Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 45: 110–134, 2018
doi: 10.1111/sjos.12291

Most Likely Transformations

TORSTEN HOTHORN
Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich
LISA MÖST
Institut für Statistik, Ludwig-Maximilians-Universität München
PETER BÜHLMANN
Seminar für Statistik, ETH Zürich

ABSTRACT. We propose and study properties of maximum likelihood estimators in the class of conditional transformation models. Based on a suitable explicit parameterization of the unconditional or conditional transformation function, we establish a cascade of increasingly complex transformation models that can be estimated, compared and analysed in the maximum likelihood framework. Models for the unconditional or conditional distribution function of any univariate response variable can be set up and estimated in the same theoretical and computational framework simply by choosing an appropriate transformation function and parameterization thereof. The ability to evaluate the distribution function directly allows us to estimate models based on the exact likelihood, especially in the presence of random censoring or truncation. For discrete and continuous responses, we establish the asymptotic normality of the proposed estimators. A reference software implementation of maximum likelihood-based estimation for conditional transformation models that allows the same flexibility as the theory developed here was employed to illustrate the wide range of possible applications.

Key words: censoring, conditional distribution function, conditional quantile function, distribution regression, transformation model, truncation

1. Introduction

In a broad sense, we can understand all statistical models as models of distributions or certain characteristics thereof, especially the mean. All distributions $\mathbb{P}_Y$ for at least ordered responses Y can be characterized by their distribution, quantile, density, odds, hazard or cumulative hazard functions. In a fully parametric setting, all these functions have been specified up to unknown parameters, and the ease of interpretation can guide us in looking at the appropriate function. In the semi-parametric and non-parametric contexts, however, the question arises how we can obtain an estimate of one of these functions without assuming much about their shape. For the direct estimation of distribution functions, we deal with monotonic functions in the unit interval, whereas for densities, we need to make sure that the estimator integrates to one. The hazard function comes with a positivity constraint, and monotonicity is required for the positive cumulative hazard function. These computationally inconvenient restrictions disappear completely only when the log-hazard function is estimated, and this explains the plethora of research papers following this path. However, the lack of any structure in the log-hazard function comes at a price. A too-erratic behaviour of estimates of the log-hazard function has to be prevented by some smoothness constraint; this makes classical likelihood inference impossible. The novel characterization and subsequent estimation of distributions via their transformation function in a broad class of transformation models that are developed in this paper can be interpreted as a compromise between structure (monotonicity) and ease of parameterization, estimation and inference. This transformation approach to modelling

and estimation allows standard likelihood inference in a large class of models that have so far commonly been dealt with by other inference procedures.

Since the introduction of transformation models based on non-linear transformations of some response variable by Box & Cox (1964), this attractive class of models has received much interest. In regression problems, transformation models can be understood as models for the conditional distribution function and are sometimes referred to as 'distribution regression', in contrast to their 'quantile regression' counterpart (Chernozhukov *et al.*, 2013). Traditionally, the models were actively studied and applied in the analysis of ordered categorical or censored responses. Recently, transformation models for the direct estimation of conditional distribution functions for arbitrary responses received interest in the context of counterfactual distributions (Chernozhukov *et al.*, 2013), probabilistic forecasting (Gneiting & Katzfuss, 2014), distribution and quantile regression (Leorato & Peracchi, 2015; Rothe & Wied, 2013), probabilistic index models (Thas *et al.*, 2012) and conditional transformation models (Hothorn *et al.*, 2014). The core idea of any transformation model is the application of a strictly monotonic transformation function h for the reformulation of an unknown distribution function $\mathbb{P}(Y \leq y)$ as $\mathbb{P}(h(Y) \leq h(y))$, where the unknown transformation function h is estimated from the data. Transformation models have received attention especially in situations where the likelihood contains terms involving the conditional distribution function $\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(h(y \mid \boldsymbol{x}))$ with inverse link function F_Z, most importantly for censored, truncated and ordered categorical responses. For partially linear transformation models with transformation function $h(y \mid \boldsymbol{x}) = h_Y(y) + h_x(\boldsymbol{x})$, much emphasis has been given to estimation procedures treating the baseline transformation h_Y (e.g. the log-cumulative baseline hazard function in the Cox model) as a high-dimensional nuisance parameter. Prominent members of these estimation procedures are the partial likelihood estimator and approaches influenced by the estimation equations introduced by Cheng *et al.* (1995). Once an estimate for the shift h_x is obtained, the baseline transformation h_Y is then typically estimated by the non-parametric maximum likelihood estimator (see, e.g. Cheng *et al.*, 1997). An overview of the extensive literature on the simultaneous non-parametric maximum likelihood estimation of h_Y and h_x, that is, estimation procedures not requiring an explicit parameterization of h_Y, for censored continuous responses is given in Zeng & Lin (2007).

An explicit parameterization of h_Y is common in models of ordinal responses (Tutz, 2012). For survival times, Kooperberg *et al.* (1995) introduced a cubic spline parameterization of the log-conditional hazard function with the possibility of response-varying effects and estimated the corresponding models by maximum likelihood. Crowther & Lambert (2014) followed up on this suggestion and used restricted cubic splines. Many authors studied penalized likelihood approaches for spline approximations of the baseline hazard function in a Cox model, for example, Ma *et al.* (2014). Less frequently, the transformation function h_Y was modelled directly. Mallick & Walker (2003), Chang *et al.* (2005) and McLain & Ghosh (2013) used Bernstein polynomials for h_Y, and Royston & Parmar (2002) proposed a maximum likelihood approach using cubic splines for modelling h_Y and also time-varying effects. The connection between these different transformation models is difficult to see because most authors present their models in the relatively narrow contexts of survival or ordinal data. The lack of a general understanding of transformation models made the development of novel approaches in this model class burdensome. Hothorn *et al.* (2014) decoupled the parameterization of the conditional transformation function $h(y \mid \boldsymbol{x})$ from the estimation procedure and showed that many interesting and novel models can be understood as transformation models. The boosting-based optimization of proper scoring rules, however, was only developed for uncensored and right-censored observations in the absence of truncation and requires the numerical approximation of the true target function. In a similar spirit, Chernozhukov *et al.* (2013) applied the connection

$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(\mathbb{1}(Y \leq y) \mid \boldsymbol{X} = \boldsymbol{x})$ for estimation in the response-varying effects transformation model $\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(h_Y(y) - \boldsymbol{x}^\top \boldsymbol{\beta}(y))$; this approach can be traced back to Foresi & Peracchi (1995).

A drawback of all but the simplest transformation models is the lack of a likelihood estimation procedure. Furthermore, although important connections to other models have been known for some time (Doksum & Gasko, 1990), it is often not easy to see how broad and powerful the class of transformation models actually is. We address these issues and embed the estimation of unconditional and conditional distribution functions of arbitrary univariate random variables under all forms of random censoring and truncation into a common theoretical and computational likelihood-based framework. In a nutshell, we show in Section 2 that all distributions can be generated by a strictly monotonic transformation of some absolute continuous random variable. The likelihood function of the transformed variable can then be characterized by this transformation function. The parameters of appropriate parameterizations of the transformation function, and thus the parameters of the conditional distribution function in which we are interested, can then be estimated by maximum likelihood under simple linear constraints that allow classical asymptotic likelihood inference, as will be shown in Section 3. Many classical and contemporary models are introduced as special cases of this framework. In particular, all transformation models sketched in this introduction can be understood and estimated in this novel likelihood-based framework. Extensions of classical and contemporary transformation models as well as some novel models are derived from our unified theoretical framework of transformation functions in Section 4, and their empirical performance is illustrated and evaluated in Section 5.

2. The likelihood of transformations

Let $(\Omega, \mathfrak{A}, \mathbb{P})$ denotes a probability space and $(\Xi, \mathfrak{C})$ a measurable space with at least ordered sample space Ξ. We are interested in inference about the distribution $\mathbb{P}_Y$ of a random variable Y, that is, the probability space $(\Xi, \mathfrak{C}, \mathbb{P}_Y)$ defined by the $\mathfrak{A} - \mathfrak{C}$ measurable function $Y : \Omega \to \Xi$. For the sake of notational simplicity, we present our results for the unconditional case first; regression models are discussed in Section 4.2. The distribution $\mathbb{P}_Y = f_Y \odot \mu$ is dominated by some measure μ and characterized by its density function f_Y, distribution function $F_Y(y) = \mathbb{P}_Y(\{\xi \in \Xi \mid \xi \leq y\})$, quantile function $F_Y^{-1}(p) = \inf\{y \in \Xi \mid F_Y(y) \geq p\}$, odds function $O_Y(y) = F_Y(y)/(1 - F_Y(y))$, hazard function $\lambda_Y(y) = f_Y(y)/(1 - F_Y(y))$ or cumulative hazard function $\lambda_Y(y) = -\log(1 - F_Y(y))$. For notational convenience, we assume strict monotonicity of F_Y, that is, $F_Y(y_1) < F_Y(y_2) \, \forall y_1 < y_2 \in \Xi$. Our aim is to obtain an estimate $\hat{F}_{Y,N}$ of the distribution function F_Y from a random sample $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} \mathbb{P}_Y$. In the following, we will show that one can always write this potentially complex distribution function F_Y as the composition of a much simpler and *a priori* specified distribution function F_Z and a strictly monotonic transformation function h. The task of estimating F_Y is then reduced to obtaining an estimate $\hat{h}_N$. The latter exercise, as we will show in this paper, is technically and conceptually attractive.

Let $(\mathbb{R}, \mathfrak{B})$ denotes the Euclidian space with Borel σ-algebra and $Z : \Omega \to \mathbb{R}$ an $\mathfrak{A} - \mathfrak{B}$ measurable function such that the distribution $\mathbb{P}_Z = f_Z \odot \mu_L$ is absolutely continuous (μ_L denotes the Lebesgue measure) in the probability space $(\mathbb{R}, \mathfrak{B}, \mathbb{P}_Z)$. Let F_Z and F_Z^{-1} denote the corresponding distribution and quantile functions. We furthermore assume $0 < f_Z(z) < \infty \, \forall z \in \mathbb{R}$, $F_Z(-\infty) = 0$ and $F_Z(\infty) = 1$ 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, standard logistic (SL) and minimum extreme value (MEV) distribution with distribution functions $F_Z(z) = \Phi(z)$, $F_Z(z) = F_{\text{SL}}(z) = (1 + \exp(-z))^{-1}$ and $F_Z(z) = F_{\text{MEV}}(z) = 1 - \exp(-\exp(z))$, respectively. In the first step, we will show that there always exists a unique and strictly monotonic transformation g such that the unknown and potentially complex distribution $\mathbb{P}_Y$ that we are interested in can be generated from the simple and known distribution $\mathbb{P}_Z$ via $\mathbb{P}_Y = \mathbb{P}_{g \circ Z}$. More formally, let $g : \mathbb{R} \to \Xi$ denotes a $\mathfrak{B} - \mathfrak{C}$ measurable function. The composition $g \circ Z$ is a random variable on $(\Xi, \mathfrak{C}, \mathbb{P}_{g \circ Z})$. We can now formulate the existence and uniqueness of g as a corollary to the probability integral transform.

Corollary 1. *For all random variables Y and Z, there exists a unique strictly monotonically increasing transformation g, such that $\mathbb{P}_Y = \mathbb{P}_{g \circ Z}$.*

Proof. Let $g = F_Y^{-1} \circ F_Z$ and $Z \sim \mathbb{P}_Z$. Then $U := F_Z(Z) \sim \text{U}[0, 1]$ and $Y = F_Y^{-1}(U) \sim \mathbb{P}_Y$ by the probability integral transform. Let $h : \Xi \to \mathbb{R}$, such that $F_Y(y) = F_Z(h(y))$. From $F_Y(y) = (F_Z \circ F_Z^{-1} \circ F_Y)(y) \iff h = F_Z^{-1} \circ F_Y$, we get the uniqueness of h and therefore g. The quantile function F_Z^{-1} and the distribution function F_Y exist by assumption and are both strictly monotonic and right continuous. Therefore, h is strictly monotonic and right continuous and so is g. □

Corollary 2. *For $\mu = \mu_L$, we have $g = h^{-1}$ and $h'(y) = \frac{\partial h(y)}{\partial y} = f_Z((F_Z^{-1} \circ F_Y)(y))^{-1} f_Y(y)$.*

This result for absolutely continuous random variables Y can be found in many textbooks (Lindsey, 1996, e.g.); Corollary 1 also covers the discrete case.

