformation models sketched in the previous section, we propose an algorithm for the estimation of conditional transformation models $\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}))$ where the estimated conditional parameter function $\hat{\boldsymbol{\vartheta}}^N$ can be either tree-structured (2) or essentially unstructured based on weights (3).

Conceptually, the model-based recursive partitioning algorithm (Zeileis, Hothorn, and Hornik 2008) for tree induction starts with the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^N_{\text{ML}}$. Deviations from such a given model that can be explained by parameter instabilities due to one or more of the predictors are investigated based on the score (or gradient) contributions. The novel "transformation trees" suggested here rely on the transformation family $\mathbb{P}_{Y,\Theta}$ whose score contributions $\boldsymbol{s}$ have relatively simple and generic forms. The score contribution of an "exact continuous" observation $y \in \mathbb{R}$ from an absolutely continuous distribution is given by the gradient of the log-density with respect to $\boldsymbol{\vartheta}$

$$\boldsymbol{s}(\boldsymbol{\vartheta} \mid Y = y) = \boldsymbol{a}(y) \frac{f_Z'(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta})}{f_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta})} + \frac{\boldsymbol{a}'(y)}{\boldsymbol{a}'(y)^\top \boldsymbol{\vartheta}}.$$

For an interval-censored observation $(\underline{y}, \bar{y}]$ the score contribution is

$$\boldsymbol{s}(\boldsymbol{\vartheta} \mid Y \in (\underline{y}, \bar{y}]) = \frac{f_Z(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\bar{y}) - f_Z(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\underline{y})}{F_Z(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta})}.$$

Under truncation to the interval $(y_l, y_r] \subset \mathcal{Y}$, one needs to substract the term $\boldsymbol{s}(\boldsymbol{\vartheta} \mid Y \in (y_l, y_r])$ from the score function.

With the transformation model and thus the likelihood and score function being available, we start tree induction with the global model $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_{\text{ML}}}$. The hypothesis of all observations $i = 1, \ldots, N$ coming from this model can be written as the independence of the $\mathbb{P}$-dimensional score contributions and all predictors, that is, $H_0 : \boldsymbol{s}(\hat{\boldsymbol{\vartheta}}^N_{\text{ML}} \mid Y) \perp \boldsymbol{X}$. Different test procedures have been suggested for assessing this hypothesis, including asymptotic M-fluctuation tests (Zeileis, Hothorn, and Hornik 2008) and permutation tests (Hothorn, Hornik, and Zeileis 2006; Zeileis and Hothorn 2013), both using appropriate multiplicity adjustment depending on the number of predictors. Rejection of H_0 leads to the implementation of a binary split in the predictor variable with most significant association to the score matrix; algorithmic details are discussed in the online appendix. Unbiasedness of a model-based tree with respect to variable selection is a consequence of splitting in the variable of highest association to the scores where association is measured by the marginal multiplicity-adjusted p-value (for details see Hothorn, Hornik, and Zeileis 2006; Zeileis, Hothorn, and Hornik 2008, and online appendix). The procedure is recursively iterated until H_0 cannot be rejected. The result is a partition of the sample space $\mathcal{X} = \dot{\bigcup}_{b=1,\ldots,B} \mathcal{B}_b$.

Based on the "transformation trees" introduced here, we construct a corresponding random forest-type algorithm as follows. A "transformation forest" is an ensemble of T transformation trees fitted to subsamples of the learning sample and, optionally, a random selection of candidate predictors available for splitting in each node of the tree. The result is a set of T partitions of the predictor sample space. The conditional parameter function of the transformation forest is defined by its nearest neighbor forest weights (3).

The question arises how the order M of the Bernstein polynomials parameterizing the transformation function h affects the conditional distribution functions $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_{\text{Tree}}(\boldsymbol{x})}$ and $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x})}$. On the one hand, the basis $\boldsymbol{a}_{\text{Bs},1}$ with $F_Z = \Phi$ only allows linear transformation functions of a standard normal and thus our models for $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ are restricted to the normal family, however, with potentially both mean and variance depending on $\boldsymbol{x}$ as the split criterion in transformation trees is sensitive to changes in both location and scale. This most simple version of transformation trees and forests based on the model $Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{N}(\mu(\boldsymbol{x}), \sigma(\boldsymbol{x})^2)$ allows both the conditional mean and the conditional variance to be inferred (the same model with

similar flexible mean and variance functions can be estimated by Bayesian additive regression trees, Pratola et al. 2020). Using a higher order M also allows modeling nonnormal distributions. With $M > 1$, the split criterion introduced in this section is able to detect changes beyond the second moment and, consequently, also higher moments of the conditional distributions $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ may vary with $\boldsymbol{x}$. An empirical comparison of transformation trees and forests with linear ($M = 1$) and nonlinear ($M > 1$) transformation function can be found in Section 6. Additional empirical properties of transformation models with larger values of M are discussed in Hothorn (2018).

In contrast to other random forest regression models, the predictive distributions obtained from a transformation forest are fully specified parametric distributions $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x})}$ and we can describe these on the scale of the distribution, quantile, density, hazard, cumulative hazard, expectile, and any other characterizing functions. Two interesting applications of these predictive distributions include prediction intervals and likelihood-based variable importances which we discuss briefly.

4.1. Prediction Intervals and Outlier Detection

For some yet unobserved response Y under predictors $\boldsymbol{x}$, a two-sided $(1-\alpha)$ prediction interval for $Y \mid \boldsymbol{X} = \boldsymbol{x}$ and some $\alpha \in (0, 0.5)$ can be obtained by numerical inversion of the conditional distribution $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x})}$ via

$$\text{PI}_\alpha(\boldsymbol{x} \mid \hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}) = \left\{ y \in \mathcal{Y} \mid \alpha/2 < \mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x})}(y) \leq 1 - \alpha/2 \right\}$$

with the property $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}(\text{PI}_\alpha(\boldsymbol{x} \mid \boldsymbol{\vartheta})) = 1 - \alpha$ at the true parameters $\boldsymbol{\vartheta}(\boldsymbol{x})$. The empirical level of the prediction interval $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}(\text{PI}_\alpha(\boldsymbol{x} \mid \hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}))$ depends on how well the parameters $\boldsymbol{\vartheta}(\boldsymbol{x})$ are approximated by the forest estimate $\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x})$. If for some observation $(y_i, \boldsymbol{x}_i)$ the corresponding prediction interval $\text{PI}_\alpha(\boldsymbol{x}_i \mid \hat{\boldsymbol{\vartheta}}^N_{\text{Forest}})$ excludes y_i, one can (at level α) suspect this observed response of being an outlier.

4.2. Permutation Variable Importance

The importance of a variable is defined as the amount of change in the risk function when the association between one predictor variable and the response is artificially broken. Permutation variable importances permute one of the predictors at a time (and thus also break the association to the remaining predictors, see Strobl et al. 2008). The risk function for transformation forests is the negative log-likelihood, thus a universally applicable formulation of variable importance for all types of measurement scales or censoring or truncation is

$$\text{VI}(j) = T^{-1} \sum_{t=1}^{T} \left(\sum_{i=1}^{N} -\ell_i \left(\hat{\boldsymbol{\vartheta}}^N_{\text{Tree},t}(\boldsymbol{x}_i) \right) - \sum_{i=1}^{N} -\ell_i \left(\hat{\boldsymbol{\vartheta}}^N_{\text{Tree},t}(\boldsymbol{x}_i^{(j)}) \right) \right),$$

where the jth variable was permuted in $\boldsymbol{x}_i^{(j)}$ for $i = 1, \ldots, N$.

5. Theoretical Aspects

The theoretical properties of random forest-type algorithms are a contemporary research problem and we refer to Biau and Scornet (2016) for an overview. In this section we discuss how these developments relate to the asymptotic behavior of transformation trees and transformation forests.

