

Conditional Transformation Models

Extended Version^{*}

Torsten Hothorn
LMU München &
Universität Zürich

Thomas Kneib
Universität Göttingen

Peter Bühlmann
ETH Zürich

Abstract

The ultimate goal of regression analysis is to obtain information about the conditional distribution of a response given a set of explanatory variables. This goal is, however, seldom achieved because most established regression models only estimate the conditional mean as a function of the explanatory variables and assume that higher moments are not affected by the regressors. The underlying reason for such a restriction is the assumption of additivity of signal and noise. We propose to relax this common assumption in the framework of transformation models. **The novel class of semiparametric regression models proposed herein allows transformation functions to depend on explanatory variables.** These transformation functions are estimated by regularised optimisation of scoring rules for probabilistic forecasts, *e.g.* the continuous ranked probability score. The corresponding estimated conditional distribution functions are consistent. **Conditional transformation models are potentially useful for describing possible heteroscedasticity, comparing spatially varying distributions, identifying extreme events, deriving prediction intervals and selecting variables beyond mean regression effects.** An empirical investigation based on a heteroscedastic varying coefficient simulation model demonstrates that semiparametric estimation of conditional distribution functions can be more beneficial than kernel-based non-parametric approaches or parametric generalised additive models for location, scale and shape.

Keywords: boosting, conditional distribution function, conditional quantile function, continuous ranked probability score, prediction intervals, structured additive regression.

1. Introduction

One of the famous “Top ten reasons to become a statistician” is that statisticians are “mean lovers” (Friedman, Friedman, and Amoo 2002), referring of course to our obsession with means. Whenever a distribution is too complex to think or expound upon, we focus on the mean as a single real number describing the centre of the distribution and block out other characteristics such as variance, skewness and kurtosis. Our willingness to simplify distributions this way is most apparent when we deal with many distributions at a time, as in a regression setting where we describe the conditional distribution $\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$ of a response $Y \in \mathbb{R}$ given different

^{*}This is an extended version of Hothorn, Kneib, and Bühlmann (2013). Please cite this reference unless you’re referring to the additional applications presented in this document.

configurations of explanatory variables $\mathbf{X} = \mathbf{x} \in \chi$. Many regression models focus on the conditional mean $\mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ and treat higher moments of the conditional distribution $\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$ as fixed or nuisance parameters that must not depend on the explanatory variables. As a consequence, model inference crucially relies on assumptions such as homoscedasticity and symmetry. Information on the scale of the response, for example prediction intervals, derived from such models also depends on these assumptions. Here, we propose a new class of conditional transformation models that allow the conditional distribution function $\mathbb{P}(Y \leq v|\mathbf{X} = \mathbf{x})$ to be estimated directly and semiparametrically under rather weak assumptions. Before we introduce this class of models in Section 2, we will attempt to set a scene of contemporary regression in the light of Gilchrist (2008) and place the major players onto this stage.

Let $Y_{\mathbf{x}} = (Y|\mathbf{X} = \mathbf{x}) \sim \mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$ denote the conditional distribution of response Y given explanatory variables $\mathbf{X} = \mathbf{x}$. We assume that $\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$ is dominated by some measure μ and has the conditional distribution function $\mathbb{P}(Y \leq v|\mathbf{X} = \mathbf{x})$. A regression model describes the distribution $\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$, or certain characteristics of it, as a function of the explanatory variables \mathbf{x} . We estimate such models based on samples of pairs of random variables (Y, \mathbf{X}) from the joint distribution $\mathbb{P}_{Y, \mathbf{X}}$. It is convenient to assume that a regression model consists of signal and noise, *i.e.* a deterministic part and an error term. In the following, we denote the error term by $Q(U)$, where $U \sim \mathcal{U}[0, 1]$ is a uniform random variable independent of \mathbf{X} and $Q: \mathbb{R} \rightarrow \mathbb{R}$ is the quantile function of an absolutely continuous distribution.

Apart from non-parametric kernel estimators of the conditional distribution function (Hall, Wolff, and Yao 1999, Hall and Müller 2003, Li and Racine 2008), there are two common ways to model the influence of the explanatory variables \mathbf{x} on the response $Y_{\mathbf{x}}$:

$$Y_{\mathbf{x}} = r(Q(U)|\mathbf{x}) \quad \text{“mean or quantile regression models” and} \quad (1)$$

$$h(Y_{\mathbf{x}}|\mathbf{x}) = Q(U) \quad \text{“transformation models”.} \quad (2)$$

For each $\mathbf{x} \in \chi$, the regression function $r(\cdot|\mathbf{x}): \mathbb{R} \rightarrow \mathbb{R}$ transforms the error term $Q(U)$ in a monotone increasing way. The inverse regression function $h(\cdot|\mathbf{x}) = r^{-1}(\cdot|\mathbf{x}): \mathbb{R} \rightarrow \mathbb{R}$ is also monotone increasing. Because h transforms the response, it is known as a transformation function, and models in the form of (2) are called transformation models.

A major assumption underlying almost all regression models of class (1) is that the regression function r is the sum of the deterministic part $r_{\mathbf{x}}: \chi \rightarrow \mathbb{R}$, which depends on the explanatory variables, and the error term:

$$r(Q(U)|\mathbf{x}) = r_{\mathbf{x}}(\mathbf{x}) + Q(U).$$

When $\mathbb{E}(Q(U)) = 0$, we get $r_{\mathbf{x}}(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$, *e.g.* linear or additive models depending on the functional form of $r_{\mathbf{x}}$. Model inference is commonly based on the normal error assumption, *i.e.* $Q(U) = \sigma\Phi^{-1}(U)$, where $\sigma > 0$ is a scale parameter and $\Phi^{-1}(U) \sim \mathcal{N}(0, 1)$. Linear heteroscedastic regression models (Carroll and Ruppert 1982) allow describing the variance as a function of the explanatory variables $Q(U) = \sigma(\mathbf{x})\tilde{Q}(U)$, where $\sigma(\mathbf{x})$ is a (usually log-linear) function of \mathbf{x} and \tilde{Q} is the quantile function of a symmetric distribution with $\mathbb{E}(\tilde{Q}(U)) = 0$. In time series analysis, GARCH models (Bollerslev 1986) share this idea. A novel semiparametric approach is extended generalised additive models, where additive functions of the explanatory variables describe location, scale and shape (GAMLSS) of a certain parametric conditional distribution of the response given the explanatory variables (Rigby and Stasinopoulos 2005).

If the assumption of a certain parametric form of the conditional distribution is questionable, $r_{\mathbf{x}}$ describes the τ quantile of $Y_{\mathbf{x}}$ when the quantile function Q is such that $Q(\tau) = 0$ for some $\tau \in (0, 1)$. This leads us to quantile regression (Koenker 2005), which is according to Stigler (2010) the “best approach to robust methods in higher dimensional linear model problems”. Estimating the complete conditional quantile function is less straightforward since we have to fit separate models for a grid of probabilities τ , and the resulting regression quantiles may cross. Solutions to this problem can be obtained by combining all quantile fits in one joint model based on, for example, location-scale models (He 1997) or quantile sheets (Schnabel and Eilers 2012), or by monotonising the estimated quantile curves using non-decreasing rearrangements (Dette and Volgushev 2008).

For transformation models (2), additivity is assumed on the scale of the inverse regression function h :

$$h(Y_{\mathbf{x}}|\mathbf{x}) = h_Y(Y_{\mathbf{x}}) + h_{\mathbf{x}}(\mathbf{x}).$$

When $\mathbb{E}(Q(U)) = 0$, we get $-h_{\mathbf{x}}(\mathbf{x}) = \mathbb{E}(h_Y(Y_{\mathbf{x}})) = \mathbb{E}(h_Y(Y)|\mathbf{X} = \mathbf{x})$. The monotone transformation function $h_Y : \mathbb{R} \rightarrow \mathbb{R}$ does not depend on \mathbf{x} and might be known in advance (Box-Cox transformation models with fixed parameters, accelerated failure time models) or is commonly treated as a nuisance parameter (Cox model, proportional odds model). One is usually interested in estimating the function $h_{\mathbf{x}} : \chi \rightarrow \mathbb{R}$, which describes the conditional mean of the *transformed* response $h_Y(Y_{\mathbf{x}})$. The class of transformation models is rich and very actively researched, most prominently in literature on the analysis of survival data. For example, the linear Weibull accelerated failure time model assumes a log transformation $h_Y(Y_{\mathbf{x}}) = \log(Y_{\mathbf{x}})$, a linear function for the conditional mean of the log-transformed response $h_{\mathbf{x}}(\mathbf{x}) = \mathbf{x}^{\top} \boldsymbol{\alpha}$, and a Weibull-distributed error term $Q(U) = \sigma Q_{\text{Weibull}}(U)$. For the Cox additive model, $h_Y(Y_{\mathbf{x}}) = \log(\Lambda(Y_{\mathbf{x}}))$ is based on the unspecified integrated baseline hazard function Λ , $h_{\mathbf{x}}(\mathbf{x}) = \sum_{j=1}^J h_{\mathbf{x},j}(\mathbf{x})$ is the sum of J smooth terms depending on the explanatory variables and $Q(U) = -\log(-\log(U))$ is the quantile function of the extreme value distribution. The proportional odds model has $h_Y(Y_{\mathbf{x}}) = \log(\Gamma(Y_{\mathbf{x}}))$, with Γ being an unknown monotone increasing function, and $Q(U) = \log(U/(1-U))$ is the quantile function of the logistic distribution. Doksum and Gasko (1990) discussed the flexibility of this class of models, and Cheng, Wei, and Ying (1995) introduced a generic algorithm for linear transformation model estimation, that is, for models with $h_{\mathbf{x}}(\mathbf{x}) = \mathbf{x}^{\top} \boldsymbol{\alpha}$, treating the transformation function h_Y as a nuisance.

In recent years, transformation models have been extended in two directions. In the first direction, more flexible forms for the conditional mean function $h_{\mathbf{x}}$ have been introduced, *e.g.* the partially linear transformation model $h_{\mathbf{x}}(\mathbf{x}) = \mathbf{x}^{\top}(0, \boldsymbol{\alpha})^{\top} + h_{\text{smooth}}(x_1)$ (where h_{smooth} is a smooth function of the first variable x_1 ; Lu and Zhang 2010), the varying coefficient model $h_{\mathbf{x}}(\mathbf{x}) = \mathbf{x}^{\top}(0, 0, \boldsymbol{\alpha})^{\top} + h_{\text{smooth}}(x_1)x_2$ (Chen and Tong 2010), random effects models (Zeng, Lin, and Yin 2005), and various approaches to additive transformation and accelerated failure time models, such as the boosting approaches by Lu and Li (2008) and Schmid and Hothorn (2008). In the second direction, a number of authors have considered algorithms that estimate h_Y and (partially) linear functions $h_{\mathbf{x}}$ simultaneously, usually by a spline expansion of h_Y (*e.g.* Shen 1998, Cheng and Wang 2011), as an alternative to the common practise of estimating h_Y *post-hoc* by some non-parametric procedure such as the Breslow estimator.

Although the transformation function h_Y is typically treated as an infinite dimensional nuisance parameter, it is important to note that h_Y contains information about higher moments

of $Y_{\mathbf{x}}$, most importantly variance and skewness. Simultaneous estimation of h_Y and $h_{\mathbf{x}}$ is therefore extremely attractive because we can obtain information about the mean and higher moments of the transformed response at the same time. However, owing to the decomposition of the regression function r or the transformation function h into both a deterministic part depending on the explanatory variables ($r_{\mathbf{x}}$ or $h_{\mathbf{x}}$) and a random part depending on the response (h_Y) or error term ($Q(U)$), higher moments of the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ must not depend on the explanatory variables in mean regression and transformation models. As a consequence, the corresponding models cannot capture heteroscedasticity or skewness that is induced by certain configurations of the explanatory variables. Therefore, we cannot detect these potentially interesting patterns, and our models will perform poorly when probability forecasts, prediction intervals or other functionals of the conditional distribution are of special interest.

Recently, [Wu, Tian, and Yu \(2010\)](#) proposed a novel transformation model for longitudinal data that partially addresses this issue. For responses and explanatory variables $\mathbf{X}(t)$ observed at time t , the model assumes

$$h(Y_{\mathbf{x}}|t, \mathbf{x}) = h_Y(Y_{\mathbf{x}}|t) + \mathbf{x}(t)^{\top} \boldsymbol{\alpha}(t).$$

Here, the transformation h_Y is conditional on time, and higher moments may vary with time. However, since h_Y does not depend on the explanatory variables \mathbf{x} , these higher moments may not vary with one or more of the explanatory variables. In the context of longitudinal data with functional explanatory variables, [Chen and Müller \(2012\)](#) consider a similar model, where the regression coefficients for functional principle components may depend on time t and the response $Y_{\mathbf{x}}$. Our contribution is a class of transformation models where the transformation function is conditional on the explanatory variables in the sense that the transformation function, and therefore higher moments of the conditional distribution of the response, may depend on potentially all explanatory variables. As a consequence, the models suggested here are able to deal with heteroscedasticity and skewness that can be regressed on the explanatory variables, and we will show that reliable estimates of the complete conditional distribution function and functionals thereof can be obtained.

We will introduce the conditional transformation models (Section 2), discuss the underlying model assumptions, and embed the estimation problem in the empirical risk minimisation framework (Section 3). For the sake of simplicity, we restrict ourselves to continuous responses Y that have been observed without censoring. Similar to other transformation models, conditional transformation models are flexible enough to deal with discrete responses and survival times, as will be discussed in later sections. We present a computationally efficient algorithm for fitting the models in Section 4. We study the asymptotic properties of the estimated conditional distribution functions in Section 5. The practical benefits of modelling the influence of explanatory variables on the variance and higher moments of the response' distribution are demonstrated in Section 6 with a special emphasis on distributional characteristics of childhood nutrition in India and on prediction intervals for birth weights of small foetus. Finally, we use a heteroscedastic varying coefficient simulation model to evaluate the empirical performance of the proposed algorithm and compare the quality of conditional distribution functions estimated by a conditional transformation model and established parametric and non-parametric procedures in Section 7.

2. Conditional Transformation Models

An attractive feature of transformation models is their close connection to the conditional distribution function. With the transformation function $h(Y_{\mathbf{x}}|\mathbf{x}) = Q(U)$, one can evaluate the conditional distribution function of response Y given the explanatory variables \mathbf{x} via

$$\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}) = \mathbb{P}(h(Y|\mathbf{x}) \leq h(v|\mathbf{x})) = F(h(v|\mathbf{x}))$$

with absolute continuous distribution function $F = Q^{-1}$. For additive transformation functions $h = h_Y + h_{\mathbf{x}}$, the conditional distribution function reads $F(h(v|\mathbf{x})) = F(h_Y(v) + h_{\mathbf{x}}(\mathbf{x}))$, *i.e.* the distribution is evaluated for a transformed and shifted version of Y . Higher moments only depend on the transformation h_Y and thus cannot be influenced by the explanatory variables. Consequently, one has to avoid the additivity in the model $h = h_Y + h_{\mathbf{x}}$ to allow the explanatory variables to impact also higher moments. We therefore suggest a novel transformation model based on an alternative additive decomposition of the transformation function h into J partial transformation functions for all $\mathbf{x} \in \chi$:

$$h(v|\mathbf{x}) = \sum_{j=1}^J h_j(v|\mathbf{x}), \quad (3)$$

where $h(v|\mathbf{x})$ is the monotone transformation function of v . In this model, the transformation function $h(Y_{\mathbf{x}}|\mathbf{x})$ and the partial transformation functions $h_j(\cdot|\mathbf{x}) : \mathbb{R} \rightarrow \mathbb{R}$ are conditional on \mathbf{x} in the sense that not only the mean of $Y_{\mathbf{x}}$ depends on the explanatory variables. For this reason, we coin models of the form (3) *Conditional Transformation Models* (CTMs). Clearly, model (3) imposes an assumption, namely additivity of the conditional distribution function on the scale of the quantile function Q :

$$Q(\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x})) = \sum_{j=1}^J h_j(v|\mathbf{x}).$$

It should be noted that here we assume additivity of the transformation function h and not additivity on the scale of the regression function r as it is common for additive mean or quantile regression models (1). Furthermore, monotonicity of h_j is sufficient but not necessary for h being monotone. Of course, we have to make further assumptions on h_j to obtain reasonable models, but these assumptions are problem specific, and we will therefore postpone these issues until Section 6. To ensure identifiability, we assume without loss of generality that the partial transformation functions are centred around zero $\mathbb{E}_Y \mathbb{E}_{\mathbf{X}} h_j(Y|\mathbf{X}) = 0$ for all $j = 1, \dots, J$ for non-systematic error terms ($\mathbb{E}(Q(U)) = 0$).