Corollary 3. *For the counting measure $\mu = \mu_C$, $h = F_Z^{-1} \circ F_Y$ is a right-continuous step function because F_Y is a right-continuous step function with steps at $y \in \Xi$.*

We now characterize the distribution F_Y by the corresponding transformation function h, set up the corresponding likelihood of such a transformation function and estimate the transformation function based on this likelihood. Let $\mathcal{H} = \{h : \Xi \to \mathbb{R} \mid \mathfrak{C} - \mathfrak{B} \text{ measurable}, h(y_1) < h(y_2)\, \forall y_1 < y_2 \in \Xi\}$ denote the space of all strictly monotonic transformation functions. With the transformation function h, we can evaluate F_Y as $F_Y(y \mid h) = F_Z(h(y))\, \forall y \in \Xi$. Therefore, we only need to study the transformation function h; the inverse transformation $g = h^{-1}$ (Bickel *et al.*, 1993, used to define a 'group model' by) is not necessary in what follows. The density for absolutely continuous variables Y ($\mu = \mu_L$) is now given by $f_Y(y \mid h) = f_Z(h(y))h'(y)$. For discrete responses Y ($\mu = \mu_C$) with finite sample space $\Xi = \{y_1, \ldots, y_K\}$, the density is

$$f_Y(y_k \mid h) = \begin{cases} F_Z(h(y_k)) & k = 1 \\ F_Z(h(y_k)) - F_Z(h(y_{k-1})) & k = 2, \ldots, K-1 \\ 1 - F_Z(h(y_{k-1})) & k = K, \end{cases}$$

and for countably infinite sample spaces $\Xi = \{y_1, y_2, y_3, \ldots\}$, we get the density

$$f_Y(y_k \mid h) = \begin{cases} F_Z(h(y_k)) & k = 1 \\ F_Z(h(y_k)) - F_Z(h(y_{k-1})) & k > 1. \end{cases}$$

With the conventions $F_Z(h(y_0)) := F_Z(h(-\infty)) := 0$ and $F_Z(h(y_K)) := F_Z(h(\infty)) := 1$, we use the more compact notation $f_Y(y_k \mid h) = F_Z(h(y_k)) - F_Z(h(y_{k-1}))$ in the sequel.

For a given transformation function h, the likelihood contribution of a datum $C = (\underline{y}, \bar{y}] \in \mathfrak{C}$ is defined in terms of the distribution function (Lindsey, 1996)

$$\mathcal{L}(h \mid Y \in C) := \int_C f_Y(y \mid h) d\mu(y) = F_Z(h(\bar{y})) - F_Z(h(\underline{y})).$$

This 'exact' definition of the likelihood applies to most practical situations of interest and, in particular, allows discrete and (conceptually) continuous as well as censored or truncated observations C. For a discrete response y_k, we have $\bar{y} = y_k$ and $\underline{y} = y_{k-1}$, such that $\mathcal{L}(h \mid Y = y_k) = f_Y(y_k \mid h) = F_Z(h(\bar{y})) - F_Z(h(\underline{y}))$. For absolutely continuous random variables Y, we almost always observe an imprecise datum $(\underline{y}, \bar{y}] \subset \mathbb{R}$ and, for short intervals $(\underline{y}, \bar{y}]$, approximate the exact likelihood $\mathcal{L}(h \mid Y \in (\underline{y}, \bar{y}])$ by the term $(\bar{y} - \underline{y}) f_Y(y \mid h)$ or simply $f_Y(y \mid h)$ with $y = (\underline{y} + \bar{y})/2$ (Lindsey, 1999). This approximation only works for relatively precise measurements, that is, short intervals. If longer intervals are observed, one speaks of 'censoring' and relies on the exact definition of the likelihood contribution instead of using the aforementioned approximation (Klein & Moeschberger, 2003). In summary, the likelihood contribution of a conceptually 'exact continuous' or left-censored, right-censored or interval-censored continuous or discrete observation $(\underline{y}, \bar{y}]$ is given by

$$\mathcal{L}(h \mid Y \in (\underline{y}, \bar{y}]) \begin{cases} \approx f_Z(h(y))h'(y) & y = (\underline{y} + \bar{y})/2 \in \Xi \quad \text{'exact continuous'} \\ = 1 - F_Z(h(\underline{y})) & y \in (\underline{y}, \infty) \cap \Xi \quad \text{'right censored'} \\ = F_Z(h(\bar{y})) & y \in (-\infty, \bar{y}] \cap \Xi \quad \text{'left censored'} \\ = F_Z(h(\bar{y})) - F_Z(h(\underline{y})) & y \in (\underline{y}, \bar{y}] \cap \Xi \quad \text{'interval censored'}, \end{cases}$$

under the assumption of random censoring. The likelihood is more complex under dependent censoring (Klein & Moeschberger, 2003), but we will not elaborate on this issue. The likelihood contribution $\mathcal{L}(h \mid Y \in (y_k, y_{k-1}])$ of an ordered factor in category y_k is equivalent to the term $\mathcal{L}(h \mid Y \in (\underline{y}, \bar{y}])$ contributed by an interval-censored observation $(\underline{y}, \bar{y}]$, when category y_k is defined by the interval $(\underline{y}, \bar{y}]$. Thus, the expression $F_Z(h(\bar{y})) - F_Z(h(\underline{y}))$ for the likelihood contribution reflects the equivalence of interval censoring and categorization at corresponding cut-off points.

For truncated observations in the interval $(y_l, y_r] \subset \Xi$, the aforementioned likelihood contribution is defined in terms of the distribution function conditional on the truncation

$$F_Y(y \mid Y \in (y_l, y_r]) = F_Z(h(y) \mid Y \in (y_l, y_r]) = \frac{F_Z(h(y))}{F_Z(h(y_r)) - F_Z(h(y_l))} \quad \forall y \in (y_l, y_r],$$

and thus, the likelihood contribution changes to (Klein & Moeschberger, 2003)

$$\frac{\mathcal{L}(h \mid Y \in (\underline{y}, \bar{y}])}{F_Z(h(y_r)) - F_Z(h(y_l))} = \frac{\mathcal{L}(h \mid Y \in (\underline{y}, \bar{y}])}{\mathcal{L}(h \mid Y \in (y_l, y_r])} \quad \text{when } y_l < \underline{y} < \bar{y} \leq y_r.$$

It is important to note that the likelihood is always *defined* in terms of a distribution function (Lindsey, 1999) and it therefore makes sense to directly model the distribution function of interest. The ability to uniquely characterize this distribution function by the transformation function h gives rise to the following definition of an estimator $\hat{h}_N$.

Definition 1 (Most likely transformation). *Let $C_1, \ldots, C_N$ denotes an independent sample of possibly randomly censored or truncated observations from $\mathbb{P}_Y$. The estimator*

$$\hat{h}_N := \arg\max_{\tilde{h} \in \mathcal{H}} \sum_{i=1}^{N} \log(\mathcal{L}(\tilde{h} \mid Y \in C_i))$$

is called the most likely transformation.

Log-concavity of f_Z ensures concavity of the log-likelihood (except when all observations are right censored) and thus ensures the existence and uniqueness of $\hat{h}_N$.

Many distributions are defined by a transformation function h, for example, the Box–Cox power exponential family (Stasinopoulos & Rigby, 2007), the sinh-arcsinh distributions (Jones & Pewsey, 2009) or the T-X family of distributions (Alzaatreh *et al.*, 2013). In what follows, we do not assume any specific form of the transformation function but parameterize h in terms of basis functions. We now introduce such a parameterization, a corresponding family of distributions, a maximum likelihood estimator and a large class of models for unconditional and conditional distributions.

3. Transformation analysis

We parameterize the transformation function $h(y)$ as a linear function of its basis-transformed argument y using a basis function $\boldsymbol{a} : \Xi \to \mathbb{R}^P$, such that $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}, \boldsymbol{\vartheta} \in \mathbb{R}^P$. The choice of the basis function $\boldsymbol{a}$ is problem specific and will be discussed in Section 4. The likelihood $\mathcal{L}$ only requires evaluation of h, and only an approximation thereof using the Lebesgue density of 'exact continuous' observations makes the evaluation of the first derivative of $h(y)$ with respect to y necessary. In this case, the derivative with respect to y is given by $h'(y) = \boldsymbol{a}'(y)^\top \boldsymbol{\vartheta}$, and we assume that $\boldsymbol{a}'$ is available. In the following, we will write $h = \boldsymbol{a}^\top \boldsymbol{\vartheta}$ and $h' = \boldsymbol{a}'^\top \boldsymbol{\vartheta}$ for the transformation function and its first derivative, omitting the argument y, and we assume that both functions are bounded away from $-\infty$ and ∞. For a specific choice of F_Z and $\boldsymbol{a}$, the transformation family of distributions consists of all distributions $\mathbb{P}_Y$ whose distribution function F_Y is given as the composition $F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta}$; this family can be formally defined as follows.

Definition 2 (Transformation family). *The distribution family* $\mathbb{P}_{Y,\Theta} = \{F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta} \mid \boldsymbol{\vartheta} \in \Theta\}$ *with parameter space* $\Theta = \{\boldsymbol{\vartheta} \in \mathbb{R}^P \mid \boldsymbol{a}^\top \boldsymbol{\vartheta} \in \mathcal{H}\}$ *is called transformation family of distributions* $\mathbb{P}_{Y,\boldsymbol{\vartheta}}$ *with transformation functions* $\boldsymbol{a}^\top \boldsymbol{\vartheta} \in \mathcal{H}$, *$\mu$-densities* $f_Y(y \mid \boldsymbol{\vartheta}), y \in \Xi$, *and error distribution function* F_Z.

The classical definition of a transformation family relies on the idea of invariant distributions, that is, only the parameters of a distribution are changed by a transformation function but the distribution itself is not changed. The normal family characterized by affine transformations is the most well-known example (e.g. Fraser, 1968; Lindsey, 1996). Here, we explicitly allow and encourage transformation functions that change the shape of the distribution. The transformation function $\boldsymbol{a}^\top \boldsymbol{\vartheta}$ is, at least in principle, flexible enough to generate any distribution function $F_Y = F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta}$ from the distribution function F_Z. We borrow the term 'error distribution' function for F_Z from Fraser (1968), because Z can be understood as an error term in some of the models discussed in Section 4. The problem of estimating the unknown transformation function h, and thus the unknown distribution function F_Y, reduces to the problem of estimating the parameter vector $\boldsymbol{\vartheta}$ through maximization of the likelihood function. We assume that the basis function $\boldsymbol{a}$ is such that the parameters $\boldsymbol{\vartheta}$ are identifiable.

Definition 3 (Maximum likelihood estimator). $\hat{\boldsymbol{\vartheta}}_N := \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} \log(\mathcal{L}(\boldsymbol{a}^\top \boldsymbol{\vartheta} \mid Y \in C_i))$

Based on the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}_N$, we define plug-in estimators of the most likely transformation function and the corresponding estimator of our target distribution F_Y as $\hat{h}_N := \boldsymbol{a}^\top \hat{\boldsymbol{\vartheta}}_N$ and $\hat{F}_{Y,N} := F_Z \circ \hat{h}_N$. Because the problem of estimating an unknown

distribution function is now embedded in the maximum likelihood framework, the asymptotic analysis benefits from standard results on the asymptotic behaviour of maximum likelihood estimators. We begin with deriving the score function and Fisher information. The score contribution of an 'exact continuous' observation $y = (\underline{y} + \bar{y})/2$ from an absolutely continuous distribution is approximated by the gradient of the log-density

$$\boldsymbol{s}(\boldsymbol{\vartheta} \mid Y \in (\underline{y}, \bar{y}]) \approx \boldsymbol{a}(y) \frac{f_Z'(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta})}{f_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta})} + \frac{\boldsymbol{a}'(y)}{\boldsymbol{a}'(y)^\top \boldsymbol{\vartheta}}. \tag{1}$$

For an interval-censored or discrete observation $\underline{y}$ and $\bar{y}$ (the constant terms $F_Z(\boldsymbol{a}(-\infty)^\top \boldsymbol{\vartheta}) = F_Z(-\infty) = 0$ and $F_Z(\boldsymbol{a}(\infty)^\top \boldsymbol{\vartheta}) = F_Z(\infty) = 1$ vanish), the score contribution is

$$\boldsymbol{s}(\boldsymbol{\vartheta} \mid Y \in (\underline{y}, \bar{y}]) = \frac{f_Z(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\bar{y}) - f_Z(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\underline{y})}{F_Z(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta})}. \tag{2}$$

For a truncated observation, the score function is $\boldsymbol{s}(\boldsymbol{\vartheta} \mid Y \in (\underline{y}, \bar{y}]) - \boldsymbol{s}(\boldsymbol{\vartheta} \mid Y \in (y_l, y_r])$.

The contribution of an 'exact continuous' observation y from an absolutely continuous distribution to the Fisher information is approximately

$$\boldsymbol{F}(\boldsymbol{\vartheta} \mid Y \in (\underline{y}, \bar{y}]) \approx -\left(\boldsymbol{a}(y) \boldsymbol{a}(y)^\top \left\{ \frac{f_Z''(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta})}{f_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta})} - \left[\frac{f_Z'(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta})}{f_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta})} \right]^2 \right\} - \frac{\boldsymbol{a}'(y) \boldsymbol{a}'(y)^\top}{(\boldsymbol{a}'(y)^\top \boldsymbol{\vartheta})^2} \right). \tag{3}$$

For a censored or discrete observation, we have the following contribution to the Fisher information

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{\vartheta} \mid Y \in (\underline{y}, \bar{y}]) = -\Bigg\{ & \frac{f_Z'(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\bar{y}) \boldsymbol{a}(\bar{y})^\top - f_Z'(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\underline{y}) \boldsymbol{a}(\underline{y})^\top}{F_Z(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta})} \\ & - \frac{[f_Z(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\bar{y}) - f_Z(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\underline{y})]}{[F_Z(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta})]^2} \\ & \times [f_Z(\boldsymbol{a}(\bar{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\bar{y})^\top - f_Z(\boldsymbol{a}(\underline{y})^\top \boldsymbol{\vartheta}) \boldsymbol{a}(\underline{y})^\top] \Bigg\}. \end{aligned} \tag{4}$$

For a truncated observation, the Fisher information is given by $\boldsymbol{F}(\boldsymbol{\vartheta} \mid Y \in (\underline{y}, \bar{y}]) - \boldsymbol{F}(\boldsymbol{\vartheta} \mid Y \in (y_l, y_r])$.