Established theoretical results on random forests (Breiman 2004; Lin and Jeon 2006; Meinshausen 2006; Biau, Devroye, and Lugosi 2008; Biau 2012; Scornet, Biau, and Vert 2015) provide a basis for the analysis of transformation forests. Analytic results on random forests for estimating conditional means with adaptive nearest neighbor weights, where estimators for the conditional mean of the form

$$\hat{\mathbb{E}}_N(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\sum_{i=1}^N w^N_{\text{Forest},i}(\boldsymbol{x}) Y_i}{\sum_{i=1}^N w^N_{\text{Forest},i}(\boldsymbol{x})}$$

were shown to be consistent in nonadaptive random forests

$$\mathbb{E}_{Y|\boldsymbol{X}=\boldsymbol{x}} \left(\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) - \hat{\mathbb{E}}_N(Y \mid \boldsymbol{X} = \boldsymbol{x}) \right)^2 \to 0$$

as $N \to \infty$ (Lin and Jeon 2006). A Glivenko–Cantelli-type result for conditional distribution functions

$$\begin{aligned} \hat{\mathbb{P}}_N(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) &= \hat{\mathbb{E}}_N(I(Y \leq y) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \frac{\sum_{i=1}^N w^N_{\text{RForest},i}(\boldsymbol{x}) I(Y_i \leq y)}{\sum_{i=1}^N w^N_{\text{RForest},i}(\boldsymbol{x})}, \end{aligned} \tag{4}$$

where the weights are obtained from Breiman and Cutler's original random forest implementation ("RForest," Breiman 2001a) is available in Meinshausen (2006).

To understand why these results carry over to transformation forests, we define the expected conditional log-likelihood given $\boldsymbol{x}$ for a fixed set of parameters $\boldsymbol{\vartheta}$ as $\ell(\boldsymbol{\vartheta} \mid \boldsymbol{X} = \boldsymbol{x}) := \mathbb{E}_{Y|\boldsymbol{X}=\boldsymbol{x}} \ell(\boldsymbol{\vartheta}, Y)$, where $\ell(\boldsymbol{\vartheta}, Y_i) = \ell_i(\boldsymbol{\vartheta})$ is the likelihood contribution from some observation Y_i. By definition, the true unknown parameter $\boldsymbol{\vartheta}(\boldsymbol{x})$ has minimal expected risk and thus maximizes the expected log-likelihood, that is, $\boldsymbol{\vartheta}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\vartheta} \in \Theta} \ell(\boldsymbol{\vartheta} \mid \boldsymbol{X} = \boldsymbol{x})$. Our random forest-type estimator of the expected conditional log-likelihood given $\boldsymbol{x}$ for a fixed set of parameters $\boldsymbol{\vartheta}$ is now

$$\hat{\ell}_N(\boldsymbol{\vartheta} \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\sum_{i=1}^N w^N_{\text{Forest},i}(\boldsymbol{x}) \ell(\boldsymbol{\vartheta}, Y_i)}{\sum_{i=1}^N w^N_{\text{Forest},i}(\boldsymbol{x})}.$$

Under the respective conditions on the distribution of $\boldsymbol{X}$ and the joint distribution of $Y, \boldsymbol{X}$ (Lin and Jeon 2006; Biau and Devroye 2010; Biau 2012), this estimator is consistent for all $\boldsymbol{\vartheta} \in \Theta$

$$\mathbb{E}_{Y|\boldsymbol{X}=\boldsymbol{x}} \left(\ell(\boldsymbol{\vartheta} \mid \boldsymbol{X} = \boldsymbol{x}) - \hat{\ell}_N(\boldsymbol{\vartheta} \mid \boldsymbol{X} = \boldsymbol{x}) \right)^2 \to 0$$

(such results were usually derived for nonadaptive random forests). This result gives us consistency of the conditional log-likelihood function

$$\hat{\ell}_N(\boldsymbol{\vartheta} \mid \boldsymbol{X} = \boldsymbol{x}) \xrightarrow{\mathbb{P}} \ell(\boldsymbol{\vartheta} \mid \boldsymbol{X} = \boldsymbol{x}) \quad \forall\, \boldsymbol{\vartheta} \in \Theta.$$

The forest conditional parameter function $\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x})$ is consistent when

$$\mathbb{P}_{\boldsymbol{\vartheta}}(\hat{\ell}_N(\boldsymbol{\vartheta}_1 \mid \boldsymbol{X} = \boldsymbol{x}) < \hat{\ell}_N(\boldsymbol{\vartheta} \mid \boldsymbol{X} = \boldsymbol{x})) \xrightarrow{\mathbb{P}} 1$$

as $N \to \infty$ for all ϑ_1 in a neighborhood of ϑ. The result $\hat{\vartheta}^N_{\text{Forest}}(x) \xrightarrow{\mathbb{P}} \vartheta(x)$ can be shown under the assumptions regarding ℓ given by Hothorn, Möst, and Bühlmann (2018), especially continuity in ϑ. Because the conditional log-likelihood $\hat{\ell}_N(\vartheta \mid X = x)$ is a conditional mean-type estimator of a transformed response Y, contemporary and future theoretical developments in the asymptotic analysis of more realistic random forest-type algorithms based on nearest neighbor weights will directly carry over to transformation forests without specific adjustments.

It is worth noting that analytic results for real-world random forest algorithms in regression models of the form $Y = r(x) + \varepsilon$, that is, under variance homogeneity, are available from Scornet (2016). This is in line with the ANOVA split criterion implemented in Breiman and Cutler's random forests (Breiman 2001a). The split procedure applied in transformation trees is, as will be illustrated empirically in the next section, able to detect changes in higher moments. Thus, transformation forests might be a way to relax the assumption of additivity of signal and noise in the future.

6. Empirical Evaluation

To evaluate transformation trees and forests empirically—in comparison to established random forest procedures—a number of artificial data-generating processes were considered. These allowed to control the type of conditional parameter function, types of effect, and model complexity in low and high dimensions. More specifically, the following four hypotheses were assessed by simulation experiments:

H1 (Type of conditional parameter regression):

H1a: Tree-structured conditional parameter function. Transformation trees and forests are able to identify subgroups associated with different transformation models $\mathbb{P}_{Y|X=x}$, that is, subgroups formed by a recursive partition (or tree) in predictor variables x.

H1b: Nonlinear conditional parameter function. Transformation forests are able to identify conditional distributions $\mathbb{P}_{Y|X=x}$ whose parameters depend on predictor variables x in a smooth nonlinear way.

H2 (Type of effect):

H2a: No effect. In a noninformative scenario with $\mathbb{P}_{Y|X=x} = \mathbb{P}_Y$ (i.e., mean and all higher moments constant) transformation trees perform almost as well as the unconditional maximum likelihood estimator without pronounced overfitting.

H2b: Location only. Transformation trees and forests perform as well as classical regression trees and forests when higher moments of the conditional distribution $\mathbb{P}_{Y|X=x}$ are constant.

H2c: Unlinked location and scale. Transformation trees and forests outperform classical regression trees and forests when higher moments of the conditional distribution $\mathbb{P}_{Y|X=x}$ vary in a way that is *not* linked to variations in the mean.

H2d: Linked location and scale. Transformation trees and forests perform as well as classical regression trees and forests when higher moments of the conditional distribution $\mathbb{P}_{Y|X=x}$ vary in a way that is linked to the mean.

H3 (Model complexity):

$P = 2$**:** Transformation trees and forests with linear transformation function h, that is, with $P = 2$ parameters, perform best for conditionally normal response variables. Transformation trees and forests with nonlinear transformation function h perform somewhat worse in this situation.

$P = 6$**:** Transformation trees and forests with nonlinear transformation function h, that is, with $P = 6$ parameters of a Bernstein polynomial of order five, outperform transformation trees and forests with linear transformation function for conditionally nonnormal response variables.

The choice $P = 6$ was motivated by the fact that resulting distributions are sufficiently flexible (see, e.g., Figure 1 depicting densities estimated by a transformation tree with Bernstein polynomials featuring $P = 6$ parameters, or Figure 5 in Hothorn 2020a). The monotonicity constraint on the transformation function h ensures that adding more parameters does not lead to overfitting, however, computing times will increase.