3. Estimating Conditional Transformation Models

The estimation of conditional distribution functions can be reformulated as a mean regression problem since $\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}) = \mathbb{E}(I(Y \leq v) | \mathbf{X} = \mathbf{x})$ for the binary event $Y \leq v$; this connection is widely used (*e.g.* by Hall and Müller 2003, Chen and Müller 2012). Similar to the approach of fitting multiple quantile regression models to obtain an estimate of the conditional quantile function, one could estimate the models $\mathbb{E}(I(Y \leq v) | \mathbf{X} = \mathbf{x})$ for a grid of v values separately. However, we instead suggest estimating conditional transformation

models by the application of an integrated loss function that allows the whole conditional distribution function to be obtained in one step.

Let ρ denote a function of measuring the loss of the probability $F(h(v|\mathbf{X}))$ for the binary event $Y \leq v$. One candidate loss function is

$$\begin{aligned} \rho_{\text{bin}}((Y \leq v, \mathbf{X}), h(v|\mathbf{X})) &:= -[I(Y \leq v) \log\{F(h(v|\mathbf{X}))\} + \\ &\quad \{1 - I(Y \leq v)\} \log\{1 - F(h(v|\mathbf{X}))\}] \geq 0, \end{aligned}$$

the negative log-likelihood of the binomial model $(Y \leq v|\mathbf{X} = \mathbf{x}) \sim B(1, F(h(v|\mathbf{x})))$ for the binary event $Y \leq v$ with link function $Q = F^{-1}$. Alternatively, one may consider the squared or absolute error losses

$$\begin{aligned} \rho_{\text{sqe}}((Y \leq v, \mathbf{X}), h(v|\mathbf{X})) &:= \frac{1}{2}|I(Y \leq v) - F(h(v|\mathbf{X}))|^2 \geq 0 \\ \rho_{\text{abe}}((Y \leq v, \mathbf{X}), h(v|\mathbf{X})) &:= |I(Y \leq v) - F(h(v|\mathbf{X}))| \geq 0. \end{aligned}$$

The squared error loss ρ_{sqe} is also known as the Brier score, and the absolute loss ρ_{abe} has been applied for assessing survival probabilities in the Cox model by [Schemper and Henderson \(2000\)](#). We define the loss function ℓ for estimating conditional transformation models as integrated loss ρ with respect to a measure μ dominating the conditional distribution $\mathbb{P}(Y \leq v|\mathbf{X} = \mathbf{x})$:

$$\ell((Y, \mathbf{X}), h) := \int \rho((Y \leq v, \mathbf{X}), h(v|\mathbf{X})) d\mu(v) \geq 0.$$

In the context of scoring rules, the loss ℓ based on ρ_{sqe} is known as the continuous ranked probability score (CPRS) or integrated Brier score and is a proper scoring rule for assessing the quality of probabilistic or distributional forecasts (see [Gneiting and Raftery 2007](#), for an overview). It seems natural to apply these scores as loss functions for model estimation, but we are only aware of the work of [Gneiting, Raftery, Westveld III, and Goldman \(2005\)](#), who directly optimise the CPRS for estimating Gaussian predictive probability density functions for continuous weather variables. In the context of non-parametric or semiparametric estimation of conditional distribution functions, minimisation of the empirical analog of the risk function

$$\mathbb{E}_{Y, \mathbf{X}} \ell((Y, \mathbf{X}), h) = \int \int \rho((y \leq v, \mathbf{x}), h(v|\mathbf{x})) d\mu(v) d\mathbb{P}_{Y, \mathbf{X}}(y, \mathbf{x}) \geq 0$$

for estimating conditional distribution functions has not yet been considered.

Model estimation based on the risk $\mathbb{E}_{Y, \mathbf{X}} \ell((Y, \mathbf{X}), h)$ is reasonable because the corresponding optimisation problem is convex and attains its minimum for the true conditional transformation function h . We summarise these facts in the following lemma, whose proof is given in the Appendix.

Lemma 1. *The risk $\mathbb{E}_{Y, \mathbf{X}} \ell((Y, \mathbf{X}), h)$ is convex in h for convex losses ρ in h . The population minimiser of $\mathbb{E}_{Y, \mathbf{X}} \ell((Y, \mathbf{X}), h)$ for $\rho = \rho_{\text{bin}}$ and $\rho = \rho_{\text{sqe}}$ is $h(v|\mathbf{x}) = Q(\mathbb{P}(Y \leq v|\mathbf{X} = \mathbf{x}))$. For $\rho = \rho_{\text{abe}}$, the minimiser is*

$$h(v|\mathbf{x}) = \begin{cases} -\infty & : \mathbb{P}(Y \leq v|\mathbf{X} = \mathbf{x}) \leq 0.5 \\ \infty & : \mathbb{P}(Y \leq v|\mathbf{X} = \mathbf{x}) > 0.5. \end{cases}$$

The corresponding empirical risk function defined by the data is

$$\hat{\mathbb{E}}_{Y,\mathbf{X}}\ell((Y, \mathbf{X}), f) = \int \int \rho((y \leq v, \mathbf{x}), h(v|\mathbf{x})) d\mu(v) d\hat{\mathbb{P}}_{Y,\mathbf{X}}(y, \mathbf{x}) \geq 0.$$

Based on an i.i.d. random sample $(Y_i, \mathbf{X}_i) \sim \mathbb{P}_{Y,\mathbf{X}}, i = 1, \dots, N$ of N observations from the joint distribution of response and explanatory variables, we define $\hat{\mathbb{P}}_{Y,\mathbf{X}}$ as the distribution putting mass $w_i > 0$ on observation i ($w_i \equiv N^{-1}$ for the empirical distribution). For computational convenience, we also approximate the measure μ by the discrete uniform measure $\hat{\mu}$, which puts mass n^{-1} on each element of the equi-distant grid $v_1 < \dots < v_n \in \mathbb{R}$ over the response space. The number of grid points n has to be sufficiently large to closely approximate the integral. The empirical risk is then

$$\begin{aligned} \hat{\mathbb{E}}_{Y,\mathbf{X}}\ell((Y, \mathbf{X}), h) &= \sum_{i=1}^N w_i n^{-1} \sum_{\iota=1}^n \rho((Y_i \leq v_\iota, \mathbf{X}_i), h(v_\iota|\mathbf{X}_i)) \\ &= n^{-1} \sum_{i=1}^N \sum_{\iota=1}^n w_i \rho((Y_i \leq v_\iota, \mathbf{X}_i), h(v_\iota|\mathbf{X}_i)). \end{aligned} \quad (4)$$

This risk is the weighted empirical risk for loss function ρ evaluated at the observations $(Y_i \leq v_\iota, \mathbf{X}_i)$ for $i = 1, \dots, N$ and $\iota = 1, \dots, n$. Consequently, we can apply algorithms for fitting generalised additive models to the binary responses $Y_i \leq v_\iota$ under loss ρ for estimating model (3). Although this seems to be rather straightforward, there are two issues to consider. First, simply expanding the observations over the grid $v_1 < \dots < v_n$ increases the computational complexity by n , which, even for moderately large sample sizes N , renders computing and storage rather burdensome. Second, unconstrained minimisation of the empirical risk, *i.e.* no smoothness of h in its first argument and h being independent of the conditioning \mathbf{x} , leads to estimates $F(\hat{h}(v|\mathbf{x})) = \hat{\mathbb{P}}(Y \leq v) = N^{-1} \sum_{i=1}^N I(Y_i \leq v)$, *i.e.* the empirical cumulative distribution function of Y for ρ_{bin} and ρ_{sqr} with $w_i = N^{-1}$. For ρ_{abe} , the empirical risk is minimised by $F(\hat{h}(v|\mathbf{x})) = 0$ for all v with $\hat{\mathbb{P}}(Y \leq v) < 0.5$ and otherwise by $F(\hat{h}(v|\mathbf{x})) = 1$. Therefore, careful regularisation is absolutely necessary to obtain reasonable models that lead to smooth conditional distribution functions (*i.e.* smoothing in the Y -direction) and that are similar for similar configurations of the explanatory variables (*i.e.* smoothing in the \mathbf{X} -direction). Instead of adding a direct penalisation term to the empirical risk, we propose in the next section a boosting algorithm for empirical risk minimisation that indirectly controls the functional form and complexity of the estimate \hat{h} .

4. Boosting Conditional Transformation Models

We propose to fit conditional transformation models (3) by a variant of component-wise boosting for minimising the empirical risk (4) with penalisation. In this class of algorithms, regularisation is achieved indirectly via the application of penalised base-learners and the complexity of the whole model is controlled by the number of boosting iterations. We refer the reader to Bühlmann and Hothorn (2007) for a detailed introduction to component-wise boosting.

For conditional transformation models, we parameterise the partial transformation functions for all $j = 1, \dots, J$ as

$$h_j(v|\mathbf{x}) = \left(\mathbf{b}_j(\mathbf{x})^\top \otimes \mathbf{b}_0(v)^\top \right) \boldsymbol{\gamma}_j \in \mathbb{R}, \quad \boldsymbol{\gamma}_j \in \mathbb{R}^{K_j K_0}, \quad (5)$$

where $\mathbf{b}_j(\mathbf{x})^\top \otimes \mathbf{b}_0(v)^\top$ denotes the tensor product of two sets of basis functions $\mathbf{b}_j : \chi \rightarrow \mathbb{R}^{K_j}$ and $\mathbf{b}_0 : \mathbb{R} \rightarrow \mathbb{R}^{K_0}$. Here, \mathbf{b}_0 is a basis along the grid of v values that determines the functional form of the response transformation. The basis \mathbf{b}_j defines how this transformation may vary with certain aspects of the explanatory variables. The tensor product may be interpreted as a generalised interaction effect (further illustrated in Section 6). For each partial transformation function h_j , we typically want to obtain an estimate that is smooth in its first argument v and smooth in the conditioning variable \mathbf{x} . Therefore, the bases are supplemented with appropriate, pre-specified penalty matrices $\mathbf{P}_j \in \mathbb{R}^{K_j \times K_j}$ and $\mathbf{P}_0 \in \mathbb{R}^{K_0 \times K_0}$, inducing the penalty matrix $\mathbf{P}_{0j} = (\lambda_0 \mathbf{P}_j \otimes \mathbf{1}_{K_0} + \lambda_j \mathbf{1}_{K_j} \otimes \mathbf{P}_0)$ with smoothing parameters $\lambda_0 \geq 0$ and $\lambda_j \geq 0$ for the tensor product basis. The base-learners corresponding to the partial transformation functions fitted to the negative gradients in each iteration of the boosting algorithm are then Ridge-type linear models with penalty matrix \mathbf{P}_{0j} . In more detail, we apply the following algorithm for fitting conditional transformation models with partial transformation functions of the form (5):

Algorithm: Boosting for Conditional Transformation Models

(Init) Initialise the parameters $\gamma_j^{[0]} \equiv 0$ for $j = 1, \dots, J$, the step-size $\nu \in (0, 1)$ and the smoothing parameters $\lambda_j, j = 0, \dots, J$. Define the grid $v_1 < Y_{(1)} < \dots < Y_{(N)} \leq v_n$. Set $m := 0$.

(Gradient)

Compute the negative gradient:

$$U_{ii} := - \left. \frac{\partial}{\partial h} \rho(Y_i \leq v_i, \mathbf{X}_i), h \right|_{h=\hat{h}_{ii}^{[m]}}$$

with $\hat{h}_{ii}^{[m]} = \sum_{j=1}^J (\mathbf{b}_j(\mathbf{X}_i)^\top \otimes \mathbf{b}_0(v_i)^\top) \gamma_j^{[m]}$.

Fit the base-learners for $j = 1, \dots, J$:

$$\hat{\beta}_j = \arg \min_{\beta \in \mathbb{R}^{K_j K_0}} \sum_{i=1}^N \sum_{i=1}^n w_i \left\{ U_{ii} - \left(\mathbf{b}_j(\mathbf{X}_i)^\top \otimes \mathbf{b}_0(v_i)^\top \right) \beta \right\}^2 + \beta^\top \mathbf{P}_{0j} \beta \quad (6)$$

with penalty matrix \mathbf{P}_{0j} .

Select the best base-learner:

$$j^* = \arg \min_{j=1, \dots, J} \sum_{i=1}^N \sum_{i=1}^n w_i \left\{ U_{ii} - \left(\mathbf{b}_j(\mathbf{X}_i)^\top \otimes \mathbf{b}_0(v_i)^\top \right) \hat{\beta}_j \right\}^2$$

(Update) the parameters $\gamma_{j^*}^{[m+1]} = \gamma_{j^*}^{[m]} + \nu \hat{\beta}_{j^*}$ and keep all other parameters fixed, *i.e.* $\gamma_j^{[m+1]} = \gamma_j^{[m]}, j \neq j^*$.

Iterate (Gradient) and (Update)

(Stop) if $m = M$. Output the final model

$$\hat{\mathbb{P}}(Y \leq v | \mathbf{X} = \mathbf{x}) = F\left(\hat{h}^{[M]}(v | \mathbf{x})\right) = F\left(\sum_{j=1}^J \left(\mathbf{b}_j(\mathbf{x})^\top \otimes \mathbf{b}_0(v)^\top\right) \boldsymbol{\gamma}_j^{[M]}\right)$$

as a function of arbitrary $v \in \mathbb{R}$ and $\mathbf{x} \in \chi$.

Before we investigate the asymptotic properties of the resulting estimates, we will discuss some details of this generic algorithm in the following.

Model Specification. The basis functions \mathbf{b}_0 and \mathbf{b}_j determine the form of the fitted model, and their choice is problem specific. In the simplest situation, in which the conditional distribution of Y given only one numeric explanatory variable x_1 shall be estimated, one could use the basis functions $\mathbf{b}_0(v) = (1, v)^\top$ and $\mathbf{b}_1(\mathbf{x}) = (1, x_1)^\top$. The corresponding base-learner is then defined by the linear function

$$((1, x_1) \otimes (1, v)) \boldsymbol{\gamma}_1 = (1, v, x_1, x_1 v) \boldsymbol{\gamma}_1.$$

For each x_1 , the transformation is linear in v with intercept $\gamma_1 + \gamma_3 x_1$ and slope $\gamma_2 + \gamma_4 x_1$, *i.e.* not only the mean may depend on x_1 but also the variance. Restricting, for example, $\mathbf{b}_0(v)$ to be constant, *i.e.* $\mathbf{b}_0(v) \equiv 1$, allows the effects of explanatory variables to be restricted to the mean alone. Assuming $\mathbf{b}_1(\mathbf{x}) \equiv 1$, on the other hand, yields a transformation function that is not affected by any explanatory variable. More flexible basis functions, *e.g.* B -spline basis functions, allow also for higher moments to depend on the explanatory variables. We illustrate appropriate choices of basis functions in Section 6.

Computational Complexity. For the estimation of base-learner parameters $\boldsymbol{\beta}_j$ in (6), it is not necessary to evaluate the Kronecker product \otimes in (5) and to compute the $nN \times K_0 K_j$ design matrix for the j th base-learner. The base-learners used here are a special form of multidimensional smooth linear array models (Currie, Durban, and Eilers 2006), where efficient algorithms for computing Ridge estimates (6) exist. The number of multiplications required for fitting the j th base-learner is approximately $c^6/(c^2/N - 1)$, instead of $N^2 c^4$ for the simplest case with $c = K_0 = K_j$ and $N = n$ (see Table 2 in Currie *et al.* 2006), and the required memory for storing the design matrices is of the order $NK_j + NK_0$, instead of NnK_jK_0 . Note that only the gradient vector is of length Nn ; all other objects can be stored in vectors or matrices growing with either N or n , and an explicit expansion of the observations $(Y_i \leq v_i, \mathbf{X}_i)$ for $i = 1, \dots, N$ and $\iota = 1, \dots, n$ is not necessary.

Choice of Tuning Parameters. The number of boosting iterations M is the most important tuning parameter determined by resampling, *e.g.* by k -fold cross-validation or bootstrapping. For the latter resampling scheme, the weights w_i in (4) are drawn from an N -dimensional multinomial distribution with constant probability parameters $p_i \equiv N^{-1}$, $i = 1, \dots, N$. The out-of-bootstrap (OOB) empirical risk with weights $w_i^{\text{OOB}} = I(w_i = 0)$ is then used as a measure to assess the quality of the distributional forecasts for a varying number of boosting iterations M . It should be noted that the loss function used to fit the models is the same function that is used as a scoring rule to assess the quality of the probabilistic forecasts of the OOB observations.

The smoothing parameters $\lambda_j, j = 0, \dots, J$ in the penalty matrices are not tuned but rather defined such that the j th base-learner has low degrees of freedom. For our computations, we simplified the penalty term to $\mathbf{P}_{0j} = \lambda_j(\mathbf{P}_j \otimes \mathbf{1}_{K_0} + \mathbf{1}_{K_j} \otimes \mathbf{P}_0)$, *i.e.* one parameter controls the smoothness in both directions. Following Hofner, Hothorn, Kneib, and Schmid (2011a), the parameters λ_j were defined such that each base-learner has the same overall low degree of freedom. Note that the degree of freedom of the estimated partial transformation function adapts to the complexity inherent in the data via the number of boosting iterations M (Bühlmann and Yu 2003). Different smoothness in the two directions can be imposed by choosing different basis functions for \mathbf{b}_0 and \mathbf{b}_j , *e.g.* a linear basis function for \mathbf{b}_j and B -splines for \mathbf{b}_0 .