We will first discuss the asymptotic properties of the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}_N$ in the parametric setting with fixed parameters $\boldsymbol{\vartheta}$ in both the discrete and continuous case. For continuous variables Y and a transformation function parameterized using a Bernstein polynomial, results for sieve maximum likelihood estimation, where the number of parameters increases with N, are then discussed in Section 3.2.

3.1. Parametric inference

Conditions on the densities of the error distribution f_Z and the basis functions $\boldsymbol{a}$ ensuring consistency and asymptotic normality of the sequence of maximum likelihood estimators $\hat{\boldsymbol{\vartheta}}_N$ and an estimator of their asymptotic covariance matrix are given in the following three theorems. Because of the full parameterization of the model, the proofs are simple standard results for likelihood asymptotics, and a more complex analysis (as required for estimation equations in the presence of a nuisance parameter h_Y, e.g. in Cheng *et al.*, 1995) is not necessary. We will restrict ourselves to absolutely continuous or discrete random variables Y, where the likelihood

is given in terms of the density $f_Y(y \mid \boldsymbol{\vartheta})$. Furthermore, we will only study the case of a correctly specified transformation $h = \boldsymbol{a}^\top \boldsymbol{\vartheta}$ and refer the reader to Hothorn *et al.* (2014), where consistency results for arbitrary h are given.

Theorem 1. *For* $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} \mathbb{P}_{Y,\boldsymbol{\vartheta}_0}$ *and under the assumptions (A1), the parameter space* Θ *is compact and (A2)* $\mathbb{E}_{\boldsymbol{\vartheta}_0}[\sup_{\boldsymbol{\vartheta}\in\Theta} |\log(f_Y(Y \mid \boldsymbol{\vartheta}))|] < \infty$ *where* $\boldsymbol{\vartheta}_0$ *is well separated:*

$$\sup_{\boldsymbol{\vartheta};|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0|\geq\epsilon} \mathbb{E}_{\boldsymbol{\vartheta}_0}[\log(f_Y(Y \mid \boldsymbol{\vartheta}))] < \mathbb{E}_{\boldsymbol{\vartheta}_0}[\log(f_Y(Y \mid \boldsymbol{\vartheta}_0))],$$

the sequence of estimators $\hat{\boldsymbol{\vartheta}}_N$ *converges to* $\boldsymbol{\vartheta}_0$ *in probability,* $\hat{\boldsymbol{\vartheta}}_N \xrightarrow{\mathbb{P}} \boldsymbol{\vartheta}_0$, *as* $N \to \infty$.

Proof. The log-likelihood is continuous in $\boldsymbol{\vartheta}$, and because of (A2), each log-likelihood contribution is dominated by an integrable function. Thus, the result follows from van der Vaart (1998) (Theorem 5.8 with example 19.7; see note at bottom of page 46). □

Remark 1. Assumption (A1) is made for convenience, and relaxations of such a condition are given in van de Geer (2000) or van der Vaart (1998). The assumptions in (A2) are rather weak: the first one holds if the functions $\boldsymbol{a}$ are not arbitrarily ill posed, and the second one holds if the function $\mathbb{E}_{\boldsymbol{\vartheta}_0}[\log(f_Y(Y \mid \boldsymbol{\vartheta}))]$ is strictly convex in $\boldsymbol{\vartheta}$ (if the assumption would not hold, we would still have convergence to the set $\arg\max_{\boldsymbol{\vartheta}} \mathbb{E}_{\boldsymbol{\vartheta}_0}[\log(f_Y(Y \mid \boldsymbol{\vartheta}))]$).

Theorem 2. *Under the assumptions of Theorem 1 and in addition (A3)*

$$\mathbb{E}_{\boldsymbol{\vartheta}_0}\left(\sup_{\boldsymbol{\vartheta}} \left\| \frac{\partial \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \right\|^2\right) < \infty,$$

(A4) $\mathbb{E}_{\boldsymbol{\vartheta}_0}(\boldsymbol{a}(Y)\boldsymbol{a}(Y)^\top)$ *and (for the absolutely continuous case* $\mu = \mu_L$ *only)* $\mathbb{E}_{\boldsymbol{\vartheta}_0}(\boldsymbol{a}'(Y)\boldsymbol{a}'(Y)^\top)$ *are non-singular, and (A5)* $0 < f_Z < \infty$, $\sup|f_Z'| < \infty$ *and* $\sup|f_Z''| < \infty$, *the sequence* $\sqrt{N}(\hat{\boldsymbol{\vartheta}}_N - \boldsymbol{\vartheta}_0)$ *is asymptotically normal with mean zero and covariance matrix*

$$\Sigma_{\boldsymbol{\vartheta}_0} = \left(\mathbb{E}_{\boldsymbol{\vartheta}_0}\left(-\frac{\partial^2 \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}\, \partial \boldsymbol{\vartheta}^\top}\right)\right)^{-1},$$

as $N \to \infty$.

Proof. Because the map $\boldsymbol{\vartheta} \mapsto \sqrt{f_Y(y \mid \boldsymbol{\vartheta})}$ is continuously differentiable in $\boldsymbol{\vartheta}$ for all y in both the discrete and absolutely continuous case and the matrix

$$\mathbb{E}_{\boldsymbol{\vartheta}_0}\left(\left[\frac{\partial \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right]\left[\frac{\partial \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right]^\top\right)$$

is continuous in $\boldsymbol{\vartheta}$ as given in (1) and (2), the transformation family $\mathbb{P}_{Y,\Theta}$ is differentiable in quadratic mean with Lemma 7.6 in van der Vaart (1998). Furthermore, assumptions (A4 and A5) ensure that the expected Fisher information matrix is non-singular at $\boldsymbol{\vartheta}_0$. With the consistency and (A3), the result follows from Theorem 5.39 in van der Vaart (1998). □

Remark 2. Assumption (A4) is valid for the densities f_Z of the normal, logistic and MEV distribution. The Fisher information (3) and (4) evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}_N$ can be used to estimate the covariance matrix $\Sigma_{\boldsymbol{\vartheta}_0}$.

Theorem 3. *Under the assumptions of Theorem 2 and assuming* $\mathbb{E}_{\vartheta_0}|\boldsymbol{F}(\vartheta_0 \mid Y)| < \infty$, *a consistent estimator for* Σ_{ϑ_0} *is given by*

$$\hat{\Sigma}_{\vartheta_0,N} = \left(N^{-1}\sum_{i=1}^{N} \boldsymbol{F}(\hat{\vartheta}_N \mid Y_i)\right)^{-1}.$$

Proof. With the law of large numbers, we have

$$N^{-1}\sum_{i=1}^{N} \boldsymbol{F}(\vartheta_0 \mid Y_i) = N^{-1}\sum_{i=1}^{N} -\frac{\partial^2 \log f_Y(Y_i \mid \vartheta)}{\partial\vartheta\,\partial\vartheta^\top} \xrightarrow{\mathbb{P}} \mathbb{E}_{\vartheta_0}\left(-\frac{\partial^2 \log f_Y(Y \mid \vartheta)}{\partial\vartheta\,\partial\vartheta^\top}\right) = \Sigma_{\vartheta_0}^{-1}.$$

Because the map $\vartheta \mapsto \boldsymbol{F}(\vartheta \mid y)$ is continuous for all y (as can be seen from (3) and (4)), the result follows with Theorem 1. □

Based on Theorems 1–3, we can perform standard likelihood inference on the model parameters ϑ. In particular, we can construct confidence intervals and confidence bands for the conditional distribution function from confidence intervals and bands for the linear functions $\boldsymbol{a}^\top\vartheta$. We complete this part by formally defining the class of transformation models.

Definition 4 (Transformation model)**.** *The triple* $(F_Z, \boldsymbol{a}, \vartheta)$ *is called transformation model.*

The transformation model $(F_Z, \boldsymbol{a}, \vartheta)$ fully defines the distribution of Y via $F_Y = F_Z \circ \boldsymbol{a}^\top\vartheta$ and thus the corresponding likelihood $\mathcal{L}(\boldsymbol{a}^\top\vartheta \mid Y \in (\underline{y}, \bar{y}])$. Our definition of transformation models as $(F_Z, \boldsymbol{a}, \vartheta)$ is strongly tied to the idea of structural inference (Fraser, 1968) and group models (Bickel *et al.*, 1993). Fraser (1968) described a measurement model $\mathbb{P}_Y$ for Y by an error distribution $\mathbb{P}_Z$ and a structural equation $Y = g \circ Z$, where g is a linear function, thereby extending the location-scale family $Y = \alpha + \sigma Z$. Group models consist of distributions generated by possibly non-linear g. The main difference to these classical approaches is that we parameterize h instead of $g = h^{-1}$. By extending the linear transformation functions g dealt with by Fraser (1968) to non-linear transformations, we approximate the potentially non-linear transformation functions $h = g^{-1} = F_Z^{-1} \circ F_Y$ by $\boldsymbol{a}^\top\vartheta$, with subsequent estimation of the parameters ϑ. For given parameters ϑ, a sample from $\mathbb{P}_Y$ can be drawn by the probability integral transform, that is, $Z_1, \ldots, Z_N \overset{\text{iid}}{\sim} \mathbb{P}_Z$ is drawn and then $Y_i = \inf\{y \in \Xi \mid \boldsymbol{a}(y)^\top \vartheta \geq Z_i\}$.

3.2. Non-parametric inference

For continuous responses Y, any unknown transformation h can be approximated by Bernstein polynomials of increasing order (Farouki, 2012). For uncensored and right-censored responses and under the same conditions for F_Z as stated in Section 3.1, McLain & Ghosh (2013) showed that the non-parametric sieve maximum likelihood estimator is consistent with rate of convergence $N^{2/5}$ for h with continuous bounded second derivatives in unconditional and linear transformation models (Section 4.3). In the latter class, the linear shift parameters $\boldsymbol{\beta}$ are asymptotically normal and semi-parametrically efficient. Numerical approximations to the observed Fisher information $\boldsymbol{F}(\hat{\vartheta}_N \mid Y \in (\underline{y}, \bar{y}])$ were shown to lead to appropriate standard errors of $\hat{\boldsymbol{\beta}}_N$ by McLain & Ghosh (2013). Hothorn *et al.* (2014) established the consistency of boosted non-parametric conditional transformation models (Section 4.2). For sieve maximum likelihood estimation in the class of conditional transformation models, the techniques employed by McLain & Ghosh (2013) require minor technical extensions, which are omitted here.

In summary, the same limiting distribution arises under both the parametric and the non-parametric paradigm for transformation functions parameterized or approximated using Bernstein polynomials, respectively. In the latter case, the target is then the best approximated transformation function with Bernstein polynomials, say $h_N^\star$ (where the index N indicates that we use a more complex approximation when N increases). If the approximation error $h_N^\star - h$ is of smaller order than the convergence rate of the estimator, the estimator's target becomes the true underlying transformation function h, and otherwise, a bias for estimating h remains.

4. Applications

The definition of transformation models tailored for specific situations 'only' requires the definition of a suitable basis function $\boldsymbol{a}$ and a choice of F_Z. In this section, we will discuss specific transformation models for unconditional and conditional distributions of ordered categorical, discrete and continuous responses Y. Note that the likelihood function $\mathcal{L}$ allows all these models to be fitted to arbitrarily censored or truncated responses; for brevity, we will not elaborate on the details.