H4 (Dimensionality):
Transformation forests stabilize transformation trees in the presence of high-dimensional noninformative predictor variables.

6.1. Data-Generating Processes

To assess **H1**–**H4** a set of data-generating processes with modular building blocks was considered, allowing to switch on or off particular features from the hypotheses.

6.1.1. H1a (Tree-Structured) Versus H1b (Nonlinear)

The mean and/or variance of a normal response variable are determined through functions of the predictors x that are either simple binary splits (Tree) or smooth nonlinear functions (Nonlin) inspired by the "Friedman 1" benchmark problem (Friedman 1991). More specifically, the conditional distributions are given by

$$Y \mid X = x \sim \text{N}\left(\mu_{\text{Tree}}(x), \sigma_{\text{Tree}}(x)^2\right), \tag{5}$$

$$Y \mid X = x \sim \text{N}\left(\mu_{\text{Nonlin}}(x), \sigma_{\text{Nonlin}}(x)^2\right). \tag{6}$$

For the tree-structured problem (5), the mean and variances are defined as

	$\mu_{\text{Tree}}(x)$	$\sigma_{\text{Tree}}(x)$
H2a	0	1
H2b	$I(x_1 > 0.5)$	1
H2c	0	$(1 + I(x_2 > 0.5))$
H2c	$I(x_1 > 0.5)$	$(1 + I(x_2 > 0.5))$

All predictor variables are independently uniform on $[0, 1]$ and in this scenario up to two informative predictors are used.

For a more complex and realistic nonlinear scenario the conditional mean function from the Friedman 1 benchmark problem is used based on five uniform informative predictor variables

$$\begin{aligned}&\text{Friedman1}(x_1, x_2, x_3, x_4, x_5)\\ &\quad = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5.\end{aligned}$$

To make this more comparable to the tree-structured case, the output of Friedman1 was scaled to the $[-1.5, 1.5]$ interval,

denoted as Friedman1*. Moreover, the same function is used for mean $\mu(x)$ and/or variance $\sigma(x)$ in H1b but based on different predictor variables:

	$\mu_{\text{Nonlin}}(\boldsymbol{x})$	$\sigma_{\text{Nonlin}}(\boldsymbol{x})$
H2a	0	1
H2b	Friedman1* $(x_1, x_2, x_3, x_4, x_5)$	1
H2c	0	exp(Friedman1* $(x_6, x_7, x_8, x_9, x_{10})$)
H2c	Friedman1* $(x_1, x_2, x_3, x_4, x_5)$	exp(Friedman1* $(x_6, x_7, x_8, x_9, x_{10})$)

Thus, there are up to ten informative predictors in this scenario.

6.1.2. H2d (Linked Location and Scale)

To link location and scale the same setup as above is considered but for a conditionally log-normal response Y' derived from models (5) and (6) as

$$\log(Y) \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{N}\left(\mu_{\text{Tree}}(\boldsymbol{x}), \sigma_{\text{Tree}}(\boldsymbol{x})^2\right), \tag{7}$$

$$\log(Y) \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{N}\left(\mu_{\text{Nonlin}}(\boldsymbol{x}), \sigma_{\text{Nonlin}}(\boldsymbol{x})^2\right). \tag{8}$$

Here, the conditional mean of the response variable Y' depends both on the underlying conditional mean $\mu(\boldsymbol{x})$ of Y and the corresponding conditional variance $\sigma(\boldsymbol{x})^2$:

$$\mathbb{E}(Y' \mid \boldsymbol{X} = \boldsymbol{x}) = \exp(\mu(\boldsymbol{x}) + \sigma(\boldsymbol{x})^2/2).$$

It is important to note that the true transformation function h in model (7) is a scaled and shifted log-transformation. Unlike the true linear transformation function h in model (5), which can be exactly fitted by the linear and Bernstein parameterizations of the transformation function in transformation trees and forests, the true log-transformation cannot be approximated easily by the basis functions $\boldsymbol{a}$. Therefore, no competitor in this simulation experiment is able to exactly recover the true data-generating process.

6.1.3. H4 (Low- Versus High-Dimensional)

In addition to the informative predictor variables described above further independent uniform noise variables were added. The low-dimensional case was defined as five additional noise variables whereas in the high-dimensional case, 500 noise variables were added.

6.2. Competitors

For testing the hypotheses H1–H4, we compared the performance of transformation trees and forests with linear and nonlinear transformation functions h to the performance of conditional inference trees (Hothorn, Hornik, and Zeileis 2006) and conditional inference forests (Strobl et al. 2007) as representatives of unbiased recursive partitioning and to Breiman and Cutler's random forests (Breiman 2001a) as an representative of exhaustive search procedures. In more detail, we compared the performance of the following methods:

CTree: Conditional inference trees with internal stopping by default parameters.

TTree: Transformation trees, either with linear ($P = 2$ parameters) or nonlinear ($P = 6$ parameters of a Bernstein polynomial) transformation functions. Tree-growing parameters are identical to those from CTree.

CForest: Conditional inference forests with `mtry` equal to one third of the number of predictor variables. Trees were grown without internal stopping until sample size constraints were met.

RForest: Breiman and Cutler's random forests with tree-growing parameters analogous to CForest (i.e., same `mtry` and stopping based on sample size constraints).

TForest: Transformation forests, either with linear ($P = 2$) or nonlinear ($P = 6$) transformation functions, and tree-growing parameters analogous to CForest and RForest.

Conditional inference trees and conditional inference forests were chosen as competitors because transformation trees and forests were implemented on top of these procedures, thus, potential differences in their empirical performance can be solely attributed to the novel split criteria proposed here. Exact tree-growing parameter specifications and a schematic overview of all competitors is given in the online appendix.

To allow a fair comparison on the same scale, trees and forests obtained from the classical methods, that is, conditional inference trees and forests and Breiman and Cutler's random forests, were used to estimate conditional parameter functions (2) and (3) in the same way as for transformation trees and forests: We first fitted trees and forests using the reference implementations of the corresponding methods and, second, computed the corresponding conditional weight functions, which allowed estimation of conditional parameter functions $\hat{\boldsymbol{\vartheta}}^N_{\text{CTree}}$, $\hat{\boldsymbol{\vartheta}}^N_{\text{CForest}}$, and $\hat{\boldsymbol{\vartheta}}^N_{\text{RForest}}$ in the third step. The combination of Breiman and Cutler's random forests with transformation models in our RForest variant is conceptually very similar to quantile regression forests (Meinshausen 2006). The same weights are used by both methods. The latter algorithm estimates a weighted empirical distribution function (4) whereas a weighted smooth conditional distribution function corresponding to a transformation model is estimated by the transformation forest.

6.3. Performance Measures

The primary performance measure is the out-of-sample negative log-likelihood because it assesses the whole predicted distribution in a proper way (Gneiting and Raftery 2007). To adjust for sampling variation, the negative log-likelihood of the true data-generating process is employed as the reference measure. More precisely, the negative log-likelihood difference, that is, the negative log-likelihood of a competitor minus the negative log-likelihood of the true data-generating process, was evaluated for the $N = 250$ observations of the validation sample. Conditional medians and prediction intervals are of additional interest and we also compared their performance by the out-of-sample check risk corresponding to the 10%, 50% (absolute error), and 90% quantiles in reference to the true data-generating process (see the online appendix). A direct comparison of coverage and lengths of prediction intervals is not considered as it would only

be valid or useful for a given configuration of the predictor variables (corresponding to maximizing forecast sharpness only subject to calibration in the proper scoring rules literature, Gneiting, Balabdaoui, and Raftery 2007).