Other parameters, such as knots or degrees of basis functions or the number n of grid points the integrated loss function ℓ is approximated with are not considered as tuning parameters. The resulting estimates are rather insensitive to their different choices. Also, we do not consider the distribution function F or the loss function ρ as tuning parameter but assume that these are part of the model specification. Different versions of F and ρ lead to different negative gradients; these are given in the Appendix.

Monotonicity. The resulting estimate $\hat{h}^{[M]}(v|\mathbf{x})$ is not automatically monotone in its first argument. Monotonicity and smoothness in the Y -direction depend on each other, and too-complex estimates tend to suffer from non-monotonicity. Empirically, based on experiments reported in Sections 6 and 7, non-monotonicity is a problem in poorly-fitting models, due to either misspecification, overfitting, or a low signal-to-noise ratio. From our point of view, inspecting the model for non-monotonicity is helpful for model diagnostics and can be dealt with by reducing model complexity. Alternatively, there are three possible modifications to the algorithm that can be implemented to enforce monotonicity: (i) fit base-learners under monotonicity constraints in (6), *e.g.* by using the iterative re-penalisation suggested by Eilers (2005) and applied to boosting by Hofner, Müller, and Hothorn (2011c), (ii) check monotonicity for each base-learner and select the best among the monotone candidates only, or (iii) select the base-learner such that it is the best one among all candidates that lead to monotone updates in $h^{[m]}$. None of these approaches had to be used for our empirical studies, in which all resulting estimates were monotone for the appropriate number of boosting iterations M .

Model Diagnostics and Overfitting. Another convenient feature of transformation models is that, with the correct model h for absolute continuous random variables Y , the errors $E_i = h(Y_i|\mathbf{X}_i), i = 1, \dots, N$ are distributed according to F . Therefore, if the observed residuals $\hat{E}_i^{[M]} = \hat{h}^{[M]}(Y_i|\mathbf{X}_i)$ are unlikely to come from distribution F , *e.g.* assessed using quantile-quantile plots or a Kolmogorov-Smirnov statistic, the model is likely to fit the data poorly. However, a good agreement between $\hat{E}_i^{[M]}$ and F does not necessarily mean that the explanatory variables describe the response well. A high correlation between the ranking of the residuals and the ranking of the responses Y_1, \dots, Y_N means that the estimated conditional distribution is very close to the unconditional empirical distribution of the responses. In this case, either the model may overfit or the response may be independent of the explanatory variables.

The fitted model may also be used to draw novel responses for given explanatory variables

using the model-based bootstrap via

$$\tilde{Y}_i = \left\{ v : Q(U_i) = \hat{h}^{[M]}(v|\mathbf{X}_i) \right\}, \quad i = 1, \dots, N$$

where U_1, \dots, U_N are i.i.d. uniform random variables. The stability of the model can now be investigated by refitting the model with observations $(\tilde{Y}_i, \mathbf{X}_i), i = 1, \dots, N$.

5. Consistency of Boosted Conditional Transformation Models

The boosting algorithm presented here is a variant of L_2 WCBoost (Bühlmann and Yu 2003) applied to dependent observations with more general base-learners. In this section, we will develop a consistency result for the squared error loss ρ_{sqe} . For simplicity, we consider the case in which the procedure is used with $F(h) = h$ as the identity function, *i.e.* the error term is uniformly distributed. Thus, we consider conditional transformation models of the form

$$\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}) = h(v|\mathbf{x}) = \sum_{j=1}^J h_j(v|x_j),$$

where the partial transformation function h_j is conditional on the j th explanatory variable in $\mathbf{x} = (x_1, \dots, x_J) \in \chi$ and $\mathbb{E}_Y(N^{-1} \sum_{i=1}^N h(Y|\mathbf{X}_i)) = 1/2$. Our analysis is for the fixed design case with deterministic explanatory variables \mathbf{X}_i or when conditioning on all \mathbf{X}_i s. A modification for the random design case could be pursued along arguments similar to those for L_2 Boosting as in Bühlmann (2006).

As in Section 4, we use a basis expansion of $h(v|\mathbf{x})$:

$$h_{N,\gamma}(v|\mathbf{x}) = \sum_{j=1}^J \left(\mathbf{b}_j(x_j)^\top \otimes \mathbf{b}_0(v)^\top \right) \boldsymbol{\gamma}_j = \sum_{j=1}^J \sum_{k_0=1}^{K_{0,N}} \sum_{k_1=1}^{K_{1,N}} \gamma_{j,k_0,k_1} b_{0,k_0}(v) b_{j,k_1}(x_j),$$

where for the sake of simplicity the number of basis functions $K_{1,N}$ is equal for all x_j .

Consider the (empirical) risk functions

$$R_{n,N}(h) = (nN)^{-1} \sum_{i=1}^N \sum_{v=1}^n (I(Y_i \leq v_i) - h(v_i|\mathbf{X}_i))^2$$

and

$$R_{n,N,\mathbb{E}}(h) = (nN)^{-1} \sum_{i=1}^N \sum_{v=1}^n \mathbb{E}[(I(Y_i \leq v_i) - h(v_i|\mathbf{X}_i))^2].$$

Denote the projected parameter by

$$\boldsymbol{\gamma}_{0,N} = \arg \min_{\boldsymbol{\gamma}} R_{n,N,\mathbb{E}}(h_{N,\gamma}). \quad (7)$$

We make the following assumptions:

(A1) The coefficient vector $\gamma_{0,N}$ is sparse and satisfies

$$\|\gamma_{0,N}\|_1 = o\left(\sqrt{\frac{N}{\log(J_N K_{0,N} K_{1,N})}}\right) \quad (N \rightarrow \infty).$$

Thereby, the dimensionality $J = J_N$ can grow with N .

(A2) The basis functions satisfy: for some $0 < C < \infty$,

$$\|b_{0,k_0}\|_\infty \leq C, \quad \|b_{j,k_1}\|_\infty \leq C \quad \forall j, k_0, k_1.$$

(A3)

$$(nN)^{-1} \sum_{i=1}^N \sum_{v=1}^n h_{\gamma_{0,N}}(v_i | \mathbf{X}_i)^2 \leq D < \infty \quad \forall n, N$$

Assumption (A1) is an ℓ_1 -norm sparsity assumption, (A2) is a mild restriction since we are modeling $I(Y \leq v)$, and (A3) requires that the signal strength does not diverge as $n, N \rightarrow \infty$.

Theorem 1. Assume (A1)-(A3). Then, for fixed n or for $n = n_N \rightarrow \infty$ ($N \rightarrow \infty$), and for $M = M_N \rightarrow \infty$ ($N \rightarrow \infty$), $M_N = o(\sqrt{N/\log(J_N K_{0,N} K_{1,N})})$:

$$(nN)^{-1} \sum_{i=1}^N \sum_{v=1}^n (h_{\hat{\gamma}^{[M]}}(v_i | \mathbf{X}_i) - h_{\gamma_{0,N}}(v_i | \mathbf{X}_i))^2 = o_P(1) \quad (N \rightarrow \infty).$$

A proof is given in the Appendix.

Convergence of $h_{\gamma_{0,N}}(v|\mathbf{x})$ to the true function $h(v|\mathbf{x})$ involves approximation theory to achieve

$$(nN)^{-1} \sum_{i=1}^N \sum_{v=1}^n (h_{\gamma_{0,N}}(v_i | \mathbf{X}_i) - h(v_i | \mathbf{X}_i))^2 = o(1) \quad (n, N \rightarrow \infty). \quad (8)$$

We typically would want to estimate the function $h(v|\mathbf{x})$ well over the whole domain, *e.g.* $[a_v, b_v] \times \chi$. This may be too ambitious if $J = \dim(\chi) = J_N$ grows with N . Hence, we restrict ourselves to the setting where the number of active variables $J_{\text{act}} < \infty$ is fixed (from the active set S), *i.e.*

$$h(v|\mathbf{x}) = \sum_{j \in S} h_j(v|x_j), \quad S \subseteq \{1, \dots, J\} \quad \text{with } |S| = J_{\text{act}}.$$

For the approximation, we typically would need $K_{0,N}, K_{1,N} \rightarrow \infty$ ($N \rightarrow \infty$) for suitable basis functions and $n = n_N \rightarrow \infty$ ($N \rightarrow \infty$); furthermore, the grid $v_1 < v_2 < \dots < v_n$ should become dense as $n = n_N \rightarrow \infty$, and also the values $\mathbf{X}_1^{\text{act}}, \dots, \mathbf{X}_N^{\text{act}}$ should become dense in $\chi_S \subseteq \chi$ as $N \rightarrow \infty$ (here, $\mathbf{X}^{\text{act}} = \{X_j; j \in S\} \in \chi_S$). If $J = J_N$ grows, but the number of active variables in the model $J_{\text{act}} < \infty$ is fixed, then some uniform approximation $h_{\gamma_{0,N}}(v|\mathbf{x}) \rightarrow h(v|\mathbf{x})$ is possible under regularity conditions.

We provide a summary for a typical situation.

Corollary 1. *Consider the setting as in Theorem 1, with $J = J_N$ potentially growing but fixed dimensionality of the active variables $J_{\text{act}} < \infty$, $n = n_N \rightarrow \infty$ ($N \rightarrow \infty$), and the functions are sufficiently regular such that (8) holds. Then, for $M = M_N$ as in Theorem 1:*

$$(nN)^{-1} \sum_{i=1}^N \sum_{i=1}^n (h_{\hat{\gamma}^{[M]}}(v_i | \mathbf{X}_i) - h(v_i | \mathbf{X}_i))^2 = o_P(1) \quad (n, N \rightarrow \infty).$$

This result states that the estimated $h_{\hat{\gamma}^{[M]}}$ are consistent for the true transformation function h .

6. Applications

In this section, we present analyses with special emphasis on higher moments of the conditional distribution, which have received less attention in previous analyses of these problems. We show that semiparametric regression using conditional transformation models is a valuable tool for detecting interesting patterns beyond the conditional mean.

“Evolution Canyon” Bacteria

The *Bacillus simplex* populations from “Evolution Canyons” I and II in Israel have recently developed into a model study of bacterial adaptation and speciation under heterogeneous environmental conditions (Sikorski and Nevo 2005). These two canyons represent similar ecological sites, 40 km apart, in which the orientation of the sun yields both a strongly sun-exposed, hot, ‘African’, south-facing slope and a cooler, mesic, lush, ‘European’, north-facing slope within a distance of only 50–400 m. Based on DNA sequences, the *B. simplex* population phylogenetically splits into two major groups, GL₁ and GL₂. These main groups are further subdivided into phylogenetic groups, the so-called ‘putative ecotypes’ (PE), which show a clear preference for one of the two slope types (Sikorski and Nevo 2005, Sikorski, Pukall, and Stackebrandt 2008b). GL₂ is composed of only PE₁ and PE₂; GL₁ consists of PE₃–PE₉.

Sikorski *et al.* (2008b) analysed the physiological properties of the bacteria that might explain their characteristic slope-type preferences. For example, the physical integrity of the cell membrane at different temperatures is crucial for cell survival, and particularly its fatty acid composition is of substantial importance. Sikorski, Brambilla, Kroppenstedt, and Tindall (2008a) compared the mean contents of the fatty acids that tolerate high and low temperatures of the *B. simplex* ecotypes. The data showed heteroscedastic variances across putative ecotypes, and Herberich, Sikorski, and Hothorn (2010) analysed the data using heteroscedasticity consistent covariance estimation techniques. Here, we aim at estimating the whole conditional distribution of the fatty acid contents (FA) of each of the six putative ecotypes PE₃–PE₇ and PE₉ from GL₁.

The simplest conditional transformation model allowing for heteroscedasticity reads

$$\mathbb{P}(\text{FA} \leq v | \text{PE} = \text{PE}_k) = \Phi(\alpha_{0,k} + \alpha_k v), \quad k \in \{3, \dots, 7, 9\}. \quad (9)$$

The base-learner is defined by a linear basis $\mathbf{b}_0(v) = (1, v)^\top$ for the grid variable and a dummy-encoding basis $\mathbf{b}_1(\text{PE}) = (I(\text{PE} = \text{PE}_3), \dots, I(\text{PE} = \text{PE}_9))^\top$ for the six putative ecotypes $\{3, \dots, 7, 9\}$. The resulting 12-dimensional parameter vector γ_1 of the tensor product base-learner then consists of separate intercept and slope parameters for each of the putative

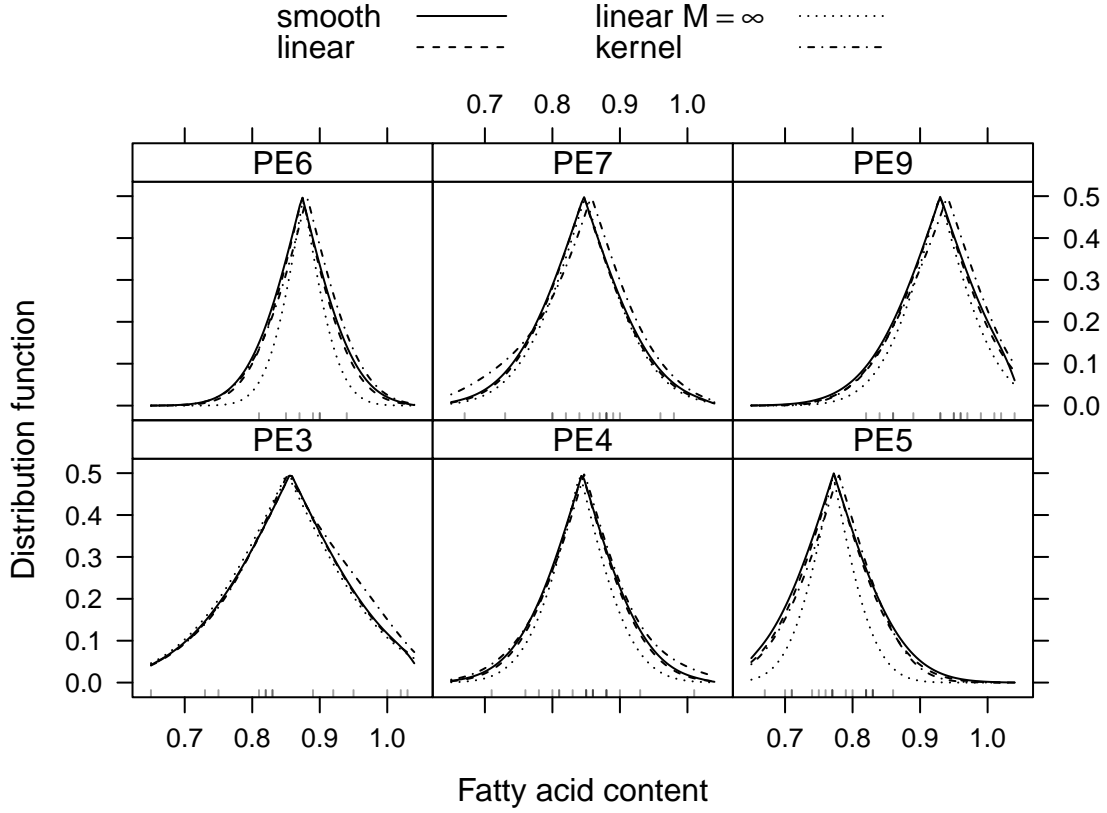


Figure 1: Evolution Canyon Bacteria. Estimated conditional distribution functions (depicted as $pI(p < 0.5) + (1 - p)I(p > 0.5)$ for probability p) of the fatty acid content of bacteria from six different putative (PE) ecotypes. Model (9) was fitted with (“linear”) and without (“linear $M = \infty$ ”) early stopping via the bootstrap whereas model (10), here denoted “smooth”, was stopped early. The observations are given as rugs.

ecotypes. Since both bases are parametric, we set the penalty matrices \mathbf{P}_0 and \mathbf{P}_1 to zero and choose $\lambda_0 = \lambda_1 = 0$, *i.e.* the base-learner does not penalise the estimates. Note that since we assume normality for the linear function $\alpha_{0,k} + \alpha_k \text{FA} \sim \mathcal{N}(0, 1)$, also the fatty acid content FA is assumed to be normal with mean *and* variance depending on the putative ecotype. Since no penalisation is defined for the base-learners, penalisation of the model parameters depends on the number of boosting iterations M . For a very large number of iterations, the algorithm converges to the estimate that is obtained by maximising the likelihood of the linear probit model fitted to the binary responses $\text{FA}_i \leq v_i$, and consequently we can fit this model by probit regression on the expanded observations in this simple setup.