4.1. Unconditional transformation models

Finite sample space For ordered categorical responses Y from a finite sample space $\Xi = \{y_1, \ldots, y_K\}$, we assign one parameter to each element of the sample space except y_K. This corresponds to the basis function $\boldsymbol{a}(y_k) = \boldsymbol{e}_{K-1}(k)$, where $\boldsymbol{e}_{K-1}(k)$ is the unit vector of length $K-1$, with its kth element being one. The transformation function h is

$$h(y_k) = \boldsymbol{e}_{K-1}(k)^\top \boldsymbol{\vartheta} = \vartheta_k \in \mathbb{R}, \quad 1 \le k < K, \quad \text{st} \quad \vartheta_1 < \cdots < \vartheta_{K-1},$$

with $h(y_K) = \infty$, and the unconditional distribution function of F_Y is $F_Y(y_k) = F_Z(\vartheta_k)$. This parameterization underlies the common proportional odds and proportional hazards model for ordered categorical data (Tutz, 2012). Note that monotonicity of h is guaranteed by the $K-2$ linear constraints $\vartheta_2 - \vartheta_1 > 0, \ldots, \vartheta_{K-1} - \vartheta_{K-2} > 0$ when constrained optimization is performed. In the absence of censoring or truncation and with $\vartheta_0 = -\infty, \vartheta_K = \infty$, we obtain the maximum likelihood estimator for $\boldsymbol{\vartheta}$ as

$$\begin{aligned}\hat{\boldsymbol{\vartheta}}_N &= \operatorname*{arg\,max}_{\vartheta_1 < \cdots < \vartheta_{K-1}} \sum_{i=1}^N \log(F_Z(\vartheta_{k(i)}) - F_Z(\vartheta_{k(i)-1})) \\ &= \left(F_Z^{-1}\left(N^{-1} \sum_{i=1}^N \mathbb{1}(Y_i \le y_1) \right), \ldots, F_Z^{-1}\left(N^{-1} \sum_{i=1}^N \mathbb{1}(Y_i \le y_{K-1}) \right) \right)^\top\end{aligned}$$

because $\hat{\pi}_k = N^{-1} \sum_{i=1}^N \mathbb{1}(Y_i = y_k), 1 \le k < K$ maximizes the equivalent multinomial (or empirical) log-likelihood $\sum_{i=1}^N \log(\pi_{k(i)})$, and we can rewrite this estimator as

$$\hat{\pi}_k = N^{-1} \left(\sum_{i=1}^N \mathbb{1}(Y_i \le y_k) - \mathbb{1}(k > 1) \sum_{i=1}^N \mathbb{1}(Y_i \le y_{k-1}) \right), 1 \le k < K.$$

The estimated distribution function $\hat{F}_{Y,N} = F_Z \circ \hat{h}_N$ is invariant with respect to F_Z.

Assumption (A4) is valid for these basis functions because we have $\mathbb{E}_{\boldsymbol{\vartheta}_0}(\boldsymbol{e}_{K-1}(Y)\boldsymbol{e}_{K-1}(Y)^\top) = \operatorname{diag}(P(Y = y_k)), 1 \le k < K$ for $Y \sim P_{Y,\boldsymbol{\vartheta}_0}$.

If we define the sample space Ξ as the set of unique observed values and the probability measure as the empirical cumulative distribution function (ECDF), putting mass N^{-1} on each

observation, we see that this particular parameterization is equivalent to an empirical likelihood approach, and we get $\hat{h}_N = F_Z^{-1} \circ \text{ECDF}$. Note that although the transformation function depends on the choice of F_Z, the estimated distribution function $\hat{F}_{Y,N} = F_Z \circ \hat{h}_N = \text{ECDF}$ does not and is simply the non-parametric empirical maximum likelihood estimator. A smoothed version of this estimator for continuous responses is discussed in the next paragraph.

Infinite sample space For continuous responses Y, the parameterization $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$, and thus also $\hat{F}_{Y,N}$, should be smooth in y; therefore, any polynomial or spline basis is a suitable choice for $\boldsymbol{a}$. For the empirical experiments in Section 5, we applied Bernstein polynomials (for an overview, see Farouki, 2012) of order M ($P = M + 1$) defined on the interval $[\underline{\iota}, \bar{\iota}]$ with

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

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. This choice is computationally attractive because strict monotonicity can be formulated as a set of M linear constraints on the parameters $\vartheta_m < \vartheta_{m+1}$ for all $m = 0, \ldots, M$ (Curtis & Ghosh, 2011). Therefore, application of constrained optimization guarantees monotonic estimates $\hat{h}_N$. The basis contains an intercept. We obtain smooth plug-in estimators for the distribution, density, hazard and cumulative hazard functions as $\hat{F}_{Y,N} = F_Z \circ \boldsymbol{a}_{\text{Bs},M}^\top \hat{\boldsymbol{\vartheta}}_N$, $\hat{f}_{Y,N} = f_Z \circ \boldsymbol{a}_{\text{Bs},M}^\top \hat{\boldsymbol{\vartheta}}_N \times \boldsymbol{a}'_{\text{Bs},M}{}^\top \hat{\boldsymbol{\vartheta}}_N$, $\hat{\lambda}_{Y,N} = \hat{f}_{Y,N}/(1 - \hat{F}_{Y,N})$ and $\hat{\Lambda}_{Y,N} = -\log(1 - \hat{F}_{Y,N})$. The estimator $\hat{F}_{Y,N} = F_Z \circ \boldsymbol{a}_{\text{Bs},M}^\top \hat{\boldsymbol{\vartheta}}_N$ must not be confused with the estimator $\hat{F}_{Y,N} = \boldsymbol{a}_{\text{Bs},M}^\top \hat{\boldsymbol{p}}$ for $Y \in [0, 1]$ obtained from the smoothed empirical distribution function with coefficients $\hat{\boldsymbol{p}}_{m+1} = \sum_{i=1}^{N} \mathbb{1}(Y_i \leq m/M)/N$ corresponding to probabilities evaluated at the quantiles m/M for $m = 0, \ldots, M$ (Babu *et al.*, 2002).

The question arises how the degree of the polynomial affects the estimated distribution function. On the one hand, the model $(\Phi, \boldsymbol{a}_{\text{Bs},1}, \boldsymbol{\vartheta})$ only allows linear transformation functions of a standard normal, and F_Y is restricted to the normal family. On the other hand, $(\Phi, \boldsymbol{a}_{\text{Bs},N-1}, \boldsymbol{\vartheta})$ has one parameter for each observation, and $\hat{F}_{Y,N}$ is the non-parametric maximum likelihood estimator ECDF, which, by the Glivenko–Cantelli lemma, converges to F_Y. In this sense, we cannot choose a 'too large' value for M. This is a consequence of the monotonicity constraint on the estimator $\boldsymbol{a}^\top \hat{\boldsymbol{\vartheta}}_N$, which, in this extreme case, just interpolates the step function $F_Z^{-1} \circ \text{ECDF}$. Empirical evidence for the insensitivity of results when M is large can be found in Hothorn (2017b) and in the discussion.

4.2. Conditional transformation models

In the following, we will discuss a cascade of increasingly complex transformation models where the transformation function h may depend on explanatory variables $\boldsymbol{X} \in \chi$. We are interested in estimating the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$. The corresponding distribution function $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ can be written as $F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h(y \mid \boldsymbol{x}))$. The transformation function $h(\cdot \mid \boldsymbol{x}) : \Xi \to \mathbb{R}$ is said to be conditional on $\boldsymbol{x}$. Following the arguments presented in the proof of Corollary 1, it is easy to see that for each $\boldsymbol{x}$, there exists a strictly monotonic transformation function $h(\cdot \mid \boldsymbol{x}) = F_Z^{-1} \circ F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}$ such that $F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(y) =$

$F_Z(h(y \mid \boldsymbol{x}))$. Because this class of conditional transformation models and suitable parameterizations was introduced by Hothorn *et al.* (2014), we will only sketch the most important aspects here.

Let $\boldsymbol{b} : \chi \to \mathbb{R}^Q$ denotes a basis transformation of the explanatory variables. The joint basis for both y and $\boldsymbol{x}$ is called $\boldsymbol{c} : \Xi \times \chi \to \mathbb{R}^{d(P,Q)}$; its dimension $d(P, Q)$ depends on the way the two basis functions $\boldsymbol{a}$ and $\boldsymbol{b}$ are combined (e.g. $\boldsymbol{c} = (\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top \in \mathbb{R}^{P+Q}$ or $\boldsymbol{c} = (\boldsymbol{a}^\top \otimes \boldsymbol{b}^\top)^\top \in \mathbb{R}^{PQ}$). The conditional transformation function is now parameterized as $h(y \mid \boldsymbol{x}) = \boldsymbol{c}(y, \boldsymbol{x})^\top \boldsymbol{\vartheta}$. One important special case is the simple transformation function $h(y \mid \boldsymbol{x}) = h_Y(y) + h_{\boldsymbol{x}}(\boldsymbol{x})$, where the explanatory variables only contribute a shift $h_{\boldsymbol{x}}(\boldsymbol{x})$ to the conditional transformation function. Often this shift is assumed to be linear in $\boldsymbol{x}$; therefore, we use the function $m(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\beta} = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$ to denote linear shifts. Here, $\boldsymbol{b}(\boldsymbol{x}) = \tilde{\boldsymbol{x}}$ is one row of the design matrix without intercept. These simple models correspond to the joint basis $\boldsymbol{c}(y, \boldsymbol{x})^\top \boldsymbol{\vartheta} = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 + \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_2$, with $h_Y(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1$ and $h_{\boldsymbol{x}}(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_2 = m(\boldsymbol{x}) = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$. The results presented in Section 3, including Theorems 1–3, carry over in the fixed design case when $\boldsymbol{a}$ is replaced by $\boldsymbol{c}$.

In the rest of this section, we will present classical models that can be embedded in the larger class of conditional transformation models and some novel models that can be implemented in this general framework.

4.3. Classical transformation models

Linear model The normal linear regression model $Y \sim \mathrm{N}(\alpha + m(\boldsymbol{x}), \sigma^2)$ with conditional distribution function $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = \Phi(\sigma^{-1}(y - (\alpha + m(\boldsymbol{x}))))$ can be understood as a transformation model with transformation function $h(y \mid \boldsymbol{x}) = y/\sigma - \alpha/\sigma - m(\boldsymbol{x})/\sigma$ parameterized via basis functions $\boldsymbol{a}(y) = (y, 1)^\top, \boldsymbol{b}(\boldsymbol{x}) = \tilde{\boldsymbol{x}}$ and $\boldsymbol{c} = (\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top$ with parameters $\boldsymbol{\vartheta} = (\sigma^{-1}, -\sigma^{-1}\alpha, -\sigma^{-1}\boldsymbol{\beta}^\top)^\top$ under the constraint $\sigma > 0$ or in more compact notation $(\Phi, (y, 1, \tilde{\boldsymbol{x}}^\top)^\top, \boldsymbol{\vartheta})$. The parameters of the model are the inverse standard deviation and the inverse negative coefficient of variation instead of the mean and variance of the original normal distribution. For 'exact continuous' observations, the likelihood $\mathcal{L}$ is equivalent to least squares, which can be maximized with respect to α and $\boldsymbol{\beta}$ without taking σ into account. This is not possible for censored or truncated observations, where we need to evaluate the conditional distribution function that depends on all parameters; this model is called Type I Tobit model (although only the likelihood changes under censoring and truncation, but the model does not). Using an alternative basis function $\boldsymbol{c}$ would allow arbitrary non-normal conditional distributions of Y, and the simple shift model $\boldsymbol{c}(y, \boldsymbol{x})^\top \boldsymbol{\vartheta} = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 + \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_2$ is then a generalization of additive models and leads to the interpretation $\mathbb{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{a}(Y)^\top \boldsymbol{\vartheta}_1) = -\boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_2$. The choice $\boldsymbol{a} = (1, \log)^\top$ implements the log-normal model for $Y > 0$. Implementation of a Bernstein basis $\boldsymbol{a} = \boldsymbol{a}_{\mathrm{Bs},M}$ allows arbitrarily shaped distributions, that is, a transition from the normal family to the transformation family, and thus likelihood inference on $\boldsymbol{\vartheta}_2$ without strict assumptions on the distribution of Y. The transformation $\boldsymbol{a}_{\mathrm{Bs},M}(y)^\top \boldsymbol{\vartheta}_1$ must increase monotonically in y. Maximization of the log-likelihood under the linear inequality constraint $\boldsymbol{D}_{M+1}\boldsymbol{\vartheta}_1 > 0$, with $\boldsymbol{D}_{M+1}$ representing first-order differences, implements this requirement.