6.4. Results: Tree-Structured Conditional Parameters (H1a)

For a normal response distribution and given the tree-structured conditional parameter function (H1a) the remaining properties of the data-generating process are varied and assessed. Figure 3 summarizes the results in terms of negative log-likelihood differences compared to the true model, using parallel coordinate displays with superimposed corresponding boxplots. These were obtained from 100 pairs of learning samples (size $N = 250$) and validation samples. The grid of panels in Figure 3 shows the type of effect in the rows (H2a–c, none, location and/or scale), the dimensionality in the columns (H4, 5 vs. 500 noise variables), and the complexity along the x-axes (H3, $P = 2$ vs. 6).

In the situation where all predictor variables were noninformative (H2a, top row of Figure 3), CTree ($P = 2$) and TTree ($P = 2$) were most resistant to overfitting; this effect is due to the test-based internal stopping of the unbiased tree methods compared here. TTree ($P = 6$) with nonlinear transformation function had slightly larger negative log-likelihood differences due to the increased model complexity (H3). Moreover, if model complexity is further increased by considering forests instead of trees, all random forest variants exhibit some more pronounced overfitting behavior.

Under the simple change in the mean (H2b, second row in Figure 3), CTree ($P = 2$) and TTree ($P = 2$) were able to detect this split best. The more complex TTree ($P = 6$) and all random forest variants performed less well in this situation. A variance change (H2c, third row in Figure 3) lead to smallest negative log-likelihood difference and thus superior performance for all transformation trees and forests as compared to the classical trees and forests which are sensitive to mean changes only. TTree ($P = 2$) performed best while none of the classical procedures seemed to be able to properly pick up this variance signal. The aggregation of multiple transformation trees leads to decreased performance, this effect was also visible in the toy example in Figure 2 (which was based on the same data-generating process (5)).

When changes in both mean and variance were present (H2c, fourth row in Figure 3), transformation forests with linear transformation function TForest ($P = 2$) performed better than all other procedures, also in the high-dimensional setup with 500 noninformative variables (H4). TTree ($P = 6$) showed some extreme outliers (H3, visible in the parallel coordinates in Figure 3) which were due to convergence problems. The corresponding transformation forests TForest ($P = 6$), however, did not experience such problems and thus seemed to stabilize the trees.

In summary, the results with respect to our hypotheses were:

H1a: Transformation trees reliably recover tree-structured conditional parameter functions in both mean and variance.

H2a: Transformation trees are rather robust to overfitting when there is no effect while transformation forests (like all other random forests) exhibit some overfitting.

H2b: Transformation trees and forests perform comparably to their classical counterparts when there are only mean effects.

H2c: Transformation trees and forests outperform their classical counterparts if there are only variance effects or variance effects that are not linked to the mean.

H3: For normal responses transformation trees and forests with linear transformation function ($P = 2$) consistently perform better than the more complex Bernstein polynomials ($P = 6$).

H4: Transformation forests stabilize the transformation trees in high-dimensional settings.

As a next step, the same simulation experiments were considered using a log-normal (rather than normal) response variable. This has two important implications: (1) changes in location and scale are linked, (2) none of the models can exactly recover the true highly skewed response distribution. Due to the latter, all methods based on linear transformations ($P = 2$) performed rather poorly and the corresponding results are not shown here.[1] The results regarding the following two hypotheses are affected:

H3: All models with complexity $P = 2$ are clearly not appropriate anymore as they cannot capture the skewness. Consequently, these models are outperformed by the more flexible Bernstein polynomials with $P = 6$.

H2d: The classic RForest ($P = 6$), that is, the combination of Breiman and Cutler's random forests with a subsequent flexible transformation model, performed almost on par with transformation trees and forests but only in the low-dimensional setup. The reason is that changes in the variance are always also linked to changes in the mean due to the skewness of the distribution.

H4: In the high-dimensional setup, however, with 500 noise predictors, TTree and TForest clearly outperform RForest. This effect was not anticipated when formulating hypotheses H1–4 above. Apparently, the splits found by exhaustive ANOVA splits in RForest only find suitable small-enough neighborhoods in low dimensions (left panels in Figure 4) but not in high dimensions (right panels).

Qualitatively the same conclusions can be drawn when assessing the competing methods based on predictions of the conditional 10% quantiles, 50% quantiles, and 90% quantiles (figures in online appendix). However, the differences are less pronounced for the 50% quantiles (medians, corresponding to the absolute errors). Note also that combining predictions of 10% and 90% quantiles amounts to 80% prediction intervals.

By and large, the empirical results in this section conformed with our hypotheses H1–4, suggesting a stable behavior of transformation trees and forests, especially with appropriate linear transformation function for normal response variables, in these very simple situations. The next section proceeds to a somewhat more realistic scenario with nonlinear conditional parameter functions defining mean and/or variance.

[1] See *https://arxiv.org/abs/1701.02110v2* for the complete results.

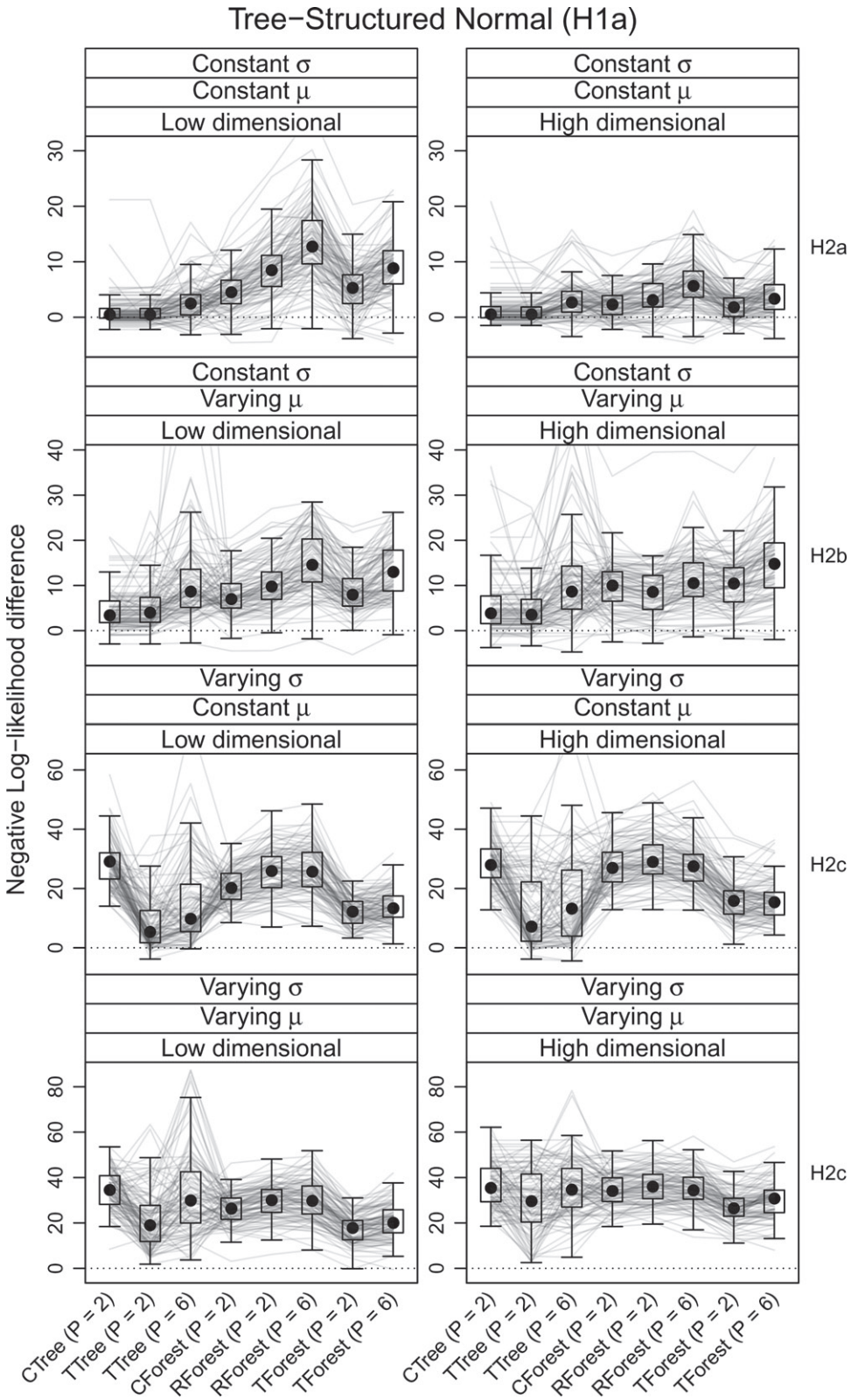

Figure 3. Simulation model (5). Negative log-likelihood differences for trees and forests in a conditional normal model with potential jumps in mean and variance. The negative log-likelihood difference was computed as the out-of-sample negative log-likelihood of each competitor minus the negative log-likelihood of the true data-generating process. Outliers were not plotted.