We can relax the normal assumption on FA by allowing for more flexible transformations in the model

$$\mathbb{P}(\text{FA} \leq v | \text{PE} = \text{PE}_k) = \Phi(h(v | \text{PE} = \text{PE}_k)). \quad (10)$$

Now $\mathbf{b}_0(v)$ is a vector of B -spline basis functions evaluated at v for some reasonable choice of knots, while \mathbf{b}_1 remains as above. Hence, instead of assuming separate linear effects for

the putative ecotypes, we now assume separate non-parametric effects parameterised in terms of B -splines. To achieve smoothness of these non-parametric effects along the v -grid, we specify the penalty matrix \mathbf{P}_0 as $\mathbf{P}_0 = \mathbf{D}^\top \mathbf{D}$ with second-order difference matrix \mathbf{D} . We do not penalise differences between the estimated functions of different putative ecotypes, *i.e.* $\mathbf{P}_1 = 0$, and therefore the penalty matrix of the tensor product is simply given by $\mathbf{P}_{01} = \lambda_0 \text{diag}(\mathbf{P}_0, \dots, \mathbf{P}_0)$. Although we still assume $h(\text{FA}|\text{PE} = \text{PE}_k) \sim \mathcal{N}(0, 1)$, the function h may now be non-linear, and thus the conditional distribution of FA given PE may be any distribution that can be generated by a monotone transformation of a standard normal. In this sense, the model (10) is non-parametric. The validity of the normal assumption on FA is plausible when the estimated functions $\hat{h}(v|\text{PE} = \text{PE}_k)$ are essentially linear in v . An alternative non-parametric estimate can be obtained by kernel smoothing for mixed data types (Li and Racine 2008, Hayfield and Racine 2008), and we compare the two fits in Figure 1.

We fitted model (9) without ($M = \infty$) and with early stopping (via 25-fold bootstrap stratified by PE) and model (10) with early stopping to the data and in addition report the result of kernel smoothing, whose bandwidth was determined via cross-validation. The conditional distribution functions of fatty acid content given the putative ecotype are depicted in Figure 1. For probabilities larger than 0.5, the estimated conditional distribution functions were mirrored at the 0.5 horizontal line to allow for easier graphical inspection of medians, variances and potential skewness.

With respect to the conditional median, the four models lead to very similar results and support the conclusion from earlier investigations (Sikorski *et al.* 2008a, Herberich *et al.* 2010) that the fatty acid content of *B. simplex* from putative ecotype PE₅ is smaller than the others and that from PE₉ is larger than the others, with the remaining ones showing no differences. The variability cannot be assumed to be constant across different putative ecotypes, but there is no indication of asymmetry. Early stopping and bandwidth choice via resampling methods lead to almost the same estimated distribution functions. Minimising the empirical risk function without early stopping (“linear $M = \infty$ ”; this is equivalent to linear probit regression on expanded observations) leads to slightly smaller variances in PE₅ and PE₆. The difficulty in discriminating between the linear model (9) and the more flexible model (10) and the lack of asymmetry in the plots indicate that a normal assumption on fatty acid content is justifiable.

Childhood Nutrition in India

Childhood undernutrition is one of the most urgent problems in developing and transition countries. To provide information not only on the nutritional status but also on health and population trends in general, Demographic and Health Surveys (DHS) conduct nationally representative surveys on fertility, family planning, maternal and child health, as well as child survival, HIV/AIDS, malaria, and nutrition. The resulting data – from more than 200 surveys in 75 countries so far – are available for research purposes at www.measuredhs.com.

Childhood nutrition is usually measured in terms of a Z score that compares the nutritional status of children in the population of interest with the nutritional status in a reference population. The nutritional status is expressed by anthropometric characteristics, *i.e.* height for age; in cases of chronic childhood undernutrition, the reduced growth rate in human development is termed stunted growth or stunting. The Z score, which compares an anthropometric

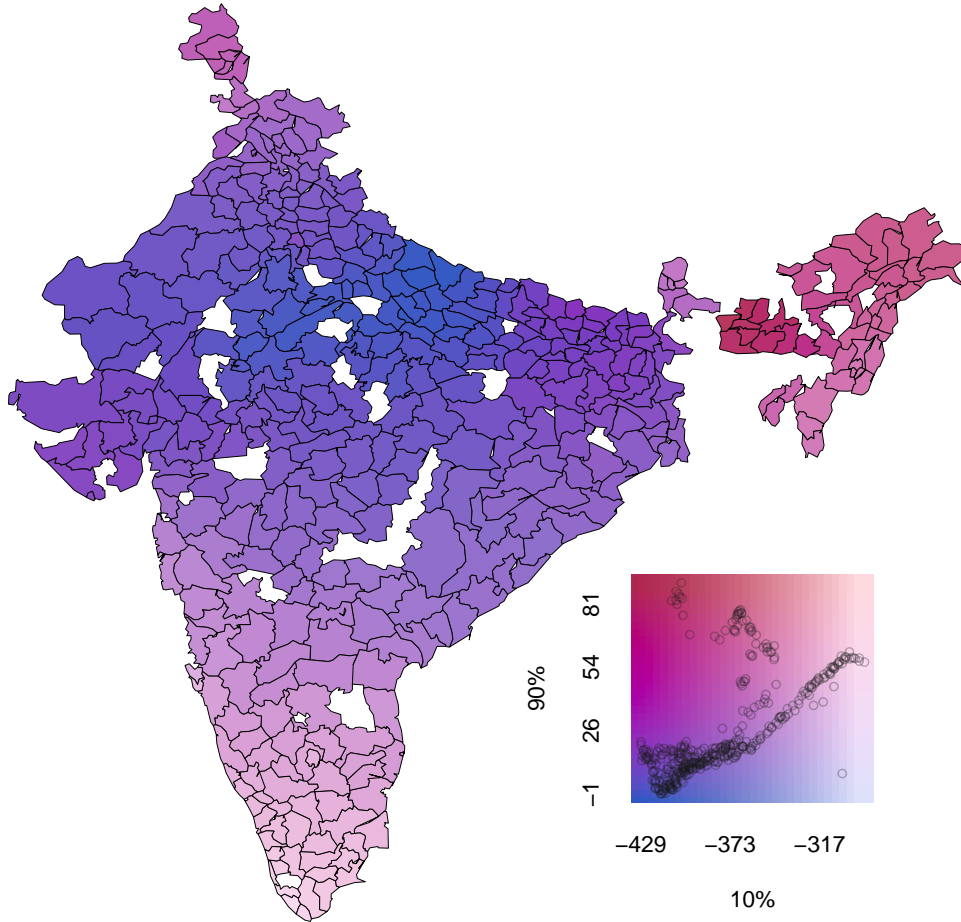


Figure 2: Childhood Nutrition in India. Colour-coded map of the 10% and 90% conditional quantiles of the Z score. Each dot in the colour legend corresponds to one district with the respective colour in the map. Blue values in the northern part of India correspond to small lower and upper quantiles. Red values, especially in the eastern Meghalaya and Assam states, indicate small lower quantiles but at the same time large upper quantiles. In the southern part of India, the lower quantiles are largest with moderate upper quantiles. White parts indicate districts with no observations.

characteristic of child i to values from a reference population, is given as

$$Z_i = \frac{AC_i - m}{s}$$

where AC denotes the anthropometric characteristic of interest and m and s correspond to median and (a robust estimate for the) standard deviation in the reference population (stratified with respect to age, gender, and some other variables). While weight might be considered as the most obvious indicator for undernutrition, we will focus on stunting, *i.e.* insufficient $AC = \text{height for age}$, in the following. Stunting provides a measure of chronic undernutrition, whereas insufficient weight for age might result from either acute or chronic undernutrition. Note that the Z score, despite its name, is not assumed to be normal.

Here we focus on estimating the whole distribution of the Z score measure for childhood nutrition in India, one of the fastest growing economies and the second-most populated country in the world. Our investigation is based on India's 1998–1999 Demographic and Health Survey (DHS, [NFHS 2000](#)) on 24,166 children visited during the survey in 412 of the 640 districts of India. The lower quantiles of this distribution can be used to assess the severity of childhood undernutrition, whereas the upper quantiles give us information about the nutritional status of children in families with above-average nutritional status.

The simplest conditional transformation model allowing for district-specific means and variances reads

$$\mathbb{P}(Z \leq v | \text{district} = k) = \Phi(\alpha_{0,k} + \alpha_k v), \quad k = 1, \dots, 412.$$

The base-learner is defined by a linear basis $\mathbf{b}_0(v) = (1, v)^\top$ for the grid variable and a dummy-encoding basis $\mathbf{b}_1(\text{district}) = (I(\text{district} = 1), \dots, I(\text{district} = k))^\top$ for the 412 districts. The resulting 824-dimensional parameter vector $\boldsymbol{\gamma}_1$ of the tensor product base-learner then consists of separate intercept and slope parameters for each of the districts of India. Note that since we assume normality for the linear function $\alpha_{0,k} + \alpha_k Z \sim \mathcal{N}(0, 1)$, also the Z score is assumed to be normal with mean *and* variance depending on the district.

We can relax the normal assumption on Z by allowing for more flexible transformations in the model

$$\mathbb{P}(Z \leq v | \text{district} = k) = \Phi(h(v | \text{district} = k)), \quad k = 1, \dots, 412. \quad (11)$$

Now $\mathbf{b}_0(v)$ is a vector of B -spline basis functions evaluated at v for some reasonable choice of knots, while \mathbf{b}_1 remains as above. Hence, instead of assuming separate linear effects for the districts, we now assume separate non-parametric effects parameterised in terms of B -splines. To achieve smoothness of these non-parametric effects along the v -grid, we specify the penalty matrix \mathbf{P}_0 as $\mathbf{P}_0 = \mathbf{D}^\top \mathbf{D}$ with second-order difference matrix \mathbf{D} . It makes sense to induce spatial smoothness on the conditional distribution functions of neighbouring districts since we do not expect the distribution of the Z score to change much from one district to its neighbouring districts. In fact, spatial smoothing is absolutely necessary in this example since otherwise we would estimate 412 separate distribution functions for the districts in India. To implement spatial smoothness of neighbouring districts, the penalty matrix \mathbf{P}_1 is chosen as an adjacency matrix, where the off-diagonal elements indicate whether two districts are neighbours (represented with a value of -1) or not (represented with a value of 0). The diagonal of the adjacency matrix contains the number of neighbours for the

corresponding district. The estimated conditional transformation function $\hat{h}(Z|\text{district} = k)$ can be interpreted as a transformation of the Z scores in district k to standard normality. Because the number of observations is large and the base-learner is fitted with penalisation, we stopped the boosting algorithm according to the in-sample empirical risk.

From the estimated conditional distribution functions, we compute the τ quantiles of the Z score for each district via

$$\hat{Q}(\tau|\text{district} = k) = \inf\{v : \Phi(\hat{h}(v|\text{district} = k)) \geq \tau\}.$$

The conditional 10% and 90% quantiles are depicted in a colour-coded map in Figure 2. The spatially smooth estimated lower and upper conditional quantiles shown simultaneously allow differentiation between three groups of districts: (A) districts with small lower and upper conditional quantiles (blue, especially in the Uttar Pradesh state), where the Z score is stochastically smaller than that of the remaining parts of India and thus all children are less well fed; (B) districts with more severe inequality, *i.e.* small lower but at the same time large upper quantiles (red, in the Meghalaya and Assam states); and (C) districts with relatively large lower and upper quantiles, which indicates a relatively good nutrition status of all children in the southern districts of India (violet, in Andhra Pradesh, Madhya Pradesh, Maharashtra, Tamil Nadu, and Kerala).

Head Circumference Growth

The Fourth Dutch Growth Study (Fredriks, van Buuren, Burgmeijer, Meulmeester, Beuker, Brugman, Roede, Verloove-Vanhorick, and Wit 2000) is a cross-sectional study that measures growth and development of the Dutch population between the ages of 0 and 22 years. The study measured, among other variables, head circumference (HC) and age of 7482 males and 7018 females. Stasinopoulos and Rigby (2007) analysed the head circumference of 7040 males with explanatory variable age using a GAMLSS model with a Box-Cox t distribution describing the first four moments of head circumference conditionally on age. The models show evidence of kurtosis, especially for older boys. We estimate the whole conditional distribution function via the conditional transformation model

$$\mathbb{P}(\text{HC} \leq v|\text{age} = x) = \Phi(h(v|\text{age} = x)).$$

The base-learner is the tensor product of B -spline basis functions $\mathbf{b}_0(v)$ for head circumference and B -spline basis functions for $\text{age}^{1/3}$. The root transformation just helps to cover the data better with equidistant knots. The penalty matrices \mathbf{P}_0 and \mathbf{P}_1 penalise second-order differences, and thus \hat{h} will be a smooth bivariate tensor product spline of head circumference and age. It is important to note that smoothing takes place in both dimensions. Consequently, the conditional distribution functions will change only slowly with age, which is a reasonable assumption. Since the number of observations is also large, we stopped the algorithm based on the in-sample empirical risk.

Figure 3 shows the data overlaid with quantile curves obtained via inversion of the estimated conditional distributions. The figure can be directly compared with Figure 16 of Stasinopoulos and Rigby (2007) and also indicates a certain asymmetry towards older boys.

Deer-vehicle Collisions

Collisions of vehicles with roe deer are a serious threat to human health and animal welfare.

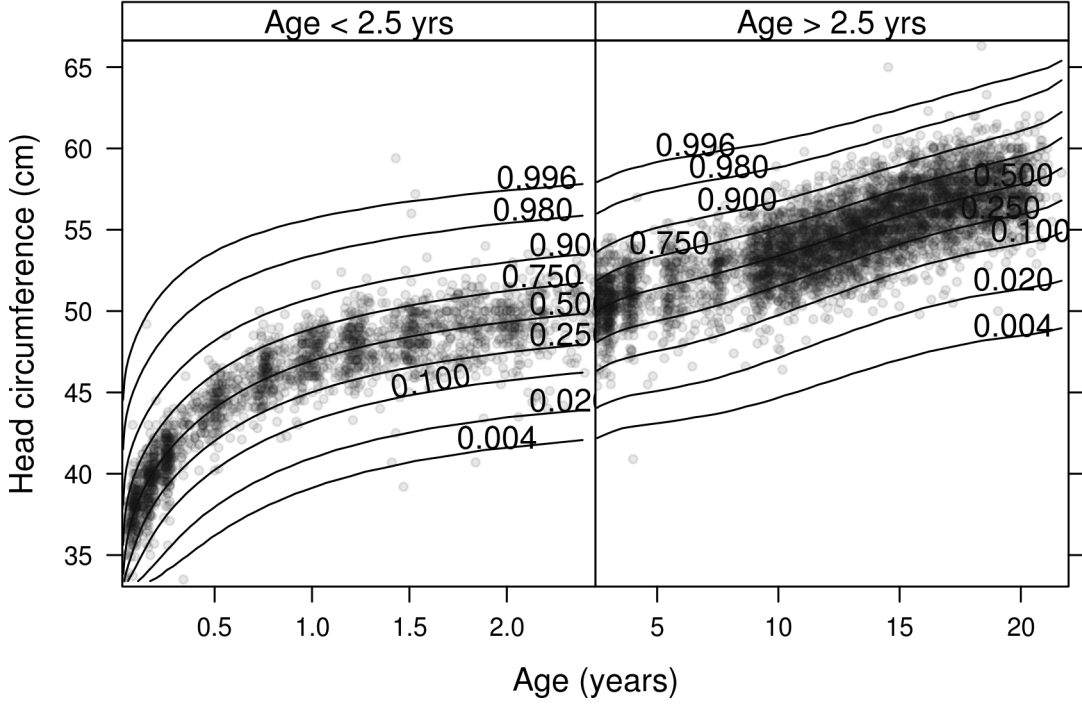


Figure 3: Head Circumference Growth. Observed head circumference and age for 7040 boys with estimated quantile curves for $\tau = 0.04, 0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98, 0.996$.

In Bavaria, Germany, more than 40,000 deer-vehicle collisions (DVCs) take place every year. [Hothorn, Brandl, and Müller \(2012\)](#) investigated the spatial distribution of the risk of deer-vehicle collisions; here we focus on the temporal aspect of the risk for two years, 2006 and 2009. For all 74,650 collisions reported to the police in these two years, we attributed each accident to the specific day of the year.

Although the number of DVCs is a discrete random variable, the distribution of the number of DVCs conditional on the day of the year can be estimated by means of an appropriate base-learner using the model

$$\mathbb{P}(\text{DVCs} \leq v | \text{day} = x_1, \text{year} = x_2) = \Phi(h_1(v | \text{day} = x_1) + h_2(v | \text{day} = x_1, \text{year} = x_2)).$$

Here, $\hat{\mu}$ is the counting measure with support v_1, \dots, v_N equal to the support of the empirical distribution of the response. Conceptually, the basis function \mathbf{b}_0 should allow for $n = N$ parameters (one for each v_i), whose first-order differences should not become too large. To restrict the number of parameters in the base-learners, we use B -splines to approximate such a discrete function on the v -grid. It should further be noted that the day of year is a discrete cyclic random variable. Therefore, we chose $\mathbf{b}_1(x_1)$ as cyclic B -splines of the day, which are obtained by a simple modification of the B -spline design matrix and the difference penalty that results from fusing the two ends of the co-domain. In analogy, a cyclic B -spline is applied to the varying coefficient term $\mathbf{b}_2(x_1, x_2) = \mathbf{b}_1(x_1) \times I(x_2 = 2009)$, which captures temporal differences between the two years and yields a cyclic B -spline of the days in 2009. Since the data are discrete, we only penalise first-order differences in both base-learners.