Continuous 'survival time' models For a continuous response $Y > 0$, the model $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(\sigma^{-1}(\log(y) - (\alpha + m(\boldsymbol{x}))))$ with basis functions $\boldsymbol{a}(y) = (1, \log(y))$ and $\boldsymbol{b}(\boldsymbol{x}) = \tilde{\boldsymbol{x}}$ and parameters $\boldsymbol{\vartheta} = (-\alpha, \sigma^{-1}, -\boldsymbol{\beta}^\top)^\top$ under the constraint $\sigma > 0$ is called the accelerated failure time (AFT) model. The model $(F_{\mathrm{MEV}}, (1, \log, \tilde{\boldsymbol{x}}^\top)^\top, (-\vartheta_1, 1, -\boldsymbol{\beta}^\top)^\top)$ with $\sigma \equiv 1$ (and thus fixed transformation function log) is the exponential AFT model because it

implies an exponential distribution of Y. When the parameter $\sigma > 0$ is estimated from the data, the model $(F_{\text{MEV}}, (1, \log, \tilde{\boldsymbol{x}}^\top)^\top, \boldsymbol{\vartheta})$ is called the Weibull model, $(F_{\text{SL}}, (1, \log, \tilde{\boldsymbol{x}}^\top)^\top, \boldsymbol{\vartheta})$ is the log-logistic AFT model and $(\Phi, (1, \log, \tilde{\boldsymbol{x}}^\top)^\top, \boldsymbol{\vartheta})$ is the log-normal AFT model. For a continuous (not necessarily positive) response Y, the model $F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(y) = F_{\text{MEV}}(h_Y(y) - m(\boldsymbol{x}))$ is called the proportional hazards, relative risk or Cox model. The transformation function h_Y equals the log-cumulative baseline hazard and is treated as a nuisance parameter in the partial likelihood framework, where only the regression coefficients $\boldsymbol{\beta}$ are estimated. Given $\hat{\boldsymbol{\beta}}$, non-parametric maximum likelihood estimators are typically applied to obtain $\hat{h}_Y$. Here, we parameterize this function as $h_Y(y) = \log(\Lambda_Y(y)) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1$ (e.g. using $\boldsymbol{a} = \boldsymbol{a}_{\text{Bs},M}$) and fit all parameters in the model $(F_{\text{MEV}}, (\boldsymbol{a}^\top, \tilde{\boldsymbol{x}}^\top)^\top, (\boldsymbol{\vartheta}_1^\top, -\boldsymbol{\beta}^\top)^\top)$ simultaneously. The model is highly popular because $m(\boldsymbol{x})$ is the log-hazard ratio to $m(\boldsymbol{0})$. For the special case of right-censored survival times, this parameterization of the Cox model was studied theoretically and empirically by McLain & Ghosh (2013). Changing the distribution function in the Cox model from F_{MEV} to F_{SL} results in the proportional odds model $(F_{\text{SL}}, (\boldsymbol{a}^\top, \tilde{\boldsymbol{x}}^\top)^\top, (\boldsymbol{\vartheta}_1^\top, -\boldsymbol{\beta}^\top)^\top)$; its name comes from the interpretation of $m(\boldsymbol{x})$ as the constant log-odds ratio of the odds $O_Y(y \mid \boldsymbol{X} = \boldsymbol{x})$ and $O_Y(y \mid \boldsymbol{x} = \boldsymbol{0})$. An additive hazards model with the conditional hazard function $\lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{x}) = \lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0}) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$ results from the choice $F_Z(z) = F_{\text{Exp}}(z) = 1 - \exp(-z)$ (Aranda-Ordaz, 1983) under the additional constraint $\lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{x}) > 0$. In this case, the function $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 > 0$ is the positive baseline cumulative hazard function $\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$.

Discrete models For ordered categorical responses $y_1 < \cdots < y_K$, the conditional distribution $F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(y_k) = F_Z(h_Y(y_k) - m(\boldsymbol{x}))$ is a transformation model with $\boldsymbol{a}(y_k) = \boldsymbol{e}_{K-1}(k)$. The model $(F_{\text{SL}}, (\boldsymbol{a}^\top, \tilde{\boldsymbol{x}}^\top)^\top, (\boldsymbol{\vartheta}_1^\top, -\boldsymbol{\beta}^\top)^\top)$ is called the discrete proportional odds model, and $(F_{\text{MEV}}, (\boldsymbol{a}^\top, \tilde{\boldsymbol{x}}^\top)^\top, (\boldsymbol{\vartheta}_1^\top, -\boldsymbol{\beta}^\top)^\top)$ is the discrete proportional hazards model. Here, $m(\boldsymbol{x})$ is the log-odds ratio or log-hazard ratio to $m(\boldsymbol{0})$ independent of k; details are given in Tutz (2012). For the special case of a binary response ($K = 2$), the transformation model $(F_{\text{SL}}, (\mathbb{1}(k = 1), \tilde{\boldsymbol{x}}^\top)^\top, (\vartheta_1, -\boldsymbol{\beta}^\top)^\top)$ is the logistic regression model, $(\Phi, (\mathbb{1}(k = 1), \tilde{\boldsymbol{x}}^\top)^\top, (\vartheta_1, -\boldsymbol{\beta}^\top)^\top)$ is the probit model and $(F_{\text{MEV}}, (\mathbb{1}(k = 1), \tilde{\boldsymbol{x}}^\top)^\top, (\vartheta_1, -\boldsymbol{\beta}^\top)^\top)$ is called the complementary log–log model. Note that the transformation function h_Y is given by the basis function $\boldsymbol{a} = \mathbb{1}(k = 1)$, that is, ϑ_1 is just the intercept. The connection between standard binary regression models and transformation models is explained in more detail by Doksum & Gasko (1990).

Linear transformation model The transformation model $(F_Z, (\boldsymbol{a}^\top, \tilde{\boldsymbol{x}}^\top)^\top, (\boldsymbol{\vartheta}_1^\top, -\boldsymbol{\beta}^\top)^\top)$ for any $\boldsymbol{a}$ and F_Z is called the linear transformation model and contains all models discussed in this section. Note that the transformation of the response $h_Y(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1$ is non-linear in all models of interest (AFT, Cox etc.) and the term 'linear' only refers to a linear shift $m(\boldsymbol{x})$ of the explanatory variables. Partially linear or additive transformation models allow non-linear shifts as part of a partially smooth basis $\boldsymbol{b}$, that is, in the form of an additive model. The number of constraints only depends on the basis $\boldsymbol{a}$ but not on the explanatory variables.

4.4. Extension of classical transformation models

A common property of all classical transformation models is the additivity of the response transformation and the shift, that is, the decomposition $h(y \mid \boldsymbol{x}) = h_Y(y) + h_{\boldsymbol{x}}(\boldsymbol{x})$ of the conditional transformation function. This assumption is relaxed by the following extensions of the classical models. Allowing for deviations from this simple model is also the key aspect for the development of novel transformation models in the rest of this section.

Discrete non-proportional odds and hazards models For ordered categorical responses, the model $F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(y_k) = F_Z(h_Y(y_k) - m_k(\boldsymbol{x}))$ allows a category-specific shift $m_k(\boldsymbol{x}) = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_k$; with F_{SL}, this cumulative model is called the non-proportional odds model, and with F_{MEV}, it is the non-proportional hazards model. Both models can be cast into the transformation model framework by defining the joint basis $\boldsymbol{c}(y_k, \boldsymbol{x}) = (\boldsymbol{a}(y_k)^\top, \boldsymbol{a}(y_k)^\top \otimes \boldsymbol{b}(\boldsymbol{x})^\top)^\top$ as the Kronecker product of the two simple basis functions $\boldsymbol{a}(y_k) = \boldsymbol{e}_{K-1}(k)$ and $\boldsymbol{b}(\boldsymbol{x}) = \tilde{\boldsymbol{x}}$ (assuming that $\boldsymbol{b}$ does not contain an intercept term). Note that the conditional transformation function $h(y \mid \boldsymbol{x})$ includes an interaction term between y and $\boldsymbol{x}$.

Time-varying effects One often studied extension of the Cox model is $F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h_Y(y) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}(y))$, where the regression coefficients $\boldsymbol{\beta}(y)$ may change with time y. The Cox model is included with $\boldsymbol{\beta}(y) \equiv \boldsymbol{\beta}$, and the model is often applied to check the proportional hazards assumption. With a smooth parameterization of time y, for example, via $\boldsymbol{a} = \boldsymbol{a}_{\text{Bs},M}$, and linear basis $\boldsymbol{b}(\boldsymbol{x}) = \tilde{\boldsymbol{x}}$, the transformation model $(F_{\text{MEV}}, (\boldsymbol{a}^\top, \boldsymbol{a}^\top \otimes \boldsymbol{b}^\top)^\top, \boldsymbol{\vartheta})$ implements this Cox model with time-varying (linear) effects. This model (with arbitrary F_Z) has also been presented in Foresi & Peracchi (1995) and is called distribution regression in Chernozhukov *et al.* (2013).

4.5. Novel transformation models

Because of the broadness of the transformation family, it is straightforward to set up new models for interesting situations by allowing more complex transformation functions $h(y \mid \boldsymbol{x})$. We will illustrate this possibility for two simple cases the independent two-sample situation and regression models for count data. The generic and most complex transformation model is called the conditional transformation model and is explained at the end of this section.

Beyond shift effects Assume we observe samples from two groups A and B and want to model the conditional distribution functions $F_{Y \mid \boldsymbol{X}=A}(y)$ and $F_{Y \mid \boldsymbol{X}=B}(y)$ of the response Y in the two groups. Based on this model, it is often interesting to infer whether the two distributions are equivalent and, if this is not the case, to characterize how they differ. Using an appropriate basis function $\boldsymbol{a}$ and the basis $\boldsymbol{b}(\boldsymbol{x}) = (1, \mathbb{1}(B))^\top$, the model $(F_Z, (\boldsymbol{a}^\top \otimes \boldsymbol{b}^\top)^\top, \boldsymbol{\vartheta})$ parameterizes the conditional transformation function as $h(y \mid A) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1$ and $h(y \mid B) = h(y \mid A) + h_{B-A}(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 + \mathbb{1}(B)\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_2$. Clearly, the second term is constant zero ($h_{B-A}(y) \equiv 0$) iff the two distributions are equivalent ($F_{Y|\boldsymbol{X}=A}(y) = F_{Y|\boldsymbol{X}=B}(y)$ for all y). For the deviation function $h_{B-A}(y) = \boldsymbol{a}^\top \boldsymbol{\vartheta}_2$, we can apply standard likelihood inference procedures for $\hat{\boldsymbol{\vartheta}}_2$ to construct a confidence band or use a test statistic like $\max(\hat{\boldsymbol{\vartheta}}_2 / \text{se}(\hat{\boldsymbol{\vartheta}}_2))$ to assess deviations from zero. If there is evidence for a group effect, we can use the model to check whether the deviation function is constant, that is, $h_{B-A}(y) \equiv c \neq 0$. In this case, the simpler model $(F_Z, (\boldsymbol{a}^\top, \mathbb{1}(B))^\top, (\boldsymbol{\vartheta}_1^\top, -\beta)^\top)$ with shift $\beta = -\boldsymbol{\vartheta}_2$ might be easier to interpret. This model actually corresponds to a normal analysis of variance model with $F_Z = \Phi$ and $\boldsymbol{a}(y)^\top = (1, y)^\top$ or the Cox proportional hazards model with $(F_{\text{MEV}}, (\boldsymbol{a}_{\text{Bs},M}^\top, \mathbb{1}(B))^\top, (\boldsymbol{\vartheta}_1^\top, -\beta)^\top)$.

Count regression 'without tears' Simple models for count data $\Xi = \{0, 1, 2, \ldots\}$ almost always suffer from over-dispersion or excess zeros. The linear transformation model $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h_Y(y) - m(\boldsymbol{x}))$ can be implemented using the basis function $\boldsymbol{a}(y) = \boldsymbol{a}_{\text{Bs},M}(\lfloor y \rfloor)$, and then the parameters of the transformation model $(F_Z, (\boldsymbol{a}^\top, \tilde{\boldsymbol{x}}^\top)^\top, \boldsymbol{\vartheta})$ are not affected by over-dispersion or under-dispersion because higher moments are handled by h_Y independently of the effects of the explanatory variables $m(\boldsymbol{x})$. If there are excess zeros, we can set up a joint transformation model $F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h_Y(y) - m(\boldsymbol{x}) + \mathbb{1}(y =$

$0)(\alpha_0 - m_0(\boldsymbol{x})))$ such that we have a two-components mixture model consisting of the count distribution $F_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h_Y(y) - m(\boldsymbol{x}))$ for $y \in \Xi$ and the probability of an excess zero

$$f_{Y \mid \boldsymbol{X}=\boldsymbol{x}}(0) = F_Z(h_Y(0) - m(\boldsymbol{x}) + (\alpha_0 - m_0(\boldsymbol{x}))) = F_Z(h_Y(0) + \alpha_0 - \tilde{\boldsymbol{x}}^\top(\boldsymbol{\beta} + \boldsymbol{\beta}_0))$$

when $m_0(\boldsymbol{x}) = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_0$. Hence, the transformation analogue to a hurdle model with hurdle at zero is the transformation model $(F_Z, (\boldsymbol{a}^\top, \tilde{\boldsymbol{x}}^\top, \mathbb{1}(y = 0), \mathbb{1}(y = 0)\tilde{\boldsymbol{x}}^\top)^\top, (\boldsymbol{\vartheta}_1^\top, \boldsymbol{\beta}^\top, \alpha_0, \boldsymbol{\beta}_0^\top)^\top)$.

Conditional transformation models When the conditional transformation function is parameterized by multiple basis functions $\boldsymbol{a}_j(y), \boldsymbol{b}_j(\boldsymbol{x}), j = 1, \ldots J$ via the joint basis

$$\boldsymbol{c} = (\boldsymbol{a}_1^\top \otimes \boldsymbol{b}_1^\top, \ldots, \boldsymbol{a}_J^\top \otimes \boldsymbol{b}_J^\top)^\top,$$

models of the class $(\cdot, \boldsymbol{c}, \boldsymbol{\vartheta})$ are called conditional transformation models with J partial transformation functions parameterized as $\boldsymbol{a}_j^\top \otimes \boldsymbol{b}_j^\top$ and include all special cases discussed in this section. It is convenient to assume monotonicity for each of the partial transformation functions; thus, the linear constraints for $\boldsymbol{a}_j$ are repeated for each basis function in $\boldsymbol{b}_j$ (detailed descriptions of linear constraints for different models in this class are available in Hothorn, 2017b). Hothorn *et al.* (2014) introduced this general model class and proposed a boosting algorithm for the estimation of transformation functions h for 'exact continuous' responses Y. In the likelihood framework presented here, conditional transformation models can be fitted under arbitrary schemes of censoring and truncation, and classical likelihood inference for the model parameters $\boldsymbol{\vartheta}$ becomes feasible. Of course, unlike in the boosting context, the number of model terms J and their complexity are limited in the likelihood world because the likelihood does not contain any penalty terms that induce smoothness in the $\boldsymbol{x}$-direction.