6.5. Results: Nonlinear Conditional Parameters (H1b)

The same hypotheses H1–4 were assessed but for nonlinear Friedman1-type conditional parameter functions instead of the tree-structured functions considered previously. More specifically, Figures 5 and 6 depict the negative log-likelihood differences based on 100 learning samples with normally distributed response ($N = 500$) and log-normally distributed

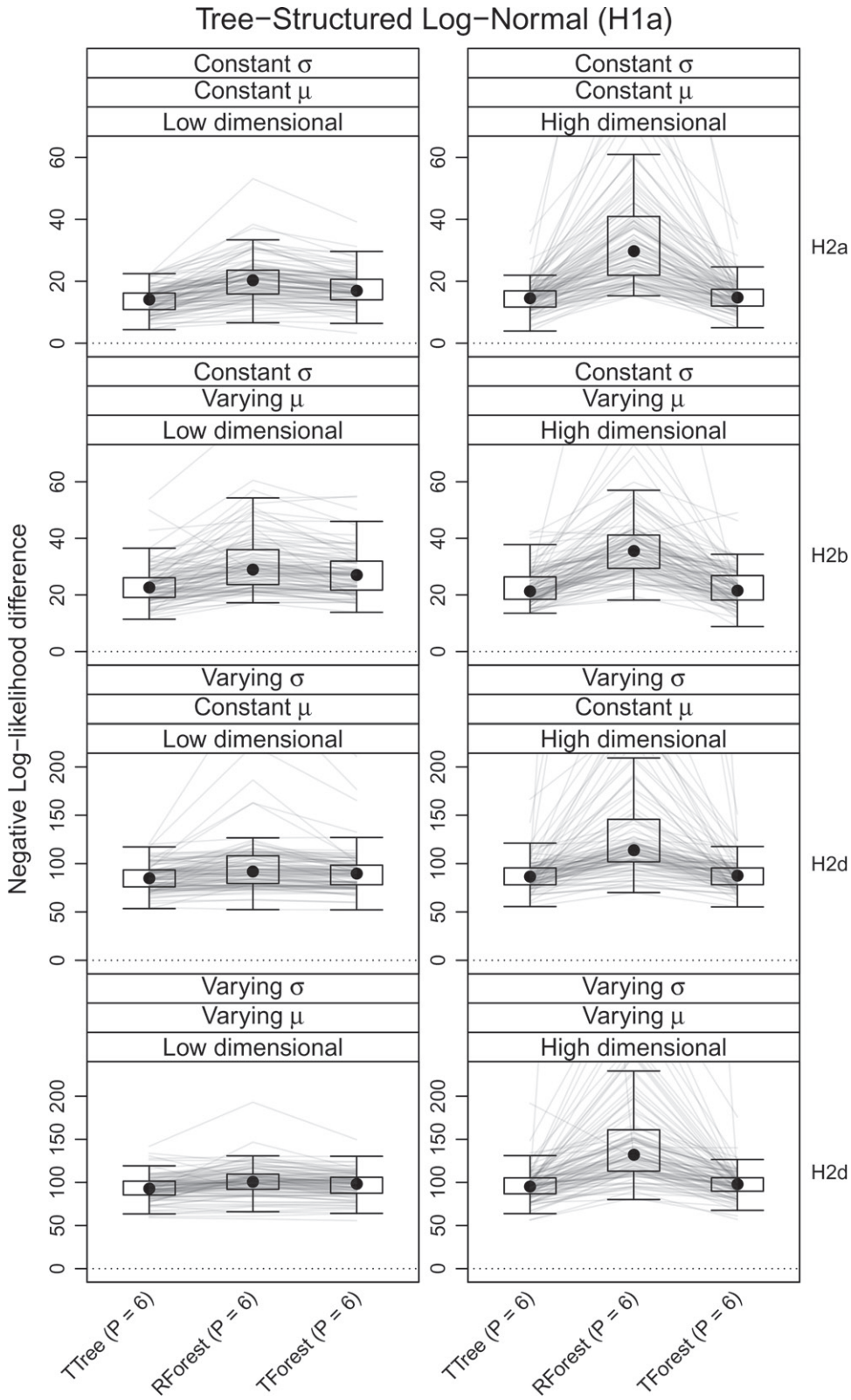

Figure 4. Simulation model (7). Negative log-likelihood differences for trees and forests in a conditional log-normal model with potential jumps in mean and variance. The negative log-likelihood difference was computed as the out-of-sample negative log-likelihood of each competitor minus by the negative log-likelihood of the true data-generating process. Outliers were not plotted.

response ($N = 2500$), respectively. We summarize the results as follows.

H1b: When a signal was present (rows 2–4), all random forest variants outperformed single trees under normality. Under nonnormality, this still holds for the random forest variants combined with flexible models ($P = 6$).

H2a: When there was no effect (top rows), CTree ($P = 2$) and TTree ($P = 2$) showed best resistance to overfitting under