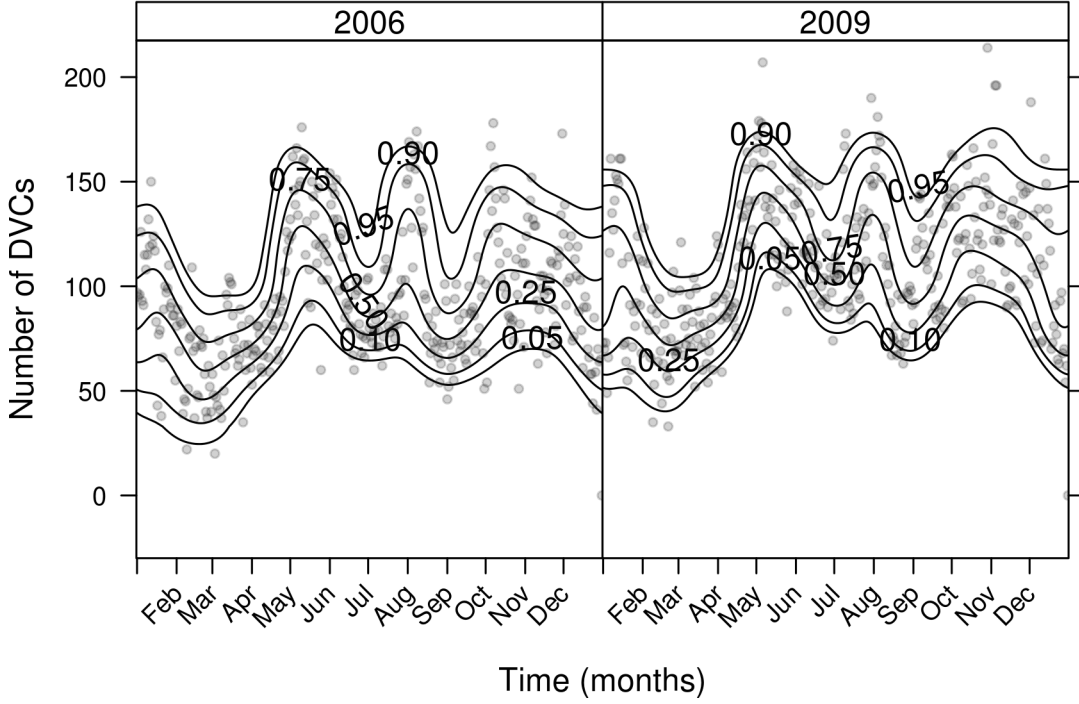


Figure 4: Deer-vehicle Collisions. Number of deer-vehicle collisions (DVCs) per day in 2006 and 2009 in Bavaria, Germany, with estimated quantile curves for $\tau = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$.

Figure 4 shows three risk peaks. The first one occurs early in May – the beginning of the growing and buck hunting season – and ends mid-June. A second and sharper peak is observed in the first week of August and corresponds to the mating season of roe deer. After a low-risk period of approximately six weeks, the risk starts to increase again at the beginning of October and slowly decreases until April for reasons yet unknown. Note that the distribution in 2009 has a larger median than that in 2006 but also shows less extreme peaks.

Birth Weight Prediction

Recent advances in neonatal medicine have lowered the threshold of survival to a gestational age of 23–24 weeks and to a birth weight of approximately 500 g. As neonatal risks of morbidity and mortality are highest in the lowest weight range, diagnostic assessment of the small foetus needs to be as precise as possible. [Schild, Maringa, Siemer, Meurer, Hart, Goecke, Schmid, Hothorn, and Hansmann \(2008\)](#) focused on this high-risk group of small foetuses (≤ 1600 g) and proposed a formula for estimating birth weight based on ultrasound imaging performed within seven days before delivery. In addition to predicting the expected birth weight given four standard 2D ultrasound parameters (HC: head circumference, FE: femur length, BPD: biparietal diameter, and AC: transverse diameter and circumference of the foetal abdomen) and three additional 3D ultrasound parameters (UA: upper arm volume, FEM: thigh volume, and ABDO: abdominal volume) $\mathbf{X} \in \mathbb{R}^7$, we aim at assessing the uncertainty of

this prediction. The data on 150 predominantly Caucasian women, collected in a prospective cohort study at the universities in Bonn and Erlangen, Germany, analysed by [Schild *et al.* \(2008\)](#), were utilised to derive 80% prediction intervals for birth weight (BW).

We begin with the linear model estimated by [Schild *et al.* \(2008\)](#)

$$\begin{aligned} \text{BW}_{\mathbf{x}} = & 656.41 + 1.832 \times \text{ABDO} + 31.198 \times \text{HC} + 5.779 \times \text{FEM} + \\ & 73.521 \times \text{FL} + 8.301 \times \text{AC} - 449.886 \times \text{BPD} + 32.534 \times \text{BPD}^2 + \\ & 77.465 \times \Phi^{-1}(U), \end{aligned}$$

and the classical prediction interval for a foetus with ultrasound parameters \mathbf{x} is then the symmetric interval around the estimated conditional mean $\hat{\mathbb{E}}(\text{BW}|\mathbf{X} = \mathbf{x})$, whose width is given by $2 \times t_{150-8,0.9} \times 77.465 \times \sqrt{1 + \text{Var}(\hat{\mathbb{E}}(\text{BW}|\mathbf{X} = \mathbf{x}))}$.

The normal assumption can be relaxed by deriving the upper and lower conditional quantiles from two quantile regression models. Linear quantile regression ([Koenker and Bassett 1978](#)) for the conditional 10%, 50% and 90% quantiles assumes that

$$\text{BW}_{\mathbf{x}} = \alpha_{0,\tau} + \mathbf{x}^\top \boldsymbol{\alpha}_\tau + Q_\tau(U), \quad \text{for } \tau = 0.1, \tau = 0.5, \text{ and } \tau = 0.9$$

with $Q_\tau(\tau) = 0$. The corresponding prediction interval for a foetus with ultrasound parameters \mathbf{x} is now $(\hat{\alpha}_{0,0.1} + \mathbf{x}^\top \hat{\boldsymbol{\alpha}}_{0.1}, \hat{\alpha}_{0,0.9} + \mathbf{x}^\top \hat{\boldsymbol{\alpha}}_{0.9})$. A more flexible description of the functional relationship between ultrasound parameters and quantiles is given by the additive quantile regression model ([Koenker, Ng, and Portnoy 1994](#))

$$\text{BW}_{\mathbf{x}} = \alpha_{0,\tau} + \sum_{j=1}^7 r_{j,\tau}(x_j) + Q_\tau(U), \quad \text{for } \tau = 0.1, \tau = 0.5, \text{ and } \tau = 0.9.$$

Here, $r_{j,\tau}$ is a quantile-specific smooth function of the j th ultrasound parameter. Parameter tuning is difficult for these models; we therefore applied a boosting approach to additive quantile regression ([Fenske, Kneib, and Hothorn 2011](#), with early stopping via 25-fold bootstrap). Prediction intervals can now be derived by $(\hat{\alpha}_{0,0.1} + \sum_{j=1}^7 \hat{r}_{j,0.1}(x_j), \hat{\alpha}_{0,0.9} + \sum_{j=1}^7 \hat{r}_{j,0.9}(x_j))$. Note that for either quantile regression model, the prediction interval is based on two separate models: one for $\tau = 0.1$ and one for $\tau = 0.9$.

Finally, we derive prediction intervals from the conditional transformation model

$$\mathbb{P}(\text{BW} \leq v | \mathbf{X} = \mathbf{x}) = \Phi \left(\sum_{j=1}^7 h_j(v|x_j) \right)$$

where, under the assumption of additivity of the transformation function h , each ultrasound parameter may influence the moments of the conditional birth weight distribution. The j th base-learner is the tensor product of B -spline basis functions $\mathbf{b}_0(v)$ for birth weight and $\mathbf{b}_j(x_j)$ are B -spline basis functions for the j th ultrasound parameter. The penalty matrices \mathbf{P}_0 and \mathbf{P}_j penalise second-order differences, and thus all estimates \hat{h}_j will be smooth bivariate tensor product splines of birth weight and the respective ultrasound parameter, with both dimensions being subject to smoothing. The number of boosting iterations was determined by 25-fold bootstrap. From the estimated conditional distribution functions, we compute the τ quantiles

of the birth weight via

$$\hat{Q}(\tau|\mathbf{X} = \mathbf{x}) = \inf \left\{ v : \Phi \left(\sum_{j=1}^7 \hat{h}_j(v|x_j) \right) \geq \tau \right\}$$

and derive the prediction interval as $(\hat{Q}(0.1|\mathbf{X} = \mathbf{x}), \hat{Q}(0.9|\mathbf{X} = \mathbf{x}))$. Note that, unlike in the quantile regression approach, the prediction interval obtained from the conditional transformation model is based on only one model that describes the whole conditional distribution of birth weight.

The observed birth weights ordered with respect to the predicted mean (linear model) or median (quantile regression and conditional transformation model) are depicted in Figure 5. In addition, the respective 80% prediction intervals are visualised by grey areas. It must be noted that, for all models, the prediction intervals are only interpretable for future observations; however, poor coverage for the learning sample also indicates poor coverage for future cases. The prediction intervals obtained from linear quantile regression indicate that the model is confident about its predictions over the whole range of birth weights. This is also the case for the additive quantile regression models for birth weights of approximately 1000 g, but the uncertainty increases for very small and larger fetuses. The intervals obtained from the linear model and the conditional transformation model appear to be similar. For birth weights between 500 and 1400 g, the prediction intervals of the conditional transformation model are symmetric around the median. This might be an indication that the normal assumption by the linear model is not completely unrealistic. The smaller interval widths that can be seen for the linear model are most likely due to the variance estimate in this case ignoring the model choice process that was performed prior to the final model fit by [Schild *et al.* \(2008\)](#). The conditional transformation model takes this variability into account. The results may also be an indication that the assumption of additivity of the transformation function rather than of the regression function (for quantile regression models) might be more appropriate for modelling birth weights.

Beyond Mean Boston Housing Values

The Boston Housing data, first published by [Harrison and Rubinfeld \(1978\)](#) and later corrected and spatially aligned by [Gilley and Pace \(1996\)](#), have become a standard test-bed for variable selection and model choice. Almost exclusively, the 13 explanatory variables have been selected with respect to their influence on the mean or median of the conditional median house value in a certain tract. Assuming a conditional transformation model, we attempt to detect dependencies of higher moments of the conditional median house value from the explanatory variables. We focus on the 12 numeric explanatory variables and ignore the binary variable coding for Charles River boundary in the conditional transformation model

$$\mathbb{P}(\text{MEDV} \leq v | \mathbf{X} = \mathbf{x}) = \Phi \left(\alpha_{\text{tract}} + h_0(v|1) + \sum_{j=1}^{12} h_j(1|x_j) + \sum_{j=1}^{12} h_j(v|x_j) \right).$$

In this model, α_{tract} is a tract-specific, spatial random effect, whose correlation structure is determined by a Markov random field defined by the neighbouring structure of the tracts capturing spatial autocorrelation and heterogeneity (similar as in the example on childhood nutrition in India). The term $h_0(v|1)$ is an unconditional transformation of the median house value,

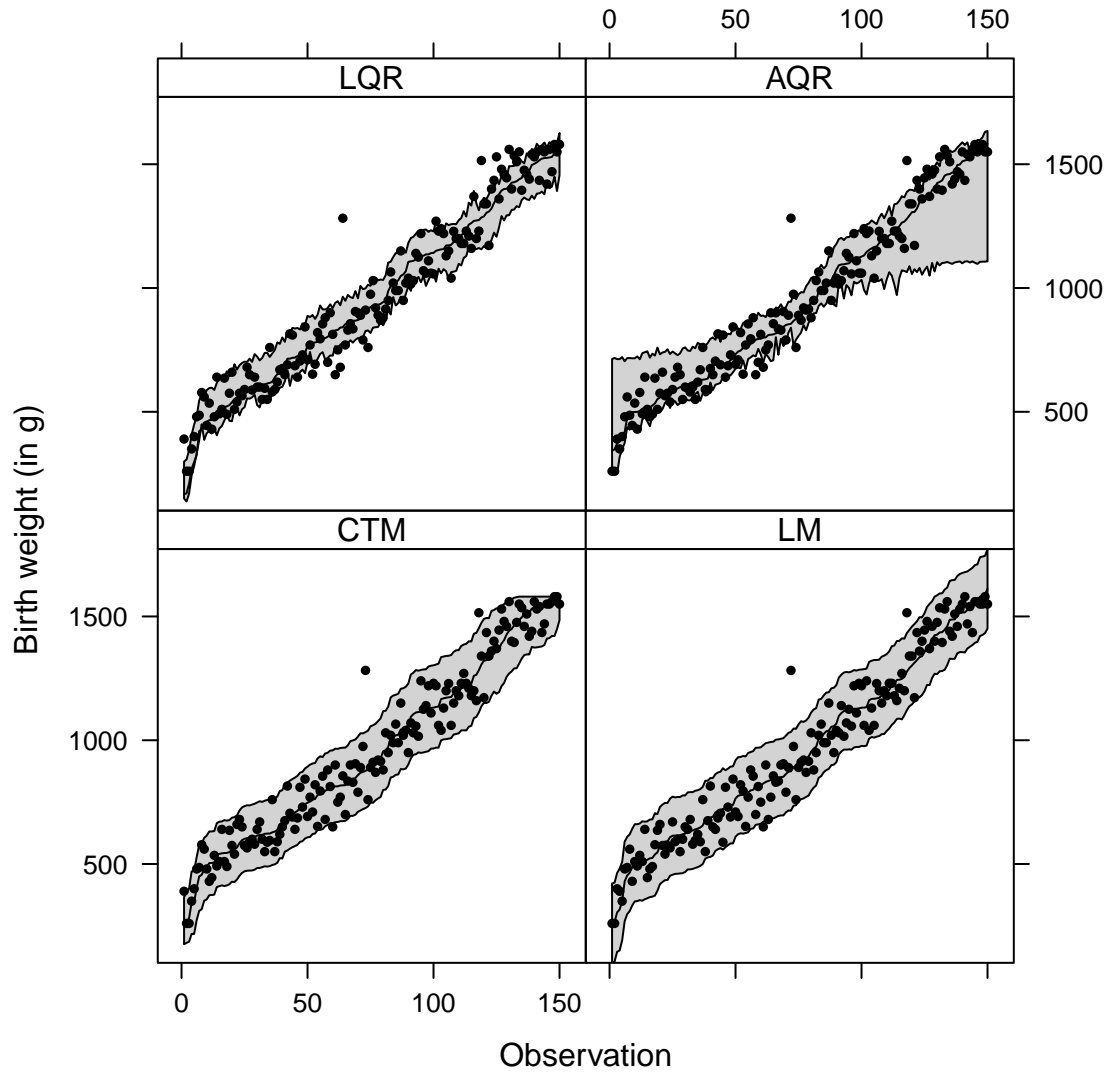


Figure 5: Birth Weight Prediction. Observed birth weights for 150 small fetuses (dots), ordered with respect to the estimated mean or median expected birth weight (central black line). The shaded area represents foetus specific 80% prediction intervals for the linear model (LM), linear quantile regression model (LQR), additive quantile regression model (AQR) and conditional transformation model (CTM).

i.e. this transformation is independent of the explanatory variables. The explanatory variables may influence the mean of the transformed median house value $h_0(\text{MEDV}|1)$ via $h_{\mathbf{x}}(\mathbf{x}) = \sum_{j=1}^{12} h_j(1|x_j)$ only or may also affect higher moments via the interaction terms $\sum_{j=1}^{12} h_j(v|x_j)$. The latter term extends the transformation model $h_0(\text{MEDV}|1) + \sum_{j=1}^{12} h_j(1|x_j)$ to a conditional transformation model. The base-learners for the transformation function $h_0(v|1)$, the effects $h_j(1|x_j)$ and the interaction terms $h_j(v|x_j)$ are constructed based on cubic B -spline basis functions supplemented with second-order difference penalty. More specifically, $\mathbf{b}_j(\mathbf{x})$ and $\mathbf{b}_0(v)$ are both represented in terms of a reparameterisation of the B -spline basis functions that allows separation of the non-linear terms into a constant, a linear effect and the non-linear (orthogonal) deviation from the linear effect, *i.e.*

$$\mathbf{b}_j(\mathbf{x}) = 1 + x_j + \tilde{\mathbf{b}}_j(x_j) \quad \text{and} \quad \mathbf{b}_0(v) = 1 + v + \tilde{\mathbf{b}}_0(v),$$

where $\tilde{\mathbf{b}}_j(x_j)$ and $\tilde{\mathbf{b}}_0(v)$ are the non-linear deviation effects. Taking the tensor product after applying the decomposition yields a decomposition into linear and non-linear main effects of x_j and v as well as linear and non-linear interaction terms (see [Fahrmeir, Kneib, and Lang 2004](#), [Kneib, Hothorn, and Tutz 2009](#), for technical details of this decomposition). The advantage of this expanded parameterisation is that the automatic model choice capabilities of the boosting algorithm allow us to flexibly determine whether linear or non-linear effects are required and whether there actually is an interaction between the transformation function and specific effects of explanatory variables.