A systematic overview of linear transformation models with potentially response-varying effects is given in Table 1. Model nomenclature and interpretation of the corresponding model parameters is mapped to specific transformation functions h and distribution functions F_Z. To the best of our knowledge, models without names have not yet been discussed in the literature, and their specific properties await closer investigation.

5. Empirical evaluation

We will illustrate the range of possible applications of likelihood-based conditional transformation models. In Section 5.2, we will present a small simulation experiment highlighting the possible advantage of indirectly modelling conditional distributions with transformation functions.

5.1. Illustrations

Density estimation: Old Faithful geyser The duration of eruptions and the waiting time between eruptions of the Old Faithful geyser in the Yellowstone National Park became a standard benchmark for non-parametric density estimation. The nine parameters of the transformation model $(\Phi, \boldsymbol{a}_{\mathrm{Bs},8}(\text{waiting}), \boldsymbol{\vartheta})$ were fitted by maximization of the approximate log-likelihood (treating the waiting times as 'exact' observations) under the eight linear constraints $\boldsymbol{D}_9\boldsymbol{\vartheta} > 0$. The model depicted in Fig. 1A reproduces the classic bimodal unconditional density of waiting time along with a kernel density estimate. It is important to note that the transformation model was fitted likelihood based, whereas the kernel density estimate relied on a cross-validated bandwidth. An unconditional density estimate for the duration of the eruptions needs to deal with censoring because exact duration times are only available for the

Table 1. *Non-exhaustive overview of conditional transformation models. Abbreviations: proportional hazards (PH), proportional odds (PO), additive hazards (AH), odds ratio (OR), hazard ratio (HR), complementary log (clog), complementary log–log (cloglog), normal linear regression model (NLRM), binary generalized linear model (BGLM), accelerated failure time (AFT)*

			F_Z			
Ξ	h	Meaning of	Φ	F_{SL}	F_{Exp}	F_{MEV}
$K = 2$			**Binary Regression**			
	$\mathbb{1}(k = 1)\vartheta_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}$		probit BGLM	logistic BGLM	clog BGLM	cloglog BGLM
		ϑ_1		$\log(O_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$	$\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$	$\log(\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$
		$\boldsymbol{\beta}$		log-OR	AH	log-HR
$K > 2$			**Polytomous Regression**			
	$\boldsymbol{e}_{K-1}(k)^\top\boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}$			discrete PO		discrete PH
		$\boldsymbol{e}_{K-1}(k)^\top\boldsymbol{\vartheta}_1$		$\log(O_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$	$\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$	$\log(\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$
		$\boldsymbol{\beta}$		log-OR	AH	log-HR
	$\boldsymbol{e}_{K-1}(k)^\top\boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}(k)$			non-PO		non-PH
		$\boldsymbol{e}_{K-1}(k)^\top\boldsymbol{\vartheta}_1$		$\log(O_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$	$\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$	$\log(\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$
$\mathbb{N}$			**Count Regression**			
	$\boldsymbol{a}_{\text{Bs},M}(\lfloor y \rfloor)^\top\boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}$					
		$\boldsymbol{a}_{\text{Bs},M}(\lfloor y \rfloor)^\top\boldsymbol{\vartheta}_1$		$\log(O_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$	$\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$	$\log(\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$
		$\boldsymbol{\beta}$		log-OR	AH	log-HR
	$\boldsymbol{a}_{\text{Bs},M}(\lfloor y \rfloor)^\top\boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}(\lfloor y \rfloor)$					
		$\boldsymbol{a}_{\text{Bs},M}(\lfloor y \rfloor)^\top\boldsymbol{\vartheta}_1$		$\log(O_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$	$\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$	$\log(\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$
$\mathbb{R}^+$			**Survival Analysis**			
	$\log(y) + \vartheta_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}$					Exponential AFT
		$\boldsymbol{\beta}$		log-OR		log-HR
	$(1, \log(y))^\top\boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}$		log-normal AFT	log-logistic AFT		Weibull AFT
		$(1, \log(y))^\top\boldsymbol{\vartheta}_1$		$\log(O_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$	$\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$	$\log(\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$
		$\boldsymbol{\beta}$		log-OR	AH	log-HR
$\mathbb{R}$			**Continuous Regression and Survival Analysis**			
	$y\vartheta_1 - \vartheta_2 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}$		NLRM			
		ϑ_1^{-2}	variance			
		$(\vartheta_2 + \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta})/\vartheta_1$	mean			
	$\boldsymbol{a}_{\text{Bs},M}(y)^\top\boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}$				Aalen AH	Cox PH
		$\boldsymbol{a}_{\text{Bs},M}(y)^\top\boldsymbol{\vartheta}_1$		$\log(O_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$	$\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$	$\log(\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$
		$\boldsymbol{\beta}$		log-OR	AH	log-HR
	$\boldsymbol{a}_{\text{Bs},M}(y)^\top\boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}(y)$			Distribution Regression		Time-varying Cox
		$\boldsymbol{a}_{\text{Bs},M}(y)^\top\boldsymbol{\vartheta}_1$		$\log(O_Y(y \mid \boldsymbol{X} = \boldsymbol{0}))$	$\Lambda_Y(y \mid \boldsymbol{X} = \boldsymbol{0})$	$\log(\Lambda_Y(y \mid \boldsymbol{x} = \boldsymbol{0}))$

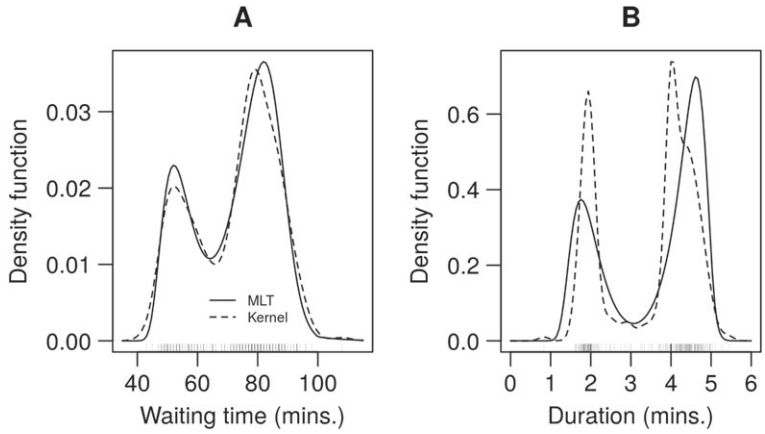

Fig. 1. Old Faithful geyser. Estimated density for waiting times (A) between and duration (B) of eruptions by the most likely transformation model (MLT) and kernel smoothing. Note that the kernel estimator was based on the imputed duration times 2, 3 and 4 for short, medium and long eruptions at night (as are the rugs in B).

daytime measurements. At night, the observations were left censored ('short' eruption), interval censored ('medium' eruption) or right censored ('long' eruption). This censoring was widely ignored in analyses of the Old Faithful data because most non-parametric kernel techniques cannot deal with censoring. We applied the transformation model $(\Phi, \boldsymbol{a}_{\text{Bs},8}(\text{duration}), \boldsymbol{\vartheta})$ based on the exact log-likelihood function under eight linear constraints and obtained the unconditional density depicted in Fig. 1B. In Hothorn (2017b), results for $M = 40$ are computed, which led to almost identical estimates of the distribution function.

Quantile regression: head circumference The Fourth Dutch Growth Study is a cross-sectional study on growth and development of the Dutch population younger than 22 years. Stasinopoulos & Rigby (2007) fitted a growth curve to head circumferences (HCs) of 7040 boys using a generalized additive models for location, scale and shape (GAMLSS) model with a Box–Cox t distribution describing the first four moments of HC conditionally on age. The model showed evidence of kurtosis, especially for older boys. We fitted the same growth curves by the conditional transformation model $(\Phi, (\boldsymbol{a}_{\text{Bs},3}(\text{HC})^\top \otimes \boldsymbol{b}_{\text{Bs},3}(\text{age}^{1/3})^\top)^\top, \boldsymbol{\vartheta})$ by maximization of the approximate log-likelihood under 3×4 linear constraints $(\boldsymbol{D}_4 \otimes \boldsymbol{I}_4)\boldsymbol{\vartheta} > 0$. Figure 2 shows the data overlaid with quantile curves obtained via inversion of the estimated conditional distributions. The figure very closely reproduces the growth curves presented in Fig. 16 of Stasinopoulos & Rigby (2007) and also indicates a certain asymmetry towards older boys.

Survival analysis: German Breast Cancer Study Group-2 trial This prospective, controlled clinical trial on the treatment of node-positive breast cancer patients was conducted by the German Breast Cancer Study Group. Out of 686 women, 246 received hormonal therapy, whereas the control group of 440 women did not. Additional variables include age, menopausal status, tumour size, tumour grade, number of positive lymph nodes, progesterone receptor and oestrogen receptor. The right-censored recurrence-free survival time is the response variable of interest.

The Cox model $(F_{\text{MEV}}, (\boldsymbol{a}_{\text{Bs},10}^\top, \mathbb{1}(\text{hormonal therapy}))^\top, \boldsymbol{\vartheta})$ implements the transformation function $h(y \mid \text{treatment}) = \boldsymbol{a}_{\text{Bs},10}(y)^\top \boldsymbol{\vartheta}_1 + \mathbb{1}(\text{hormonal therapy})\beta$, where $\boldsymbol{a}_{\text{Bs},10}^\top \boldsymbol{\vartheta}_1$ is the

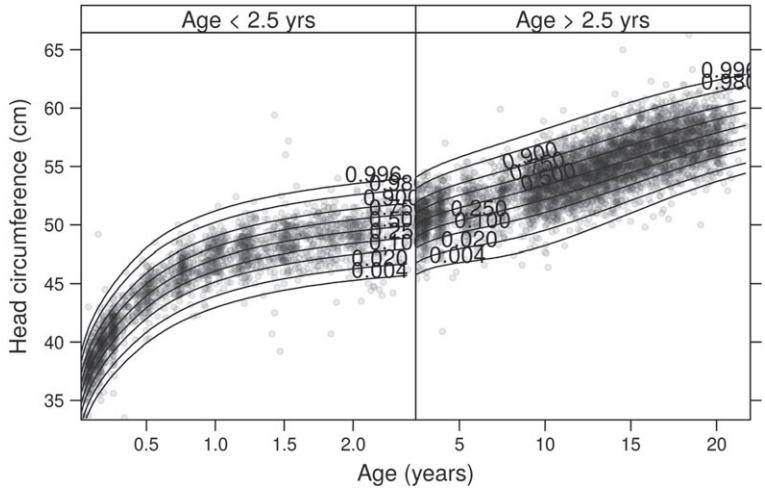

Fig. 2. Head circumference growth. Observed head circumference and age of 7040 boys with estimated quantile curves for $p = 0.04, 0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98, 0.996$.

log-cumulative baseline hazard function parameterized by a Bernstein polynomial and $\beta \in \mathbb{R}$ is the log-hazard ratio of hormonal therapy. This is the classical Cox model with one treatment parameter β but with fully parameterized baseline transformation function, which was fitted by the exact log-likelihood under ten linear constraints. The model assumes proportional hazards, an assumption whose appropriateness we wanted to assess using the non-proportional hazards model $(F_{\text{MEV}}, (\boldsymbol{a}_{\text{Bs},10}^\top \otimes (1, \mathbb{1}(\text{hormonal therapy})))^\top, \boldsymbol{\vartheta})$ with the transformation function

$$h(y \mid \text{treatment}) = \boldsymbol{a}_{\text{Bs},10}(y)^\top \boldsymbol{\vartheta}_1 + \mathbb{1}(\text{hormonal therapy}) \boldsymbol{a}_{\text{Bs},10}(y)^\top \boldsymbol{\vartheta}_2.$$

The function $\boldsymbol{a}_{\text{Bs},10}(y)^\top \boldsymbol{\vartheta}_2$ is the time-varying difference of the log-hazard functions of women without and with hormonal therapy and can be interpreted as the deviation from a constant log-hazard ratio treatment effect of hormonal therapy. Under the null hypothesis of no treatment effect, we would expect $\boldsymbol{\vartheta}_2 \equiv \mathbf{0}$. This monotonic deviation function adds ten linear constraints $\boldsymbol{D}_{11}\boldsymbol{\vartheta}_1 + \boldsymbol{D}_{11}\boldsymbol{\vartheta}_2 > 0$, which also ensure monotonicity of the transformation function for treated patients. We first compared the fitted survivor functions obtained from the model including a time-varying treatment effect with the Kaplan–Meier estimators in both treatment groups. Figure 3A shows a nicely smoothed version of the survivor functions obtained from this transformation model. Figure 3B shows the time-varying treatment effect $\boldsymbol{a}_{\text{Bs},10}(y)^\top \hat{\boldsymbol{\vartheta}}_2$, together with a 95% confidence band computed from the joint normal distribution of $\hat{\boldsymbol{\vartheta}}_2$ for a grid over time; the method is much simpler than other methods for inference on time-varying effects (e.g. Sun *et al*., 2009). The 95% confidence interval around the log-hazard ratio $\hat{\beta}$ is also plotted, and as the latter is fully covered by the confidence band for the time-varying treatment effect, there is no reason to question the treatment effect computed under the proportional hazards assumption.