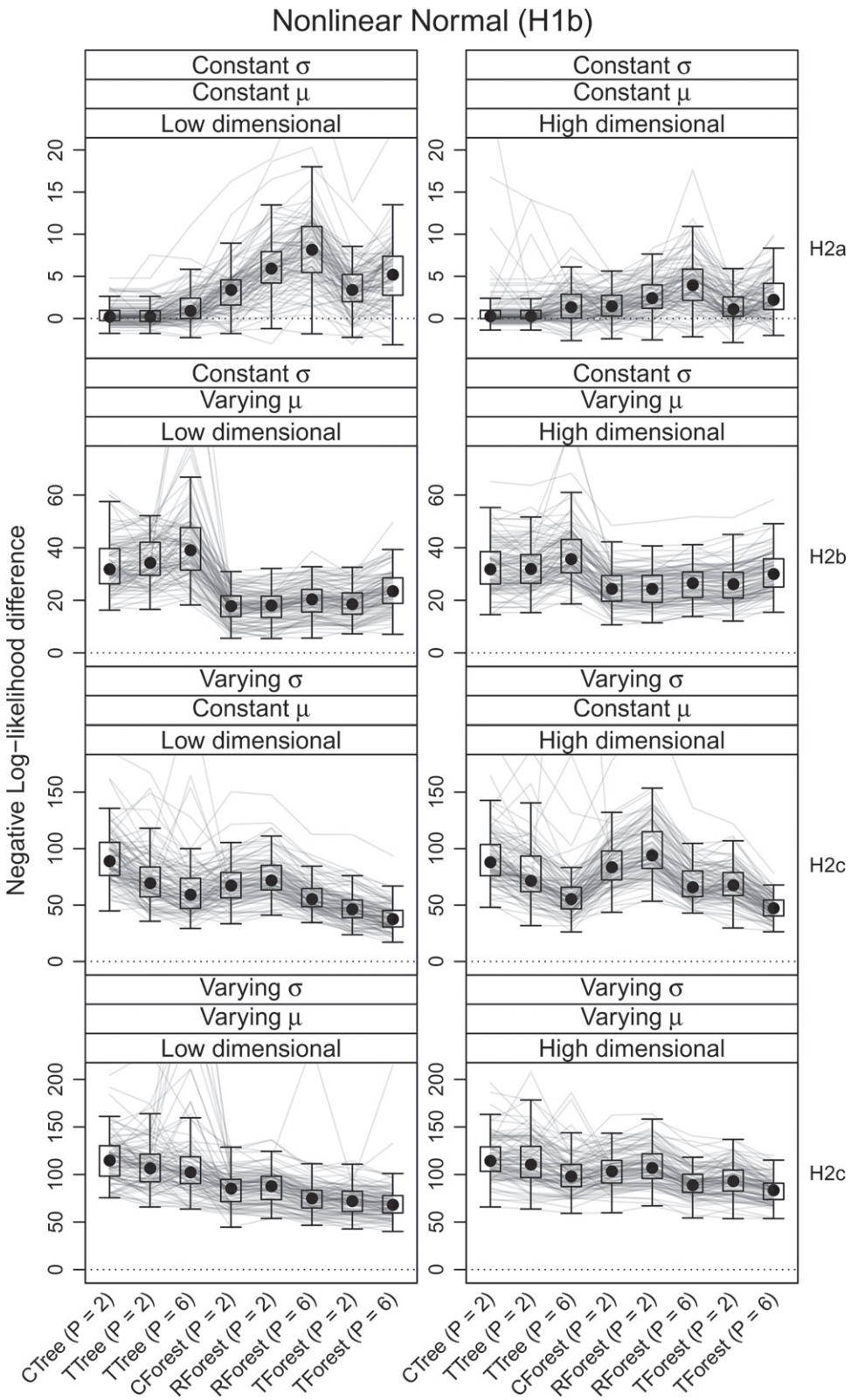

Figure 5. Simulation model (6). Negative log-likelihood differences for trees and forests in a conditional normal model with nonlinear functions defining mean and variance. The negative log-likelihood difference was computed as the out-of-sample negative log-likelihood of each competitor minus the negative log-likelihood of the true data-generating process. Outliers were not plotted.

normality. Under nonnormality, TTree ($P = 6$) still shows this behavior but the corresponding forests also perform similarly well.

H2b: All forest variants performed similarly well when predictor variables only had an effect on the mean (second rows).

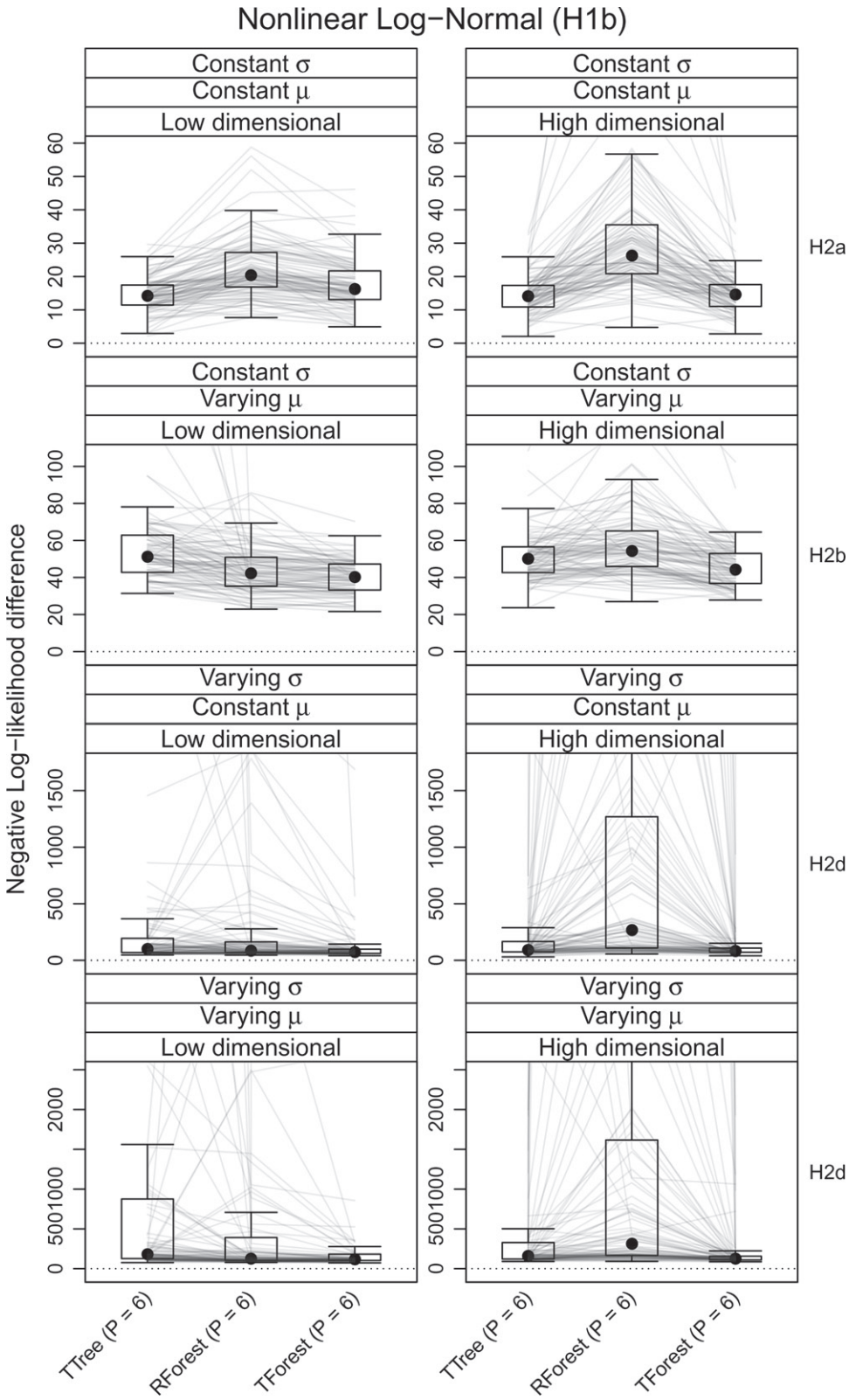

Figure 6. Simulation model (8). Negative log-likelihood differences for trees and forests in a conditional log-normal model with nonlinear functions defining mean and variance. The negative log-likelihood difference was computed as the out-of-sample negative log-likelihood of each competitor minus the negative log-likelihood of the true data-generating process. Outliers were not plotted.

H2c: Under normality, transformation forests performed best when some of the predictor variables also affected the variance (rows 3–4), where the classical procedures were not able to capture these changes appropriately.

H2d: Under nonnormality, transformation forests (with $P = 6$) still performed best (rows 3–4). However, in low dimensions the classical RForest also performs well albeit with a much larger variance than TForest.

H3: Under nonnormality, all trees and forests combined with flexible Bernstein polynomials ($P = 6$) clearly outperform all other methods. Under normality, the flexible models with $P = 6$ were sometimes slightly worse than the $P = 2$ models but often also a little bit better.

H4: In many situations, the picture in low-dimensional settings (left column) is quite similar to that in high-dimensional scenarios (right column). However, sometimes it can be seen that transformation forests stabilize transformation trees in the presence of high-dimensional noninformative predictor variables. Overall, the improvement of TForest over RForest seemed more pronounced in high dimensions.

As before, qualitatively the same patterns could be observed for the corresponding 10%, 50%, and 90% check risks (figures in online appendix) and thus 80% prediction intervals. In summary, our hypotheses H1–4 were found to describe the behavior of transformation trees and forests in this more complex setup well. The loss of using an overly complex model, such as a transformation model with $P = 6$, was tolerable in the simple normal setups but the gains, especially when parameters of a skewed response depend on the predictor variables, was found to be substantial.