Censored observations were dealt with by choosing inverse probability of censoring weights w_i for the empirical risk function (4) derived from the Kaplan-Meier estimate of the censoring distribution. The stability selection procedure ([Meinshausen and Bühlmann 2010](#)) selected three variables that have an influence on the conditional distribution of the median housing value (MEDV), namely per capita crime (CRIM), average numbers of rooms per dwelling (RM), and percentage values of lower status population (LSTAT). After variable selection, we refitted a conditional transformation model of the simpler form

$$\begin{aligned} \mathbb{P}(\text{MEDV} \leq v | \text{CRIM}, \text{RM}, \text{LSTAT}) \\ = \Phi(h_{\text{CRIM}}(v | \text{CRIM}) + h_{\text{RM}}(v | \text{RM}) + h_{\text{LSTAT}}(v | \text{LSTAT})), \end{aligned}$$

where the base-learners are tensor products of B -spline bases. The fitted functions can be conveniently depicted in the observation space. For example, a scatter plot of MEDV and CRIM and a grey-level image of the bivariate function $\hat{h}_{\text{CRIM}}(\text{MEDV} | \text{CRIM})$ can be viewed in the same coordinate system. Similar to the mirrored distribution functions in [Figure 1](#), we show negative absolute values of the fitted functions \hat{h} for easier interpretation.

[Figure 6](#) indicates that the percentage values of lower status population (LSTAT) lead to smaller values of the median housing value at almost constant variance. However, the conditional distribution will be skewed towards higher MEDV values. For tracts with small average numbers of rooms per dwelling (RM), the median housing value is small and increases with increasing numbers of rooms. The same applies to the variability, since the estimated function $\hat{h}_{\text{RM}}(\text{MEDV} | \text{RM})$ shows more spread for larger values of RM. Per capita crime seems to have an effect on variability and skewness, since for larger crime values, the distribution will be heavily skewed and less variable than small per capita crime values. However, compared to the other two variables, the influence is only of marginal value due to small absolute contributions of this model term to the full model.

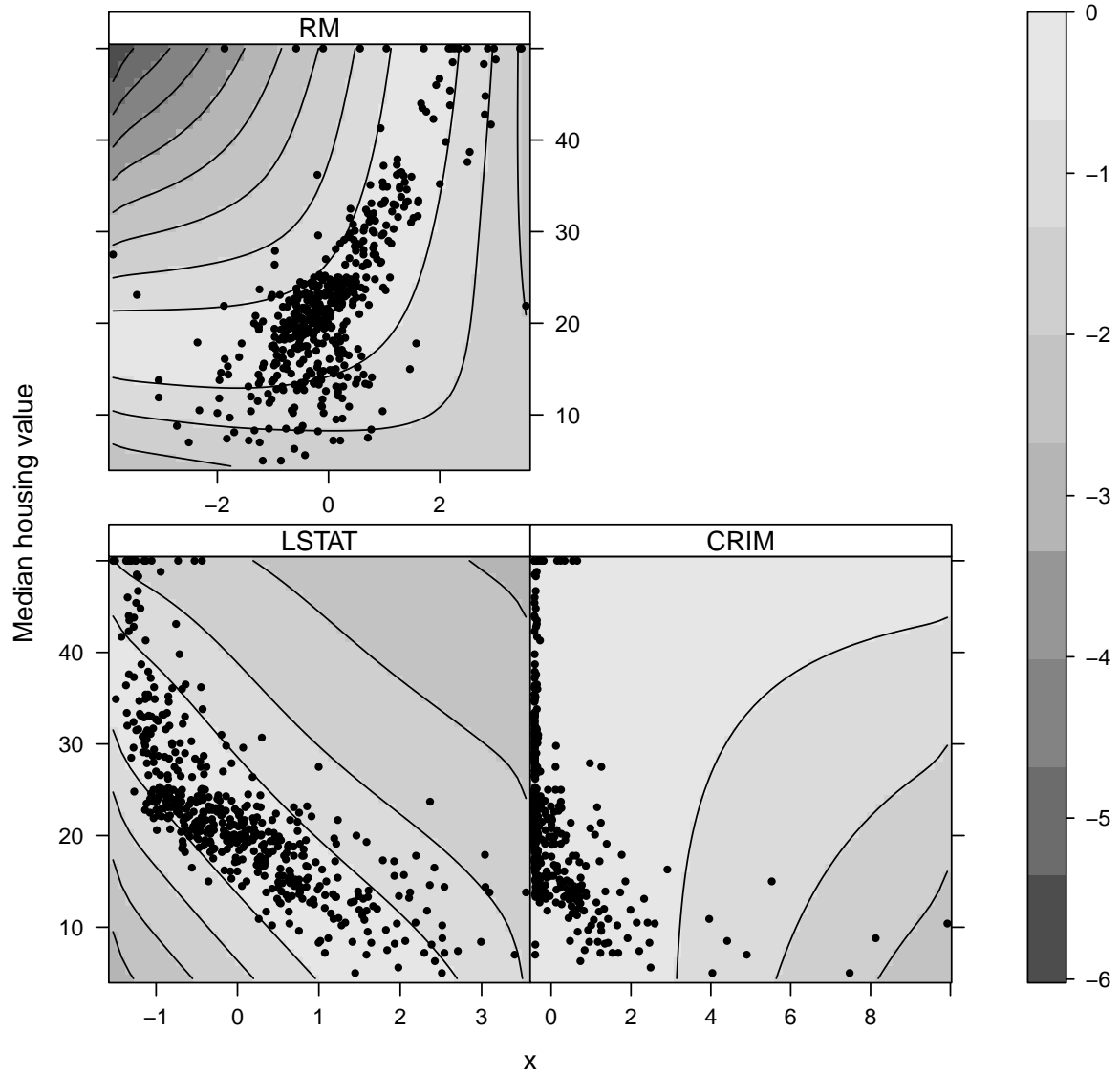


Figure 6: Beyond Mean Boston Housing Values. Conditional transformation model for the three selected variables per capita crime (CRIM), average numbers of rooms per dwelling (RM) and percentage values of lower status population (LSTAT). Each panel depicts the data as scatter plots along with the corresponding negative absolute values of the estimated transformation function at the probit scale. The explanatory variables were standardised prior to analysis.

7. Empirical Evaluation

We shall compare the empirical performance of conditional transformation models fitted by means of the proposed boosting algorithm to two competitors. Conditional transformation models are semiparametric models in the sense that we assume a certain distribution for the transformed responses and additivity of the model terms on the scale of the corresponding quantile function. Therefore, it is natural to compare the goodness of the estimated conditional distribution functions to a fully parametric approach and a non-parametric estimation technique.

For the sake of simplicity, we study a model in which two explanatory variables influence both the conditional expectation and the conditional variance of a normal distributed response Y . The error term $\Phi^{-1}(U)$ is standard normal, and, to obtain normal responses, we restrict the possible transformations to linear functions:

$$\begin{aligned}\Phi^{-1}(U) &= h(Y_{\mathbf{x}}|\mathbf{x}) = \sum_j h_j(Y_{\mathbf{x}}|\mathbf{x}) = \sum_j b_j(\mathbf{x})Y_{\mathbf{x}} - a_j(\mathbf{x}) \\ &= Y_{\mathbf{x}} \sum_j b_j(\mathbf{x}) - \sum_j a_j(\mathbf{x}) \\ \Leftrightarrow Y_{\mathbf{x}} &= \frac{\Phi^{-1}(U) + \sum_j a_j(\mathbf{x})}{\sum_j b_j(\mathbf{x})} \sim \mathcal{N}\left(\frac{\sum_j a_j(\mathbf{x})}{\sum_j b_j(\mathbf{x})}, \left(\sum_j b_j(\mathbf{x})\right)^{-2}\right).\end{aligned}$$

Although the partial transformation functions are linear in $Y_{\mathbf{x}}$, the expectation and variance depend on the explanatory variables in a non-linear way. The choices $X_1 \sim \mathcal{U}[0, 1]$, $X_2 \sim \mathcal{U}[-2, 2]$, $a_1(\mathbf{x}) = 0$, $a_2(\mathbf{x}) = x_2$, and $b_1(\mathbf{x}) = x_1$, $b_2(\mathbf{x}) = 0.5$ lead to the heteroscedastic varying coefficient model

$$Y_{\mathbf{x}} = \frac{1}{x_1 + 0.5}x_2 + \frac{1}{x_1 + 0.5}\Phi^{-1}(U), \quad (12)$$

where the variance of $Y_{\mathbf{x}}$ ranges between 0.44 and 4 depending on X_1 . This model can be fitted in the GAMLSS framework under the assumptions that the expectation 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 use a kernel estimator for estimating the conditional distribution function of $Y_{\mathbf{x}}$ as a function of the two explanatory variables.

The conditional transformation model

$$\mathbb{P}(Y \leq v | X_1 = x_1, X_2 = x_2) = \Phi(h(v|x_1, x_2)) = \Phi(h_1(v|x_1) + h_2(v|x_2))$$

is a semiparametric compromise between these two extremes. The error distribution is assumed to be standard normal and additivity of the transformation function h is also part of the model specification. The base-learners are tensor products of B -spline basis functions $\mathbf{b}_0(v)$ for Y and B -spline basis functions for X_1 and X_2 , respectively. The penalty matrices \mathbf{P}_0 , \mathbf{P}_1 and \mathbf{P}_2 penalise second-order differences, and thus \hat{h}_j will be smooth bivariate tensor product splines of the response and explanatory variables X_1 and X_2 . Smoothing takes place in both dimensions.

For all three approaches, we obtain estimates of $\mathbb{P}(Y \leq v | X_1 = x_1, X_2 = x_2)$ over a grid on x_1, x_2 and compute the mean absolute deviation (MAD) of the true and estimated probabilities

$$\text{MAD}(x_1, x_2) = \frac{1}{n} \sum_v |\mathbb{P}(Y \leq v | X_1 = x_1, X_2 = x_2) - \hat{\mathbb{P}}(Y \leq v | X_1 = x_1, X_2 = x_2)|$$

for each pair of x_1 and x_2 . Then, the minimum, the median, and the maximum of the MAD values over this grid are computed as summary statistics. This procedure was repeated for 100 random samples of size $N = 200$ drawn from model (12). Cross-validation was used to determine the bandwidths for the kernel-based methods, for details see Hayfield and Racine (2008). The boosting-based estimation of GAMLSS models (Mayr, Fenske, Hofner, Kneib, and Schmid 2012a) turned out to be more stable than the reference implementation (package `gamlss`, Stasinopoulos, Rigby, and Akantziliotou 2011), and we therefore fitted the GAMLSS models by the dedicated boosting algorithm. For GAMLSS and conditional transformation models fitted by boosting, the number of boosting iterations was determined via sample-splitting. To investigate the stability of the three procedures under non-informative explanatory variables, we added $p = 1, \dots, 5$ uniformly distributed variables without association to the response to the data and included them as potential explanatory variables in the three models. The case $p = 0$ corresponds to model (12).

Figure 7 shows the empirical distributions of the minimum, median and maximum MAD for the three competitors. For $p = 0$, GAMLSS and conditional transformation models perform on par with respect to the median MAD, although GAMLSS shows a somewhat larger variability. The median MAD is slightly smaller than 0.02 for both procedures, which indicates that the true conditional distribution function can be fitted precisely. The maximal MAD is smallest for conditional transformation models and can be quite large for GAMLSS. In contrast, for some configurations of the explanatory variables, GAMLSS seems to offer better estimates with respect to the minimal MAD. The kernel estimator leads to the largest median MAD values but seems more robust than GAMLSS with respect to the maximal MAD. These results are remarkably robust in the presence of up to five non-informative explanatory variables, although of course the MAD increases with p .

The general theme that GAMLSS on average performs as well as conditional transformation models in the special case of model (12) but is associated with a larger variability might be explained by the independent estimation of the functions for the expectation and variance, *i.e.* GAMLSS does not “know” that the varying coefficient term and the variance term are actually the same. The inferior performance of the kernel estimator might be explained by the technical difficulties associated with bandwidth choice. The tuning parameters for the two boosting approaches are easier to choose. Our general impression is that the kernel-estimated conditional distribution functions are more erratic than the smooth functions obtained with boosting for conditional transformation models (analysis of simulation data not shown here). Since conditional transformation models are also an alternative to quantile regression models, it would be interesting to compare both approaches. At this point, it is important to recall that the two models assume additivity of the effects of X_1 and X_2 , however, on different scales as explained in Section 2. Consequently, the heteroscedastic varying coefficient model (12) cannot be fitted in a straightforward way using standard linear or additive quantile regression. Though, the estimation problem can be slightly reformulated by describing the τ -quantile of $Y_{\mathbf{x}}$ as the sum of a varying coefficient term $r_1(x_1)x_2$ and a smooth function $r_2(x_1)$. This model, implemented using the boosting approach to additive quantile regression with varying

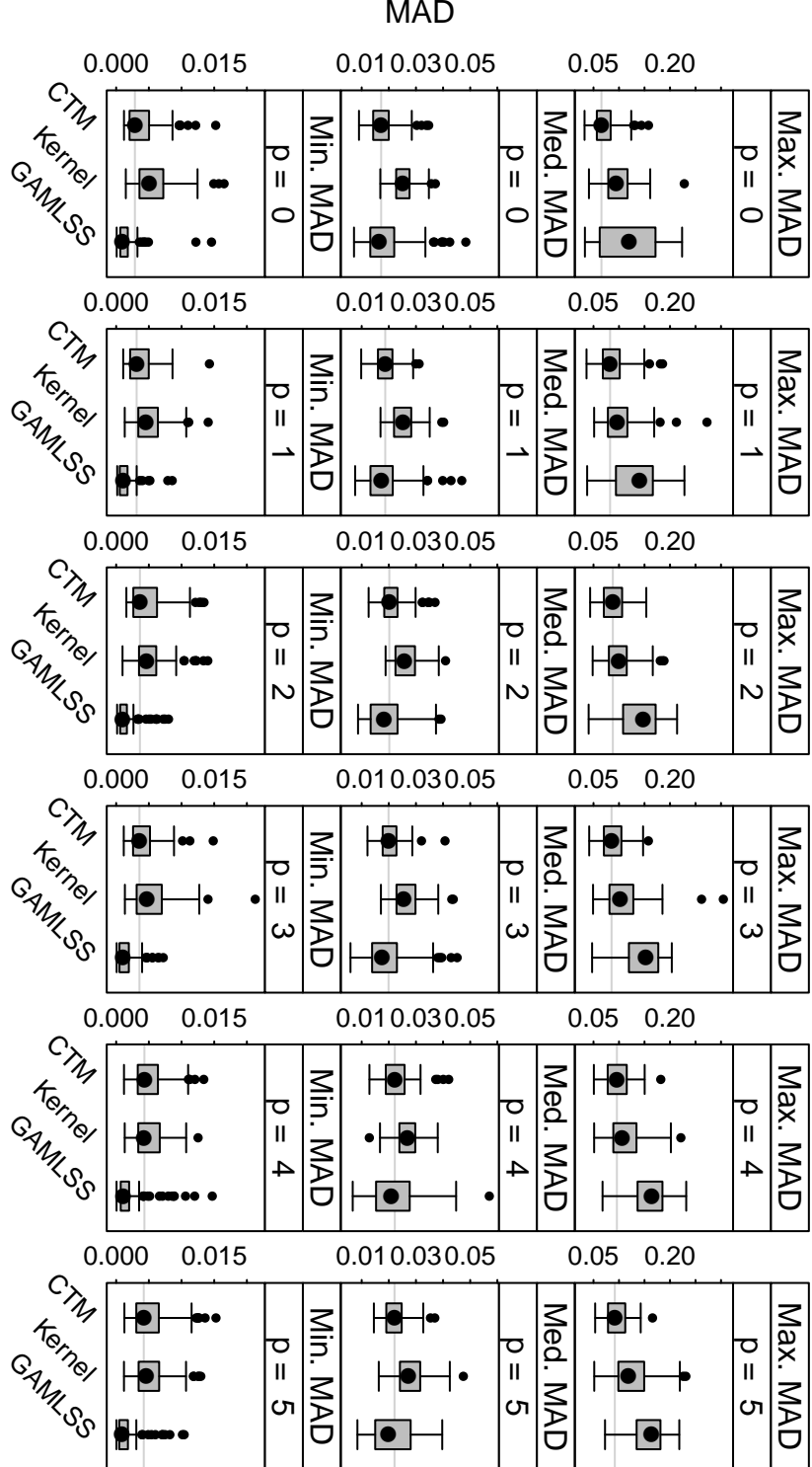


Figure 7: Empirical Evaluation. Minimum, median, and maximum of the mean absolute deviation (MAD) between true and estimated probabilities for conditional transformation models (CTM), non-parametric kernel distribution function estimation (Kernel), and generalised additive models for location, scale, 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 CTM.