In the second step, we allowed an age-varying treatment effect to be included in the model $(F_{\text{MEV}}, (\boldsymbol{a}_{\text{Bs},10}(y)^\top \otimes (\mathbb{1}(\text{hormonal therapy}), 1 - \mathbb{1}(\text{hormonal therapy})) \otimes \boldsymbol{b}_{\text{Bs},3}(\text{age})^\top)^\top, \boldsymbol{\vartheta})$. For both treatment groups, we estimated a conditional transformation function of survival time y given age parameterized as the tensor basis of two Bernstein bases. Each of the two basis functions comes with 10×3 linear constraints; therefore, the model was fitted under 60 linear constraints. Figure 4 allows an assessment of the prognostic and predictive properties of age. As the survivor functions were clearly larger for all patients treated with hormones, the positive

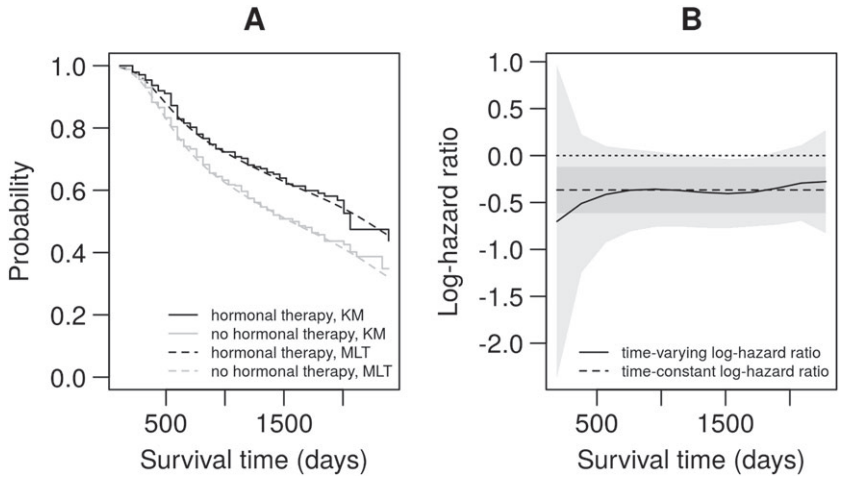

Fig. 3. German Breast Cancer Study Group-2. Estimated survivor functions by the most likely transformation model (MLT) and the Kaplan–Meier (KM) estimator in the two treatment groups (A). Verification of proportional hazards (B): the log-hazard ratio $\hat{\beta}$ (dashed line) with 95% confidence interval (dark grey) is fully covered by a 95% confidence band for the time-varying treatment effect (the time-varying log-hazard ratio is in light grey; the estimate is the solid line) computed from a non-proportional hazards model.

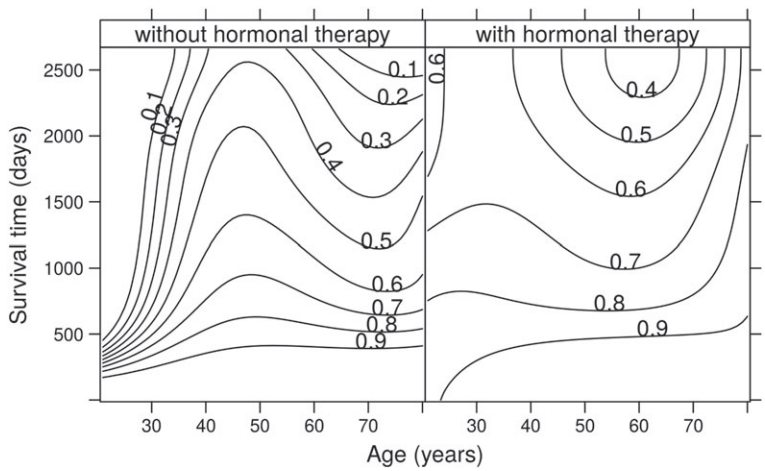

Fig. 4. German Breast Cancer Study Group-2. Prognostic and predictive effect of age. The contours depict the conditional survivor functions given treatment and age of the patient.

treatment effect applied to all patients. However, the size of the treatment effect varied greatly. The effect was most pronounced for women younger than 30 years and levelled off a little for older patients. In general, the survival times were longest for women between 40 and 60 years old. Younger women suffered the highest risk; for women older than 60 years, the risk started to increase again. This effect was shifted towards younger women when hormonal treatment was applied.

5.2. Simulation experiment

The transformation family includes linear as well as very flexible models, and we therefore illustrate the potential gain of modelling a transformation function h by comparing a very simple transformation model with a fully parametric approach and to a non-parametric approach using a data-generating process introduced by Hothorn *et al.* (2014).

In the transformation model $(\Phi, ((1, y) \otimes (1, \boldsymbol{x}^\top))^\top, \boldsymbol{\vartheta})$, two explanatory variables $\boldsymbol{x} = (x_1, x_2)^\top$ influence both the conditional mean and the conditional variance of a normal response Y. Although the transformation function is linear in y with three linear constraints, the mean and variance of Y given $\boldsymbol{x}$ depend on $\boldsymbol{x}$ in a non-linear way. The choices $x_1 \sim \mathrm{U}[0, 1], x_2 \sim \mathrm{U}[-2, 2]$ with $\boldsymbol{\vartheta} = (0, 0, -1, .5, 1, 0)$ lead to the heteroscedastic varying coefficient model

$$Y = \frac{1}{x_1 + 0.5} x_2 + \frac{1}{x_1 + 0.5} \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, 1), \tag{5}$$

where the variance of Y ranges between 0.44 and 4 depending on x_1. This model can be fitted in the GAMLSS framework under the assumptions that the mean of the normal response depends on a smoothly varying regression coefficient $(x_1 + 0.5)^{-1}$ for x_2 and that the variance is a smooth function of x_1. This model is therefore fully parametric. As a non-parametric counterpart, we used a kernel estimator for estimating the conditional distribution function of Y as a function of the two explanatory variables.

From the transformation model, the GAMLSS and kernel estimators, we obtained estimates of $F_{Y \mid \boldsymbol{X} = \boldsymbol{x}}(y)$ over a grid on y, x_1, x_2 and computed the mean absolute deviation (MAD) of the true and estimated probabilities

$$\mathrm{MAD}(x_1, x_2) = \frac{1}{n} \sum_y |F_{Y \mid \boldsymbol{X} = \boldsymbol{x}}(y) - \hat{F}_{Y \mid \boldsymbol{X} = \boldsymbol{x}, N}(y)|$$

for each pair of x_1 and x_2. Then, the minimum, the median and the maximum of the MAD values for all x_1 and x_2 were computed as summary statistics. The most likely transformation approach and its two competitors were estimated and evaluated for 100 random samples of size $N = 200$ drawn from model (5). Cross-validation was used to determine the bandwidths for the kernel-based estimators (function `npcdist()` in package **np**; for details, see Hayfield & Racine, 2008). We fitted the GAMLSS models by boosting; the number of boosting iterations was determined via sample splitting (Mayr *et al.*, 2012). To investigate the stability of the three procedures under non-informative explanatory variables, we added to the data $p = 1, \ldots, 5$ uniformly distributed variables without association to the response and included them as potential explanatory variables in the three models. The case $p = 0$ corresponds to model (5).

Figure 5 shows the empirical distributions of the minimum, median and maximum MAD for the three competitors. Except for the minimum MAD in the absence of any irrelevant explanatory variables ($p = 0$), the conditional distributions fitted by the transformation models were closer to the true conditional distribution function by means of the MAD. This result was obtained because the transformation model only had to estimate a simple transformation function, whereas the other two procedures had a difficult time approximating this simple transformation model on another scale. However, the comparison illustrates the potential improvement one can achieve when fitting simple models for the transformation function instead of more complex models for the mean (GAMLSS) or distribution function (Kernel). The kernel estimator led to the largest median MAD values but seemed more robust than GAMLSS with respect to the maximum MAD. These results were remarkably robust in the

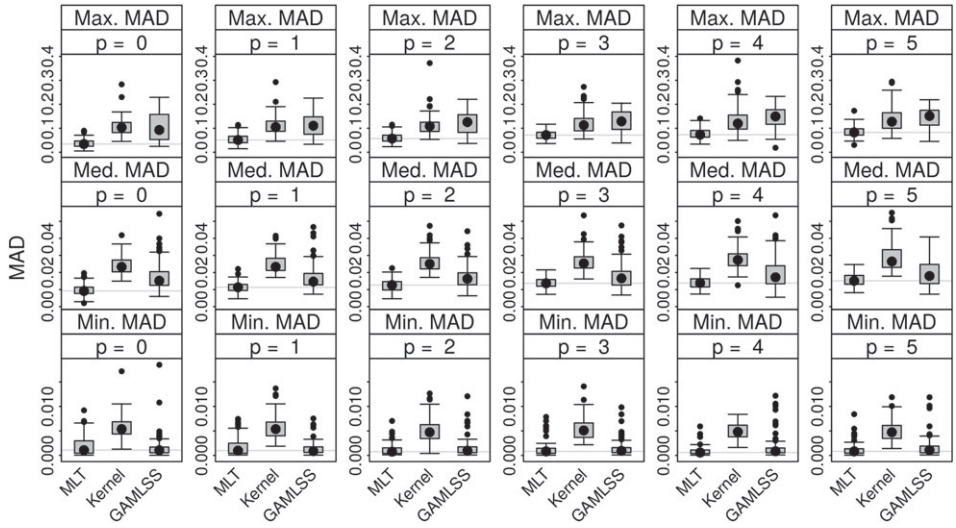

Fig. 5. Empirical evaluation. Minimum, median and maximum of the mean absolute deviation (MAD) between true and estimated probabilities for most likely transformation (MLT) models, non-parametric kernel distribution function estimation (Kernel) and generalized additive models for location, scale and shape (GAMLSS) for 100 random samples. Values on the ordinate can be interpreted as absolute differences of probabilities. The grey horizontal lines correspond to the median of MLT.

presence of up to five non-informative explanatory variables, although of course the MAD increased with the number of non-informative variables p.

6. Discussion

The contribution of a likelihood approach for the general class of conditional transformation models is interesting both from a theoretical and a practical perspective. With the range of simple to very complex transformation functions introduced in Section 4 and illustrated in Section 5, it becomes possible to understand classical parametric, semi-parametric and non-parametric models as special cases of the same model class. Thus, analytic comparisons between models of different complexity become possible. The transformation family $P_{Y,\Theta}$, the corresponding likelihood function and the most likely transformation estimator are easy to understand. This makes the approach appealing also from a teaching perspective. Connections between standard parametric models (e.g. the normal linear model) and potentially complex models for survival or ordinal data can be outlined in very simple notation, placing emphasis on the modelling of (conditional) distributions instead of just modelling (conditional) means. Computationally, the log-likelihood $\log \circ \mathcal{L}$ is linear in the number of observations N and, for contributions of 'exact continuous' responses, only requires the evaluation of the derivative h' of the transformation function h instead of integrals thereof.

Based on the general understanding of transformation models outlined in this paper, it will be interesting to study these models outside the strict likelihood world. A mixed transformation model for cluster data (Cai *et al.*, 2002; Zeng *et al.*, 2005; Choi & Huang, 2012) is often based on the transformation function $h(y \mid \boldsymbol{x}, i) = h_Y(y) + \delta_i + h_{\boldsymbol{x}}(\boldsymbol{x})$ with random intercept (or 'frailty' term) δ_i for the ith observational unit. Conceptually, a more complex deviation from the global model could be formulated as $h(y \mid \boldsymbol{x}, i) = h_Y(y) + h_Y(y, i) + h_{\boldsymbol{x}}(\boldsymbol{x})$, that is, each observational unit is assigned its own 'baseline' transformation $h_Y(y) + h_Y(y, i)$, where the

second term is an integral zero deviation from h_Y. For longitudinal data with possibly time-varying explanatory variables, the model $h(y \mid \boldsymbol{x}(t), t) = h_Y(y, t) + \boldsymbol{x}(t)\boldsymbol{\beta}(t)$ (Ding *et al.*, 2012; Wu & Tian, 2013) can also be understood as a mixed version of a conditional transformation model. The penalized log-likelihood $\log(\mathcal{L}(h \mid y)) - \text{pen}(\boldsymbol{\beta})$ for the linear transformation model $h(y \mid \boldsymbol{x}) = h_Y(y) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$ leads to Ridge-type or Lasso-type regularized models, depending on the form of the penalty term. Priors for all model parameters $\boldsymbol{\vartheta}$ allow a fully Bayesian treatment of transformation models.