6.6. Illustration: Swiss Body Mass Indices

Finally, we illustrate the applicability of transformation trees and forests in a realistic situation by modeling the conditional body mass index (BMI = weight (in kg)/height (in m)^2) distribution for Switzerland, based on 16,427 individuals aged between 18 and 74 years from the 2012 Swiss Health Survey (see Lohse et al. 2017, for a detailed description of the study). The predictor variables $\boldsymbol{x}$ included *sex* (male, female), *age* (in years), level of education (*edu*: I/mandatory, II/secondary, III/tertiary), *nationality* (Swiss/foreign), *region* (German/Romansh, French, Italian), and a number of lifestyle variables: *smoking* (never, former, light, moderate, heavy), fruit and vegetable consumption (*frveg*: high, low), *alcohol* intake (gram per day), and physical activity (*physcat*: 0, 1–2, >2 days per week). The transformation model

$$\mathbb{P}(\text{BMI} \le y \mid \text{sex}, \text{age}, \dots) = \Phi\left(\boldsymbol{a}_{\text{Bs},5}(y)^\top \boldsymbol{\vartheta}(\text{sex}, \text{age}, \dots)\right),$$

parameterizes the conditional transformation function by Bernstein polynomials of order $M = 5$ (i.e., $P = 6$ parameters). The parameters $\boldsymbol{\vartheta}$ of the polynomial may depend on the predictor variables in a potentially complex way, featuring interactions and nonlinearities. Transformation trees and forest allow such conditional parameter functions $\boldsymbol{\vartheta}(\boldsymbol{x})$, and thus the corresponding conditional BMI distributions, to be estimated in a data-driven way without any a priori specification. (A comparison of transformation forests to models with structured transformation functions is given in Hothorn (2018).)

The in-sample negative log-likelihood of the tree presented in Figure 1 was 43,079.42. The first split was in sex, within which three age groups were selected for both females (≤34, (34, 51]) and males (≤25, (25, 36], >36). Increasing age was mostly associated with increasing mean BMI whereas the sex difference also affected variance and skewness which were both higher for women. Further splits were in education as well as the lifestyle variables alcohol intake (*alcohol*) and physical activity (*physcat*). While these also affected mean BMI, the differences in variance were often even more pronounced, for example, node 5 versus 6, node 8 versus 9, or node 13 versus 14. These insights are interesting, but this transformation tree model is, of course, very rough.

A transformation forest allows smoother conditional parameter functions $\boldsymbol{\vartheta}(\boldsymbol{x})$ to be estimated. The negative log-likelihood was 42520.18 and thus an improvement over the negative log-likelihood 43079.42 of the transformation tree (see online appendix for cross-validation results). To gain some insight into the complex transformation forest model, a partial dependency plot for conditional deciles is shown in Figure 7. This visualizes

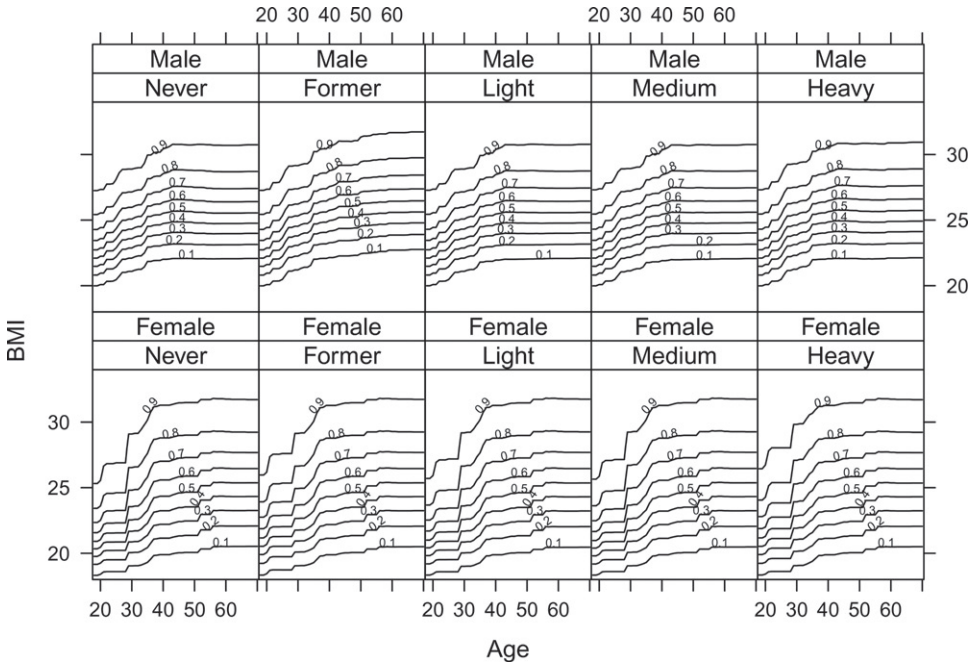

Figure 7. Body mass index (BMI). Partial dependency plot of conditional deciles estimated by a transformation forest with nonlinear transformation function.

the association between sex, smoking, age, and BMI. In general, the median BMI increased with age, as did the BMI variance. For males, there seemed to be a level-effect whose onset depended on smoking category. Females tended to higher BMI values, and the variance was larger compared to males. There seemed to be a bump in BMI values for females, roughly around 30 years. This corresponds to mothers giving birth to their first child around this age. It is important to note that the right-skewness of the conditional BMI distributions renders conditional normal distributions inappropriate, even under variance heterogeneity.

7. Discussion

Transformation forests, as well as the underlying transformation trees, can be understood as adaptive local likelihood estimators in the rather general parametric transformation family of distributions. Owing to possible interactions and nonlinear effects in an essentially unstructured conditional parameter function $\boldsymbol{\vartheta}(\boldsymbol{x})$, the resulting conditional distributions of the response may depend on the predictors in a very general way. The ability to model the impact of some predictors on the whole conditional distribution simultaneously, including its mean but also higher moments, is a unique feature of this novel member of the random forest family. The likelihood approach taken here also directly allows the procedures to be applied to randomly censored, truncated, or discrete observations (Hothorn, Möst, and Bühlmann 2018). A similar extension of random forests which is based on score functions was recently proposed by Athey, Tibshirani, and Wager (2019). Unlike transformation trees looking at all parameters $\boldsymbol{\vartheta}$ simultaneously, these "generalized random forests" apply CART-like regression trees to one single contrast of interest, for example a predictor-varying treatment effect.

The algorithmic internals of transformation trees are rooted in conditional inference trees (Hothorn, Hornik, and Zeileis 2006) and model-based recursive partitioning (Zeileis, Hothorn, and Hornik 2008) and inherit the unbiased variable selection property from these ancestors. Transformation forests also allow for unbiased variable importances (Strobl et al. 2007), including the internal handling of missing predictor variables (Hapfelmeier et al. 2014). An open-source implementation of transformation trees and transformation forests is available in the R add-on package `trtf` (Hothorn 2020b, see online appendix).

Transformation trees and forests make it easy to combine the flexible transformation function with classical additive effects, simply by adding them to the transformation function. A prominent application case are treatment effects, for example, in medicine, economics, or marketing, that can be included as $\mathbb{P}(Y \leq y \mid \text{treatment}) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} + \beta \cdot \text{treatment})$ featuring a binary treatment indicator variable. This model can also be combined with transformation forests in the same way as before, via their likelihood contributions $\ell_i(\boldsymbol{\vartheta}, \beta)$ and weights $w_i^N(\boldsymbol{x})$. This yields not only a conditional parameter function for the transformation parameters $\boldsymbol{\vartheta}(\boldsymbol{x})$ but additionally a personalized treatment effect $\beta(\boldsymbol{x})$. Recently, there has been increasing interest in using random forest algorithms for estimating such personalized treatment effects (Foster, Taylor, and Ruberg 2011; Seibold, Zeileis, and Hothorn 2016, 2018; Wager and Athey 2018) and transformation trees and forests can readily couple this with the flexibility of transformation models: Korepanova et al. (2020) provide empirical results in the context of transformation survival forests.