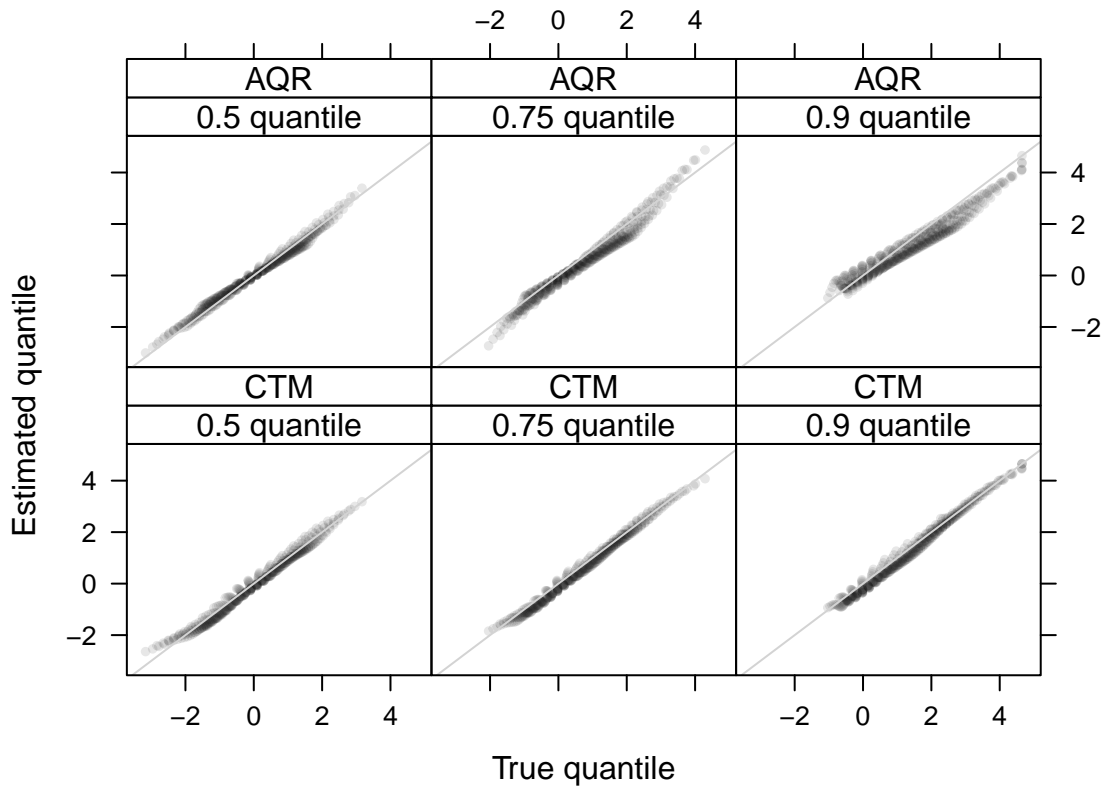


Figure 8: Comparison of conditional transformation models and additive quantile regression. Scatterplot of true versus estimated quantiles obtained from one conditional transformation model (CTM) and from three additive quantile regression (AQR) models fitted to 200 observations drawn from the heteroscedastic varying coefficient model (12).

coefficients introduced by [Fenske et al. \(2011\)](#), allows the estimation of conditional τ -quantiles. We fitted three such quantile regression models (for $\tau = 0.5, 0.75, 0.9$) to a sample of size $N = 200$ from model (12) and determined the optimal number of boosting iterations by the out-of-bootstrap empirical risk of the check function. To give an impression, we compare these estimated τ -quantiles with the corresponding conditional quantiles obtained by inverting the estimated conditional distribution function from a conditional transformation model. Figure 8 displays scatterplots of the true conditional quantiles over a grid of x_1 and x_2 values with the corresponding estimated quantiles derived from one conditional transformation model and the three additive quantile regression models for $\tau = 0.5, 0.75, 0.9$, the latter models including the varying coefficient term. It seems that, in this example, both approaches recover the true quantiles equally good.

8. Discussion

In his book *Quantile Regression*, [Koenker \(2005\)](#) puts transformation models in the “twilight

zone of quantile regression” and suggests that estimating conditional distribution functions by means of transformation models might be an alternative to the direct estimation of conditional quantile functions. We undertook the “worthwhile exercise” (Koenker 2005, Section 8.1.1.) and developed a semiparametric framework for the estimation of conditional distribution functions by conditional transformation models that allows higher moments of the conditional distribution to depend on the explanatory variables.

Because the empirical risk function (4) is equivalent to well-established risk functions for binary data, there are many potentially interesting algorithms that can be used to fit conditional transformation models, although dependent observations have to be dealt with. We chose a component-wise boosting approach mainly because of its divide-and-conquer strategy, which allows a very efficient fitting of base-learners that depend on the response and on one or more explanatory variables at the same time via linear array models. In addition, the algorithm is general, and different base-learners appropriate for the model at hand can be easily implemented as illustrated in Section 6. Also very attractive are the model choice properties, as for example illustrated by means of the Boston Housing data in Section 6. The theoretical and empirical properties investigated in Sections 5 and 7 indicated that the method is applicable to more high-dimensional situations as well. While boosting became popular owing to its success in fitting simple models under challenging circumstances – especially linear or additive models for high-dimensional explanatory variables – the attractiveness of this class of algorithms for fitting challenging models in simple circumstances has been only rarely recognised. Exceptions are Ridgeway (2002) and Sexton and Laake (2012), who study boosting algorithms for fitting density functions. Lu and Li (2008), Schmid and Hothorn (2008) and Schmid, Hothorn, Maloney, Weller, and Potapov (2011) proposed boosting algorithms for transformation models that treat the transformation function h_Y as a nuisance parameter. In the same model framework, Tutz and Groll (2013) propose a likelihood-boosting approach for fitting cumulative and sequential models for ordinal responses.

Boosting algorithms for estimating conditional quantiles by minimising the check function have been introduced by Kriegler and Berk (2010), Fenske *et al.* (2011), and Zheng (2012). The computation of prediction intervals based on pairs of such models is rather straightforward (Mayr, Hothorn, and Fenske 2012b). Our approach to the estimation of the conditional distribution function has the advantage that one model fits the whole distribution, which can then be used to derive arbitrary functionals from. The quantile score representation of the continuous ranked probability score (see Gneiting and Ranjan 2011) might be a basis to develop a boosting technique similar to the one described in this paper for the estimation of full conditional quantile functions. The main difference between transformation and quantile regression models that we have to keep in mind is that additivity is assumed on two different scales. From a practical point of view, diagnostic tools to assess on which of these scales it is more appropriate to assume an additive model would be very important.

The applications presented in Section 6 showed that conditional transformation models are generic, and one can, by choosing appropriate base-learners, fit models that are specific to the problem at hand. An empirical evaluation showed that the estimated conditional distribution functions are on average as good as the estimates obtained from a parametric approach (GAMLSS) that relies on more assumptions. In comparison to non-parametric kernel distribution estimators, conditional transformation models are more adaptable, for example to spatial or temporal data. The performance of the semiparametric models compared to that of the non-parametric competitor was considerably better at the small price of the assumption

of additivity of the transformation function.

It will be interesting to further study conditional transformation models with respect to the following extensions. Instead of making assumptions about the quantile function Q representing the error distribution, it would be possible to fit the corresponding distribution function by techniques introduced for single index models (Tutz and Petry 2012). To reduce the complexity, one could restrict the partial transformation functions h_j to linear functions, *i.e.* only the first two moments are allowed to be influenced by the explanatory variables. Furthermore, when the response $Y = (Y_1, Y_2)$ is bivariate, the bivariate conditional distribution function $\mathbb{P}(Y_1 \leq v_1, Y_2 \leq v_2 | \mathbf{X} = \mathbf{x})$ can be estimated by minimising the risk function

$$\int \int \rho\{((y_1 \leq v_1 \wedge y_2 \leq v_2), \mathbf{x}), h(v_1, v_2 | \mathbf{x})\} d\mu(v_1) d\mu(v_2) d\hat{\mathbb{P}}_{Y, \mathbf{X}}(y_1, y_2, \mathbf{x})$$

in h . It might also be interesting to allow only certain moments to depend on certain explanatory variables. For example, a partially conditional transformation model of the form

$$h(v | \mathbf{x}) = \sum_{j=1}^{J_1} h_j(v | \mathbf{x}) + \sum_{j=J_1+1}^J h_j(1 | \mathbf{x})$$

describes the conditional expectation $\sum_{j=J_1+1}^J h_j(1 | \mathbf{x})$ of a transformation $\sum_{j=1}^{J_1} h_j(Y_{\mathbf{x}} | \mathbf{x})$.

Researchers pay most attention to the conditional distribution function in applications where the response is a survival time or has been censored owing to other reasons. In survival analysis, it is common to deal with the conditional survivor function $S(Y_{\mathbf{x}} | \mathbf{x}) = 1 - F(Y_{\mathbf{x}} | \mathbf{x})$ for survival time Y . In the seemingly simpler situation of an uncensored continuous response, many data analysts focus on the conditional mean and do not bother with the conditional distribution function at all. Most of the literature on transformation models therefore deals with the censored case. The easiest way to deal with right-censoring is to minimise the risk function weighted by inverse probability of censoring weights, as in the generalised estimating equations approach by Cheng *et al.* (1995), where parameters of a linear transformation model are estimated by the root of a V-statistic defined by all binary indicators $I(Y_i \geq Y_i), i, i = 1, \dots, N$. Accelerated failure time models fitted by boosting of an inverse probability of censoring weighted risk function have been described by Hothorn, Bühlmann, Dudoit, Molinaro, and van der Laan (2006), and future research awaits the investigation of the performance of conditional transformation models under censoring.

Computational Details

Conditional transformation models were fitted using an implementation of component-wise boosting in package **mboost** (version 2.1-2, Hothorn, Bühlmann, Kneib, Schmid, and Hofner 2011). Package **gamboostLSS** (version 1.0-3, Hofner, Mayr, Fenske, and Schmid 2011b) was used to fit GAMLSS models and kernel distribution estimation was performed using package **np** (version 0.40-13, Hayfield and Racine 2011). Linear quantile regression was computed using package **quantreg** (version 4.79, Koenker 2011). All computations were performed using R version 2.13.2 (R Development Core Team 2011).

Throughout Section 6, we used the loss function defined by ρ_{bin} and modelled non-linear functions by cubic B -spline bases with 20 equi-distant knots. For further computational

details we refer the reader to the R code that implements the analyses presented in Sections 6 and 7, which is available in an experimental R package **ctm** at <http://R-forge.R-project.org/projects/ctm>. The results presented in this paper can be reproduced using this package, except for the birth weight data, which are not publically available.

Acknowledgements

We would like to thank Gareth O. Roberts, an anonymous associate editor and referee for their careful review of an initial version of this paper. Andreas Mayr implemented the `Normal()` family for **gamboostLSS** on our request, Achim Zeileis helped finding the right colors for Figure 2 and Ronald Schild provided us with the birth weight data. We are indepted to Paul Eilers, Tilman Gneiting and Roger Koenker for their comments on a draft version and thank Karen A. Brune for improving the language. Financial support by Deutsche Forschungsgemeinschaft (grant HO 3242/4-1) is gratefully acknowledged.

References

- Bollerslev T (1986). “Generalized Autoregressive Conditional Heteroskedasticity.” *Journal of Econometrics*, **31**(3), 307–327.
- Bühlmann P (2006). “Boosting for High-dimensional Linear Models.” *The Annals of Statistics*, **34**(2), 559–583.
- Bühlmann P, Hothorn T (2007). “Boosting Algorithms: Regularization, Prediction and Model Fitting.” *Statistical Science*, **22**(4), 477–505. With discussion.
- Bühlmann P, van de Geer S (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Verlag, Berlin, Heidelberg, Germany.
- Bühlmann P, Yu B (2003). “Boosting with the L_2 Loss: Regression and Classification.” *Journal of the American Statistical Association*, **98**, 324–339.
- Carroll RJ, Ruppert D (1982). “Robust Estimation in Heterscedastic Linear Models.” *The Annals of Statistics*, **10**(2), 429–441.
- Chen K, Müller HG (2012). “Conditional Quantile Analysis When Covariates Are Functions, With Application to Growth Data.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(1), 67–89.
- Chen K, Tong X (2010). “Varying Coefficient Transformation Models With Censored Data.” *Biometrika*, **97**(4), 969–976.
- Cheng G, Wang X (2011). “Semiparametric Additive Transformation Model Under Current Status Data.” *Electronic Journal of Statistics*, **5**, 1735–1764.
- Cheng SC, Wei LJ, Ying Z (1995). “Analysis of Transformation Models With Censored Data.” *Biometrika*, **82**(4), 835–845.

- Currie ID, Durban M, Eilers PHC (2006). “Generalized Linear Array Models With Applications to Multidimensional Smoothing.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(2), 259–280.
- Dette H, Volgushev S (2008). “Non-crossing Non-parametric Estimates of Quantile Curves.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(3), 609–627.
- Doksum KA, Gasko M (1990). “On a Correspondence Between Models in Binary Regression Analysis and in Survival Analysis.” *International Statistical Review*, **58**(3), 243–252.
- Eilers PHC (2005). “Unimodal Smoothing.” *Journal of Chemometrics*, **19**(5-7), 317–328.
- Fahrmeir L, Kneib T, Lang S (2004). “Penalized Structured Additive Regression for Space-Time Data: A Bayesian Perspective.” *Statistica Sinica*, **14**(3), 731–761.
- Fenske N, Kneib T, Hothorn T (2011). “Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression.” *Journal of the American Statistical Association*, **106**(494), 494–510.
- Fredriks AM, van Buuren S, Burgmeijer RJF, Meulmeester JF, Beuker RJ, Brugman E, Roede MJ, Verloove-Vanhorick SP, Wit J (2000). “Continuing Positive Secular Growth Change in The Netherlands 1955–1997.” *Pediatric Research*, **47**(3), 316–323.
- Friedman HH, Friedman LW, Amoo T (2002). “Using Humor in the Introductory Statistics Course.” *Journal of Statistics Education*, **10**(3). URL <http://www.amstat.org/publications/jse/v10n3/friedman.html>.
- Gilchrist W (2008). “Regression Revisited.” *International Statistical Review*, **76**(3), 401–418.
- Gilley OW, Pace RK (1996). “On the Harrison and Rubinfeld Data.” *Journal of Environmental Economics and Management*, **31**, 403–405.
- Gneiting T, Raftery AE (2007). “Strictly Proper Scoring Rules, Prediction, and Estimation.” *Journal of the American Statistical Association*, **102**(477), 359–378.
- Gneiting T, Raftery AE, Westveld III AH, Goldman T (2005). “Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation.” *Monthly Weather Review*, **133**(5), 1098–1118.
- Gneiting T, Ranjan R (2011). “Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules.” *Journal of Business & Economic Statistics*, **29**(3), 411–422.
- Hall P, Müller HG (2003). “Order-preserving Nonparametric Regression, With Applications to Conditional Distribution and Quantile Function Estimation.” *Journal of the American Statistical Association*, **98**(463), 598–608.
- Hall P, Wolff RCL, Yao Q (1999). “Methods for Estimating a Conditional Distribution Function.” *Journal of the American Statistical Association*, **94**(445), 154–163.
- Harrison D, Rubinfeld DL (1978). “Hedonic Prices and the Demand for Clean Air.” *Journal of Environmental Economics and Management*, **5**, 81–102.