It is possible to relax the assumption that F_Z is known. The simultaneous estimation of F_Z in the model $P(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(h_Y(y) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta})$ was studied by Horowitz (1996) and later extended by Linton *et al.* (2008) to non-linear functions $h_{\boldsymbol{x}}$ with parametric baseline transformation h_Y and kernel estimates for F_Z and $h_{\boldsymbol{x}}$. For AFT models, Zhang & Davidian (2008) applied smooth approximations for the density f_Z in an exact censored likelihood estimation procedure. In a similar set-up, Huang (2014) proposed a method to jointly estimate the mean function and the error distribution in a generalized linear model. The estimation of F_Z is noteworthy in additive models of the form $h_Y + h_{\boldsymbol{x}}$ because these models assume additivity of the contributions of y and $\boldsymbol{x}$ on the scale of $F_Z^{-1}(P(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}))$. If this model assumption seems questionable, one can either allow unknown F_Z or move to a transformation model featuring a more complex transformation function.

From this point of view, the distribution function F_Z in flexible transformation models is only a computational device mapping the unbounded transformation function h into the unit interval strictly monotonically, making the evaluation of the likelihood easy. Then, F_Z has no further meaning or interpretation as error distribution. A compromise could be the family of distributions $F_Z(z \mid \rho) = 1 - (1 + \rho \exp(z))^{-1/\rho}$ for $\rho > 0$ (suggested by McLain & Ghosh, 2013) with simultaneous maximum likelihood estimation of $\boldsymbol{\vartheta}$ and ρ for additive transformation functions $h = h_Y + h_{\boldsymbol{x}}$, as these models are flexible and still relatively easy to interpret.

In light of the empirical results discussed in this paper and the theoretical work of McLain & Ghosh (2013) on a Cox model with log-cumulative baseline hazard function parameterized in terms of a Bernstein polynomial with increasing order M, one might ask where the boundaries among parametric, semi-parametric and non-parametric statistics lie. The question how the order M affects results practically has been repeatedly raised; therefore, we will close our discussion by looking at a Cox model with increasing M for the German Breast Cancer Study Group-2 data. All eight baseline variables were included in the linear predictor, and we fitted the model with orders $M = 1, \ldots, 30, 35, 40, 45, 50$ of the Bernstein polynomial parameterizing the log-cumulative baseline hazard function. In Fig. 6A, the log-cumulative baseline hazard functions start with a linear function ($M = 1$) and quickly approach a function that is essentially a smoothed version of the Nelson-Aalen-Breslow estimator plotted in red. In Fig. 6B, the trajectories of the estimated regression coefficients become very similar to the partial likelihood estimates as M increased. For $M \geq 10$, for instance, the results of the 'semi-parametric' and the 'fully parametric' Cox models are practically equivalent. An extensive collection of such head-to-head comparisons of most likely transformations with their classical counterparts can be found in Hothorn (2017b). Our work for this paper and practical experience with its reference software implementation convinced us that rethinking classical models in terms of fully parametric transformations is intellectually and practically a fruitful exercise.

6.1. Computational details

A reference implementation of most likely transformation models is available in the **mlt** package (Hothorn, 2017a). All data analyses can be reproduced in the dynamic document

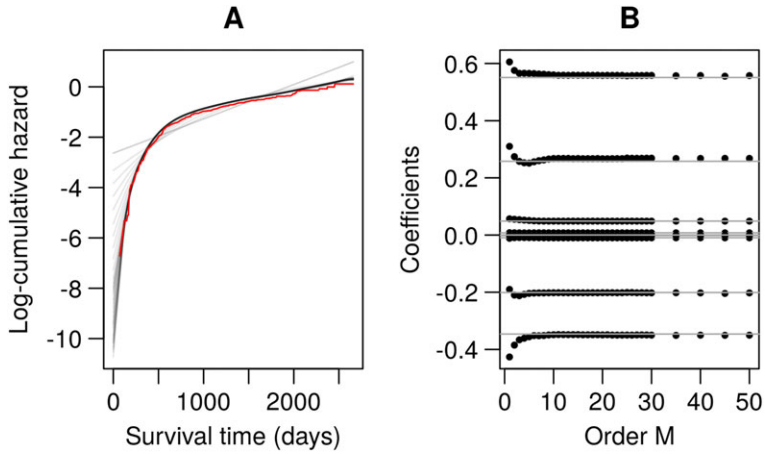

Fig. 6. German Breast Cancer Study Group-2. Comparison of exact and partial likelihood for order $\boldsymbol{M} = 1, \ldots, 30, 35, 40, 45, 50$ of the Bernstein polynomial approximating the log-cumulative baseline hazard function h_Y. The estimated log-cumulative baseline hazard functions for varying $\boldsymbol{M}$ are shown in grey, and the Nelson-Aalen-Breslow estimator is shown in red (A). The right panel (B) shows the trajectories of the regression coefficients $\hat{\boldsymbol{\beta}}$ obtained for varying $\boldsymbol{M}$, which are represented as dots. The horizontal lines represent the partial likelihood estimates.

Hothorn (2017b). Augmented Lagrangian Minimization implemented in the `auglag()` function of package **alabama** (Varadhan, 2015) was used for optimizing the log-likelihood. Package **gamboostLSS** (version 1.2-2, Hofner *et al.*, 2016) was used to fit GAMLSS models and kernel density, and distribution estimation was performed using package **np** (version 0.60-2, Racine & Hayfield, 2014). All computations were performed using R version 3.4.0 (R Core Team, 2017). Additional applications are described in an extended version of this paper (Hothorn *et al.*, 2017).

Acknowledgements

Torsten Hothorn received financial support by Deutsche Forschungsgemeinschaft under grant number HO 3242/4-1. We thank Karen A. Brune for improving the language.

References

Alzaatreh, A., Lee, C. & Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron* **71**, 63–79.

Aranda-Ordaz, F. J. (1983). An extension of the proportional hazards model for grouped data. *Biometrics* **39**, 109–117.

Babu, G. J., Canty, A. J. & Chaubey, Y. P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *J. Stat. Plan. Infer.* **105**, 377–392.

Bickel, P. J., Klaassen, C. A. J., Ritov, Ya'acov & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*, The Johns Hopkins University Press, Baltimore, U.S.A. and London, U.K.

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

Cai, T., Cheng, S. C. & Wei, L. J. (2002). Semiparametric mixed-effects models for clustered failure time data. *J. Am. Stat. Assoc.*, 514–522.

Chang, I., Hsiung, C. A., Wu, Y.-J. & Yany, C.-C. (2005). Bayesian survival analysis using Bernstein polynomials. *Scand. J. Statist.* **32**, 447–466.

Cheng, S. C., Wei, L. J. & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.
Cheng, S. C., Wei, L. J. & Ying, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *J. Am. Stat. Assoc.* **92**, 227–235.
Chernozhukov, V., Fernández-Val, I. & Melly, B. (2013). Inference on counterfactual distributions. *Econometrica* **81**, 2205–2268.
Choi, S. & Huang, X. (2012). A general class of semiparametric transformation frailty models for nonproportional hazards survival data. *Biometrics* **68**, 1126–1135.
Crowther, M. J. & Lambert, P. C. (2014). A general framework for parametric survival analysis. *Stat. Med.* **33**, 5280–5297.
Curtis, S. M. & Ghosh, S. K. (2011). A variable selection approach to monotonic regression with Bernstein polynomials. *J. Appl. Stat.* **38**, 961–976.
Ding, A. A., Tian, S., Yu, Y. & Guo, H. (2012). A class of discrete transformation survival models with application to default probability prediction. *J. Am. Stat. Assoc.* **107**, 990–1003.
Doksum, K. A. & Gasko, M. (1990). On a correspondence between models in binary regression analysis and in survival analysis. *Int. Stat. Rev.* **58**, 243–252.
Farouki, R. T. (2012). The Bernstein polynomial basis: A centennial retrospective. *Comput. Aided Geom. D.* **29**, 379–419.
Foresi, S. & Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *J. Am. Stat. Assoc.* **90**, 451–466.
Fraser, D. A. S. (1968). *The structure of inference*, John Wiley & Sons, New York, U.S.A.
Gneiting, T. & Katzfuss, M. (2014). Probabilistic forecasting. *Annu. Rev. Stat. Appl.* **1**, 125–151.
Hayfield, T. & Racine, J. S. (2008). Nonparametric econometrics: The np package. *J. Stat. Software* **27**, 1–32.
Hofner, B., Mayr, A., Fenske, N. & Schmid, M. (2016). *Gamboostlss: Boosting methods for 'GAMLSS'*, R package version 1.2-2. Available at: https://CRAN.R-project.org/package=gamboostLSS.
Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* **64**, 103–137.
Hothorn, T. (2017a). *MLT: Most likely transformations*, R package version 0.1-3. Available at: https://CRAN.R-project.org/package=mlt.
Hothorn, T. (2017b). *MLT: Most likely transformations: The mlt package*, R package vignette version 0.1-5. Available at: https://CRAN.R-project.org/package=mlt.docreg.
Hothorn, T., Kneib, T. & Bühlmann, P. (2014). Conditional transformation models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **76**, 3–27.
Hothorn, T., Möst, L. & Bühlmann, P. (2017). Most likely transformations. Tech. Rep. arXiv 1508.06749, Available at: https://arxiv.org/abs/1508.06749.
Huang, A. (2014). Joint estimation of the mean and error distribution in generalized linear models. *J. Am. Stat. Assoc.* **109**, 186–196.
Jones, M. C. & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika* **96**, 761–780.
Klein, J. P. & Moeschberger, M. K. (2003). *Survival analysis. Techniques for censored and truncated data*, (2nd edn.), Springer, New York, U.S.A.
Kooperberg, C., Stone, C. J. & Truong, Y. K. (1995). Hazard regression. *J. Am. Stat. Assoc.* **90**, 78–94.
Leorato, S. & Peracchi, F. (2015). Comparing distribution and quantile regression. Tech. Rep. 1511, Einaudi Institute for Economics and Finance, Rome, Italy Available at https://ideas.repec.org/p/eie/wpaper/1511.html.
Lindsey, J. K. (1996). *Parametric statistical inference*, Clarendon Press, Oxford, UK.
Lindsey, J. K. (1999). Some statistical heresies. *J. R. Stat. Soc. Ser. D (The Statistic.)* **48**, 1–40.
Linton, O., Sperlich, S. & van K., I. (2008). Estimation of a semiparametric transformation model. *Ann. Stat.* **36**, 686–718.
Ma, J., Heritier, S. & Lô, S. N. (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Comput. Stat. Data Anal.* **74**, 142–156.
Mallick, B. K. & Walker, S. (2003). A Bayesian semiparametric transformation model incorporating frailties. *J. Stat. Plann. Infer.* **112**, 159–174.
Mayr, A., Fenske, N., Hofner, B., Kneib, T. & Schmid, M. (2012). GAMLSS for high-dimensional data – a flexible approach based on boosting. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **61**, 403–427.
McLain, A. C. & Ghosh, S. K. (2013). Efficient sieve maximum likelihood estimation of time-transformation models. *J. Stat. Theor. Pract.* **7**, 285–303.
R Core Team (2017). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/.

Racine, J. S. & Hayfield, T. (2014). *NP: Nonparametric kernel smoothing methods for mixed data types*, R package version 0.60-2. Available at: https://CRAN.R-project.org/package=np.

Rothe, C. & Wied, D. (2013). Misspecification testing in a class of conditional distributional models. *J. Am. Stat. Assoc.* **108**, 314–324.

Royston, P. & Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat. Med.* **21**, 2175–2197.

Stasinopoulos, D. M. & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Software* **23**, 1–46.

Sun, Y., Sundaram, R. & Zhao, Y. (2009). Empirical likelihood inference for the Cox model with time-dependent coefficients via local partial likelihood. *Scand. J. Statist.* **36**, 444–462.

Thas, O., Neve, J. D., Clement, L. & Ottoy, J.-P. (2012). Probabilistic index models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **74**, 623–671.

Tutz, G. (2012). *Regression for categorical data*, Cambridge University Press, New York, U.S.A.

van de Geer, S. (2000). *Empirical processes in m-estimation*, Cambridge University Press, Cambridge, UK.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Cambridge University Press, Cambridge, UK.

Varadhan, R. (2015). *Alabama: Constrained nonlinear optimization*, R package version 2015.3-1. Available at: https://CRAN.R-project.org/package=alabama.

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

Zeng, D. & Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **69**, 507–564.

Zeng, D., Lin, D. Y. & Yin, G. (2005). Maximum likelihood estimation for the proportional odds model with random effects. *J. Am. Stat. Assoc.* **100**, 470–483.

Zhang, M. & Davidian, M. (2008). Smooth semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics* **64**, 567–576.

Received November 2016, in final form June 2017

Torsten Hothorn, Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich.

E-mail: torsten.hothorn@uzh.ch