Breiman and Cutler's random forests are typically understood as being part of the machine learning inspired "algorithmic modeling culture" whereas additive models are rooted in the "data modeling culture" (Breiman 2001b). Transformation forests featuring predictor-varying effects $\boldsymbol{\beta}(\boldsymbol{x})$ of additional variables $\boldsymbol{u}$ (e.g., a priori known confounders in observational studies) offer a compromise between these two extreme points of view. The model $\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}) + \boldsymbol{u}^\top \boldsymbol{\beta}(\boldsymbol{x}))$ is structured additive in $\boldsymbol{u}$ with essentially unstructured intercept function $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x})$ and regression effects $\boldsymbol{\beta}(\boldsymbol{x})$ which, again, can be estimated by a transformation forest. This direct extension to the additional parameters $\boldsymbol{\beta}$ is possible because transformation trees and forests were specifically designed to detect and model general patterns of parameter instabilities. Transformation forests relying on the nonlinear Cox model $\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = 1 - \exp(-\exp(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} + \beta(\boldsymbol{x})))$ can be implemented by splitting scores with respect to β but not with respect to the nuisance parameters $\boldsymbol{\vartheta}$. Such a forest is much closer to the data modeling culture, because $\beta(\boldsymbol{x})$ has a clear interpretation as a log-hazard ratio and $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ is a log-cumulative baseline hazard function (see Korepanova et al. 2020, for details). Other examples of smooth transitions from simple "data models" to complex "algorithmic models" within the transformation family are discussed in Hothorn (2018).

Supplementary Materials

Computational details and software, algorithmic variants and their computational complexity, additional empirical results.

Acknowledgments

We would like to thank Heidi Seibold and Nicolai Meinshausen for discussions on an initial version. Comments by the handling editor and an anonymous referee were very helpful for revising the article.

Funding

Torsten Hothorn acknowledges financial support by the Swiss National Science Foundation (grant numbers IZSEZ0_177091 and 200021_184603).

ORCID

Torsten Hothorn http://orcid.org/0000-0001-8301-0471
Achim Zeileis http://orcid.org/0000-0003-0918-3766

References

Athey, S., Tibshirani, J., and Wager, S. (2019), "Generalized Random Forests," *The Annals of Statistics*, 47, 1148–1178. [1184,1195]

Biau, G. (2012), "Analysis of a Random Forests Model," *Journal of Machine Learning Research*, 13, 1063–1095. [1181,1186]

Biau, G., and Devroye, L. (2010), "On the Layered Nearest Neighbour Estimate, the Bagged Nearest Neighbour Estimate and the Random

Forest Method in Regression and Classification," *Journal of Multivariate Analysis*, 101, 2499–2518. [1186]

Biau, G., Devroye, L., and Lugosi, G. (2008), "Consistency of Random Forests and Other Averaging Classifiers," *Journal of Machine Learning Research*, 9, 2015–2033. [1181,1186]

Biau, G., and Scornet, E. (2016), "A Random Forest Guided Tour," *Test*, 25, 197–227. [1186]

Bloniarz, A., Wu, C., Yu, B., and Talwalkar, A. (2016), "Supervised Neighborhoods for Distributed Nonparametric Regression," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 1450–1459, available at *http://proceedings.mlr.press/v51/bloniarz16.pdf*. [1184]

Breiman, L. (2001a), "Random Forests," *Machine Learning*, 45, 5–32. [1186,1187,1188]

——— (2001b), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–231. [1195]

——— (2004), "Consistency for a Simple Model of Random Forests," Technical Report 670, Statistics Department, UCB, California, available at *http://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf*. [1186]

Brillinger, D. R. (1977), "Discussion of Stone (1977)," *The Annals of Statistics*, 5, 622–623. [1183]

Criminisi, A., Shotton, J., and Konukoglu, E. (2012), "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning," *Foundations and Trends in Computer Graphics and Vision*, 7, 81–227. [1182]

Curtis, S. M., and Ghosh, S. K. (2011), "A Variable Selection Approach to Monotonic Regression With Bernstein Polynomials," *Journal of Applied Statistics*, 38, 961–976. [1185]

Farouki, R. T. (2012), "The Bernstein Polynomial Basis: A Centennial Retrospective," *Computer Aided Geometric Design*, 29, 379–419. [1185]

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011), "Subgroup Identification From Randomized Clinical Trial Data," *Statistics in Medicine*, 30, 2867–2880. [1195]

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–67. [1187]

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007), "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society*, Series B, 69, 243–268. [1189]

Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. [1188]

Hapfelmeier, A., Hothorn, T., Ulm, K., and Strobl, C. (2014), "A New Variable Importance Measure for Random Forests With Missing Data," *Statistics and Computing*, 24, 21–34. [1195]

Hothorn, T. (2018), "Top-Down Transformation Choice," *Statistical Modelling*, 18, 274–298. [1186,1194,1195]

——— (2020a), "Most Likely Transformations: The **mlt** Package," *Journal of Statistical Software*, 92, 1–68. [1187]

——— (2020b), "trtf: Transformation Trees and Forests," R Package Version 0.3-7, available at *https://CRAN.R-project.org/package=trtf*. [1195]

Hothorn, T., Hornik, K., and Zeileis, A. (2006), "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15, 651–674. [1185,1188,1195]

Hothorn, T., Kneib, T., and Bühlmann, P. (2014), "Conditional Transformation Models," *Journal of the Royal Statistical Society*, Series B, 76, 3–27. [1182,1184]

Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004), "Bagging Survival Trees," *Statistics in Medicine*, 23, 77–91. [1181,1184]

Hothorn, T., Möst, L., and Bühlmann, P. (2018), "Most Likely Transformations," *Scandinavian Journal of Statistics*, 45, 110–134. [1182,1184,1185,1187,1195]

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random Survival Forests," *The Annals of Applied Statistics*, 2, 841–860. [1182]

Korepanova, N., Seibold, H., Steffen, V., and Hothorn, T. (2020), "Survival Forests Under Test: Impact of the Proportional Hazards Assumption on Prognostic and Predictive Forests for ALS Survival," *Statistical Methods in Medical Research*, 29, 1403–1419. [1195]

Lin, Y., and Jeon, Y. (2006), "Random Forests and Adaptive Nearest Neighbors," *Journal of the American Statistical Association*, 101, 578–590. [1182,1184,1186]

Lindsey, J. K. (1996), *Parametric Statistical Inference*, Oxford: Clarendon Press. [1184]

Lohse, T., Rohrmann, S., Faeh, D., and Hothorn, T. (2017), "Continuous Outcome Logistic Regression for Analyzing Body Mass Index Distributions," *F1000Research*, 6, 1933. [1194]

Meinshausen, N. (2006), "Quantile Regression Forests," *Journal of Machine Learning Research*, 7, 983–999. [1181,1182,1184,1186,1188]

Pinson, P. (2013), "Wind Energy: Forecasting Challenges for Its Operational Management," *Statistical Science*, 28, 564–585. [1181]

Pratola, M., Chipman, H., George, E. I., and McCulloch, R. (2020), "Heteroscedastic BART via Multiplicative Regression Trees," *Journal of Computational and Graphical Statistics*, 29, 405–417. [1186]

Scornet, E. (2016), "On the Asymptotics of Random Forests," *Journal of Multivariate Analysis*, 146, 72–83. [1187]

Scornet, E., Biau, G., and Vert, J.-P. (2015), "Consistency of Random Forests," *The Annals of Statistics*, 43, 1716–1741. [1181,1186]

Seibold, H., Zeileis, A., and Hothorn, T. (2016), "Model-Based Recursive Partitioning for Subgroup Analyses," *International Journal of Biostatistics*, 12, 45–63. [1195]

——— (2018), "Individual Treatment Effect Prediction for Amyotrophic Lateral Sclerosis Patients," *Statistical Methods in Medical Research*, 27, 3104–3125. [1184,1195]

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008), "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, 9, 1–11. [1186]

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007), "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics*, 8, 25. [1188,1195]

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567. [1183]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [1184,1195]

Zeileis, A., and Hothorn, T. (2013), "A Toolbox of Permutation Tests for Structural Change," *Statistical Papers*, 54, 931–954. [1185]

Zeileis, A., Hothorn, T., and Hornik, K. (2008), "Model-Based Recursive Partitioning," *Journal of Computational and Graphical Statistics*, 17, 492–514. [1182,1183,1184,1185,1195]