- Hayfield T, Racine JS (2008). “Nonparametric Econometrics: The np Package.” *Journal of Statistical Software*, **27**(5), 1–32. URL <http://www.jstatsoft.org/v27/i05>.
- Hayfield T, Racine JS (2011). *np: Nonparametric Kernel Smoothing Methods for Mixed Data Types*. R package version 0.40-12, URL <http://CRAN.R-project.org/package=np>.
- He X (1997). “Quantile Curves Without Crossing.” *The American Statistician*, **51**(2), 186–192.
- Herberich E, Sikorski J, Hothorn T (2010). “A Robust Procedure for Comparing Multiple Means Under Heteroscedasticity in Unbalanced Designs.” *PLoS ONE*, **5**(4), e9788.
- Hofner B, Hothorn T, Kneib T, Schmid M (2011a). “A Framework for Unbiased Model Selection Based on Boosting.” *Journal of Computational and Graphical Statistics*, **20**(4), 956–971.
- Hofner B, Mayr A, Fenske N, Schmid M (2011b). *gamboostLSS: Boosting Methods for GAMLSS Models*. R package version 1.0-3/r39, URL <http://R-Forge.R-project.org/projects/gamboostlss/>.
- Hofner B, Müller J, Hothorn T (2011c). “Monotonicity-Constrained Species Distribution Models.” *Ecology*, **92**(10), 1895–1901.
- Hothorn T, Brandl R, Müller J (2012). “Large-scale Model-based Assessment of Deer-vehicle Collision Risk.” *PLoS ONE*, **7**(2), e29510.
- Hothorn T, Bühlmann P, Dudoit S, Molinaro A, van der Laan MJ (2006). “Survival Ensembles.” *Biostatistics*, **7**(3), 355–373.
- Hothorn T, Bühlmann P, Kneib T, Schmid M, Hofner B (2011). *mboost: Model-Based Boosting*. R package version 2.1-1, URL <http://CRAN.R-project.org/package=mboost>.
- Hothorn T, Kneib T, Bühlmann P (2013). “Conditional Transformation Models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Accepted 2012-11-27.
- Kneib T, Hothorn T, Tutz G (2009). “Variable Selection and Model Choice in Geoadditive Regression Models.” *Biometrics*, **65**(2), 626–634.
- Koenker R (2005). *Quantile Regression*. Economic Society Monographs. Cambridge University Press, New York.
- Koenker R (2011). *quantreg: Quantile Regression*. R package version 4.76, URL <http://CRAN.R-project.org/package=quantreg>.
- Koenker R, Bassett G (1978). “Regression Quantiles.” *Econometrica*, **46**(1), 33–50.
- Koenker R, Ng P, Portnoy S (1994). “Quantile Smoothing Splines.” *Biometrika*, **81**(4), 673–680.
- Kriegler B, Berk R (2010). “Small Area Estimation of the Homeless in Los Angeles: An Application of Cost-sensitive Stochastic Gradient Boosting.” *The Annals of Applied Statistics*, **4**(3), 1234–1255.

- Li Q, Racine JS (2008). “Nonparametric Estimation of Conditional CDF and Quantile Functions With Mixed Categorical and Continuous Data.” *Journal of Business and Economic Statistics*, **26**(4), 423–434.
- Lu W, Li L (2008). “Boosting Method for Nonlinear Transformation Models With Censored Survival Data.” *Biostatistics*, **9**(4), 658–667.
- Lu W, Zhang HH (2010). “On Estimation of Partially Linear Transformation Models.” *Journal of the American Statistical Association*, **105**(490), 683–691.
- Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012a). “GAMLSS for High-dimensional Data—A Flexible Approach Based on Boosting.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**(3), 403–427.
- Mayr A, Hothorn T, Fenske N (2012b). “Prediction Intervals for Future BMI Values of Individual Children—A Non-parametric Approach by Quantile Boosting.” *BMC Medical Research Methodology*, **12**, 6.
- Meinshausen N, Bühlmann P (2010). “Stability Selection.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 417–473.
- NFHS (2000). “National Family Health Survey (NFHS-2), 1998–1999: India.” *International Institute for Population Sciences (IIPS) and ORC Macro*. URL <http://www.measuredhs.com/pubs>.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Ridgeway G (2002). “Looking for Lumps: Boosting and Bagging for Density Estimation.” *Computational Statistics & Data Analysis*, **38**(4), 379–392.
- Rigby RA, Stasinopoulos DM (2005). “Generalized Additive Models for Location, Scale and Shape.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(3), 507–554.
- Schemper M, Henderson R (2000). “Predictive Accuracy and Explained Variation in Cox Regression.” *Biometrics*, **56**(1), 249–255.
- Schild RL, Maringa M, Siemer J, Meurer B, Hart N, Goecke TW, Schmid M, Hothorn T, Hansmann ME (2008). “Weight Estimation By Three-dimensional Ultrasound Imaging in the Small Fetus.” *Ultrasound in Obstetrics and Gynecology*, **32**(2), 168–175.
- Schmid M, Hothorn T (2008). “Flexible Boosting of Accelerated Failure Time Models.” *BMC Bioinformatics*, **9**, 269.
- Schmid M, Hothorn T, Maloney KO, Weller DE, Potapov S (2011). “Geoadditive Regression Modeling of Stream Biological Condition.” *Environmental and Ecological Statistics*, **18**(4), 709–733.
- Schnabel SK, Eilers PHC (2012). “Simultaneous Estimation of Quantile Curves Using Quantile Sheets.” *AStA Advances in Statistical Analysis*. doi:10.1007/s10182-012-0198-1. Accepted 2012-06-21.

- Sexton J, Laake P (2012). “Boosted Coefficient Models.” *Statistics and Computing*, **22**(4), 867–876.
- Shen X (1998). “Proportional Odds Regression and Sieve Maximum Likelihood Estimation.” *Biometrika*, **85**(1), 165–177.
- Sikorski J, Brambilla E, Kroppenstedt RM, Tindall BJ (2008a). “The Temperature Adaptive Fatty Acid Content in *Bacillus simplex* Strains from “Evolution Canyon”, Israel.” *Microbiology*, **154**(8), 2416–2426.
- Sikorski J, Nevo E (2005). “Adaptation and Incipient Sympatric Speciation of *Bacillus simplex* under Microclimatic Contrast at “Evolution Canyons” I and II, Israel.” *Proceedings of the National Academy of Sciences, U.S.A.*, **102**(44), 15924–15929.
- Sikorski J, Pukall R, Stackebrandt E (2008b). “Carbon Source Utilization Patterns of *Bacillus simplex* Ecotypes Do Not Reflect Their Adaptation to Ecologically Divergent Slopes in “Evolution Canyon”, Israel.” *FEMS Microbiology Ecology*, **66**(1), 38–44.
- Stasinopoulos DM, Rigby RA (2007). “Generalized Additive Models for Location Scale and Shape (GAMLSS) in R.” *Journal of Statistical Software*, **23**(7), 1–46. URL <http://www.jstatsoft.org/v23/i07>.
- Stasinopoulos M, Rigby B, Akantziliotou C (2011). *gamlss: Generalized Additive Models for Location Scale and Shape*. R package version 4.1-1, URL <http://CRAN.R-project.org/package=gamlss>.
- Stigler SM (2010). “The Changing History of Robustness.” *The American Statistician*, **64**(4), 277–281.
- Tutz G, Groll A (2013). “Likelihood-based Boosting in Binary and Ordinal Random Effects Models.” *Journal of Computational and Graphical Statistics*. doi:10.1080/10618600.2012.694769. In press.
- Tutz G, Petry S (2012). “Nonparametric Estimation of the Link Function Including Variable Selection.” *Statistics and Computing*, **22**(2), 545–561.
- van de Geer S (2000). *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK.
- Wu C, Tian X, Yu J (2010). “Nonparametric Estimation for Time-varying Transformation Models With Longitudinal Data.” *Journal of Nonparametric Statistics*, **22**(2), 133–147.
- Zeng D, Lin DY, Yin G (2005). “Maximum Likelihood Estimation for the Proportional Odds Model With Random Effects.” *Journal of the American Statistical Association*, **100**(470), 470–483.
- Zheng S (2012). “QBoost: Predicting Quantiles with Boosting for Regression and Binary Classification.” *Expert Systems with Applications*, **39**(2), 1687–1697.

Appendix

Proofs

Proof. Proof of Lemma 1

Convexity: If the loss function ρ is convex in its second argument, so is the loss function ℓ

$$\ell((Y, \mathbf{X}), \alpha h + (1 - \alpha)g) \leq \alpha \ell((Y, \mathbf{X}), h) + (1 - \alpha) \ell((Y, \mathbf{X}), g),$$

with $g(\cdot|\mathbf{x}) : \mathbb{R} \rightarrow \mathbb{R}$ being a monotone increasing transformation function and $\alpha \in [0, 1]$, because of the convexity of ρ and the monotonicity and linearity of the Lebesgue integral.

Population Minimisers: Let f denote the density of F . With iterated expectation we have

$$\begin{aligned} \mathbb{E}_{Y, \mathbf{X}} \ell((Y, \mathbf{X}), h) &= \int \int \int \rho((y \leq v, \mathbf{x}), h(v|\mathbf{x})) d\mu(v) d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) \\ &= \int \int \int \underbrace{\rho((y \leq v, \mathbf{x}), h(v|\mathbf{x}))}_{=: A_{v, \mathbf{x}}(h(v|\mathbf{x}))} d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) d\mu(v) d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) \end{aligned}$$

and the risk is minimal when $A_{v, \mathbf{x}}(h(v|\mathbf{x}))$ is minimal for the scalar $h(v|\mathbf{x})$ for all v and \mathbf{x} , *i.e.* when

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial A_{v, \mathbf{x}}(h(v|\mathbf{x}))}{\partial h(v|\mathbf{x})} \\ &= \int \frac{\partial}{\partial h(v|\mathbf{x})} \rho((y \leq v, \mathbf{x}), h(v|\mathbf{x})) d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\ &\stackrel{\rho=\rho_{\text{sqe}}}{=} \int (I(y \leq v) - F(h(v|\mathbf{x}))) f(h(v|\mathbf{x})) d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\ &= f(h(v|\mathbf{x})) \left\{ \int I(y \leq v) d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) - F(h(v|\mathbf{x})) \right\} \\ &= f(h(v|\mathbf{x})) \{ \mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}) - F(h(v|\mathbf{x})) \} \end{aligned}$$

which for $f(h(v|\mathbf{x})) > 0$ is zero for $h(v|\mathbf{x}) = F^{-1}(\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}))$. Similar, for $\rho = \rho_{\text{bin}}$ the term

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial A_{v, \mathbf{x}}(h(v|\mathbf{x}))}{\partial h(v|\mathbf{x})} \\ &= \int - \left\{ \frac{I(y \leq v)}{F(h(v|\mathbf{x}))} f(h(v|\mathbf{x})) - \frac{1 - I(y \leq v)}{1 - F(h(v|\mathbf{x}))} f(h(v|\mathbf{x})) \right\} d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\ &= f(h(v|\mathbf{x})) \left\{ \frac{\int 1 - I(y \leq v) d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y)}{1 - F(h(v|\mathbf{x}))} - \frac{\int I(y \leq v) d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y)}{F(h(v|\mathbf{x}))} \right\} \\ &= f(h(v|\mathbf{x})) \left\{ \frac{1 - \mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x})}{1 - F(h(v|\mathbf{x}))} - \frac{\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x})}{F(h(v|\mathbf{x}))} \right\} \end{aligned}$$

is zero for $h(v|\mathbf{x}) = F^{-1}(\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}))$ when $f(h(v|\mathbf{x})) > 0$.

For the absolute error, note that

$$\rho_{\text{abe}}((Y \leq v, \mathbf{X}), h(v|\mathbf{X})) = I(Y \leq v) \{1 - F(h(v|\mathbf{X}))\} + \{1 - I(Y \leq v)\} F(h(v|\mathbf{X}))$$

and thus

$$\begin{aligned}
A_{v,\mathbf{x}}(h(v|\mathbf{x})) &= \int \rho_{\text{abe}}((y \leq v, \mathbf{x}), h(v|\mathbf{x})) d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\
&= \int I(Y \leq v) \{1 - F(h(v|\mathbf{X}))\} + \{1 - I(Y \leq v)\} F(h(v|\mathbf{X})) d\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\
&= \{1 - F(h(v|\mathbf{X}))\} \mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}) + F(h(v|\mathbf{X})) \{1 - \mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x})\}
\end{aligned}$$

This expression attains its minimal value of $\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x})$ for $\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}) \leq 0.5$ when $F(h(v|\mathbf{X})) = 0$. For $\mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x}) > 0.5$, the minimum $1 - \mathbb{P}(Y \leq v | \mathbf{X} = \mathbf{x})$ is attained when $F(h(v|\mathbf{X})) = 1$. Thus, absolute error will lead to too extreme estimated values of h and corresponding conditional distribution functions. \square

Proof. Proof of Theorem 1

We use a modified argument of a proof presented in Section 12.8.2. in [Bühlmann and van de Geer \(2011\)](#). Formally, we can write

$$\begin{aligned}
I(Y_i \leq v_i) &= h_{\gamma_{0,N}}(v_i | \mathbf{X}_i) + \varepsilon_{ii}, \\
\varepsilon_{ii} &= I(Y_i \leq v_i) - h_{\gamma_{0,N}}(v_i | \mathbf{X}_i) \quad (i = 1, \dots, N, \quad i = 1, \dots, n).
\end{aligned}$$

The errors ε_{ii} have reasonable properties, as discussed in (13) below.

There are two issues that need to be addressed. First, we define the inner products of functions h and g

$$(h, g)_{n,N,\mathbb{E}} = n^{-1} \sum_{i=1}^n \mathbb{E}[h(v_i | \mathbf{X}) g(v_i | \mathbf{X})]$$

and

$$(h, g)_{n,N} = n^{-1} N^{-1} \sum_{i=1}^N \sum_{i=1}^n h(v_i | \mathbf{X}_i) g(v_i | \mathbf{X}_i).$$

The proof in [Bühlmann and van de Geer \(2011\)](#) can then be used with the scalar product $(h, g)_{n,N}$.

Secondly, for controlling the probabilistic part of the proof, we need to show that the analogue of formula (12.26) in [Bühlmann and van de Geer \(2011\)](#) holds. This translates to deriving a bound for

$$\max_{j,k_0,k_1} (b_{0,k_0} b_{j,k_1}, \varepsilon)_{n,N} = \max_{j,k_0,k_1} (nN)^{-1} \sum_{i=1}^N \sum_{i=1}^n b_{0,k_0}(v_i) b_{j,k_1}(X_j) \varepsilon_{ii}.$$

Because $h_{\gamma_{0,N}}$ is the projection of $I(Y \leq v_i)$ ($i = 1, \dots, n$) onto the basis functions $b_{0,k_0}(v_i) b_{j,k_1}(X_j)$ ($i = 1, \dots, n$) with respect to the $\|\cdot\|_{n,N,\mathbb{E}}$ -norm (see (7)), and owing to the definition of ε_{ii} , we have

$$n^{-1} \sum_{i=1}^n \mathbb{E}[\varepsilon_{ii} b_{0,k_0}(v_i) b_{j,k_1}(X_j)] = 0 \quad \forall j, k_0, k_1. \quad (13)$$

Therefore,

$$\begin{aligned}
(nN)^{-1} \sum_{i=1}^N \sum_{i=1}^n b_{0,k_0}(v_i) b_{j,k_1}(X_j) \varepsilon_{ii} &= N^{-1} \sum_{i=1}^N Z_i(j, k_0, k_1), \\
\mathbb{E}[Z_i(j, k_0, k_1)] &= 0.
\end{aligned}$$

Furthermore, owing to the boundedness assumption in (A1), $\|Z_i(j, k_0, k_1)\|_\infty \leq C_1$ for some constant $C_1 < \infty \forall i, j, k_0, k_1$. Applying Hoeffding's inequality, for independent (but not necessarily identically distributed) random variables (van de Geer 2000, Lem. 3.5) and using the union bound, we obtain

$$\max_{j, k_0, k_1} (b_{0, k_0} b_{j, k_1}, \varepsilon)_{n, N} = O_P \left(\sqrt{\log(J_N K_{0, N} K_{1, N}) / N} \right).$$

This, together with the proof from Section 12.8.2 in Bühlmann and van de Geer (2011) completes the proof of Theorem 1. \square

Gradients

Gradients for different loss functions ρ and arbitrary absolute continuous distribution functions F with density function f :

$$\begin{aligned} U_{ii} & \stackrel{\rho=\rho_{\text{bin}}}{=} \left\{ \frac{I(Y_i \leq v_i)}{F(\hat{h}_{ii}^{[m]})} - \frac{1 - I(Y_i \leq v_i)}{1 + F(\hat{h}_{ii}^{[m]})} \right\} f(\hat{h}_{ii}^{[m]}) \\ U_{ii} & \stackrel{\rho=\rho_{\text{sqe}}}{=} \left\{ I(Y_i \leq v_i) - F(\hat{h}_{ii}^{[m]}) \right\} f(\hat{h}_{ii}^{[m]}) \\ U_{ii} & \stackrel{\rho=\rho_{\text{abe}}}{=} \left\{ I(Y_i \leq v_i) - (1 - I(Y_i \leq v_i)) \right\} f(\hat{h}_{ii}^{[m]}) \end{aligned}$$

Affiliation:

Torsten Hothorn

Institut für Statistik

Ludwig-Maximilians-Universität München

Ludwigstraße 33

DE-80539 München, Germany

E-mail: Torsten.Hothorn@R-project.org

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

ISPM, Abteilung Biostatistik

Universität Zürich

Hirschengraben 84

CH-8001 Zürich, Switzerland

Thomas Kneib

Wirtschaftswissenschaftliche Fakultät

Georg-August-Universität Göttingen

Platz der Göttinger Sieben 5

DE-37073 Göttingen, Germany

Peter Bühlmann

Seminar für Statistik

ETH Zürich

CH-8092 Zürich, Switzerland