# Bayesian Conditional Transformation Models

Manuel Carlan[*], Thomas Kneib[†]
Chair of Statistics, University of Göttingen
and
Nadja Klein[‡]
Chair of Statistics and Data Science, Humboldt-Universität zu Berlin

**Abstract**

Recent developments in statistical regression methodology shift away from pure mean regression towards distributional regression models. One important strand thereof is that of conditional transformation models (CTMs). CTMs infer the entire conditional distribution directly by applying a transformation function to the response conditionally on a set of covariates towards a simple log-concave reference distribution. Thereby, CTMs allow not only variance, kurtosis or skewness but the complete conditional distribution to depend on the explanatory variables. We propose a Bayesian notion of conditional transformation models (BCTMs) focusing on exactly observed continuous responses, but also incorporating extensions to randomly censored and discrete responses. Rather than relying on Bernstein polynomials that have been considered in likelihood-based CTMs, we implement a spline-based parametrization for monotonic effects that are supplemented with smoothness priors. Furthermore, we are able to benefit from the Bayesian paradigm via easily obtainable credible intervals and other quantities without relying on large sample approximations. A simulation study demonstrates the competitiveness of our approach against its likelihood-based counterpart but also Bayesian additive models of location, scale and shape and Bayesian quantile regression. Two applications illustrate the versatility of BCTMs in problems involving real world data, again including the comparison with various types of competitors.

*Keywords: Conditional distribution function; distributional regression; Hamiltonian Monte Carlo; monotonicity constraint; penalized splines; No-U-Turn Sampler.*

[*] The work of Manuel Carlan was supported by DFG via the research training group 1644.

[†] Thomas Kneib received financial support from the DFG within the research project KN 922/9-1.

[‡] Nadja Klein gratefully acknowledges funding from the DFG through the Emmy Noether grant KL 3037/1-1. The authors would like to thank the Editor, Associate Editor and two referees for many valuable comments that lead to a significant improvement of our original submission. nadja.klein@hu-berlin.de

# 1 Introduction

Regression is omnipresent in many statistical applications and an ongoing field in recent research on statistical methods. While, in principle, interest always lies in describing the conditional distribution $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ of a response $Y$ given a set of explanatory variables $\boldsymbol{X}$ with observed realisations $\boldsymbol{x}$, most traditional approaches target the conditional expectation $\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$ as the only characteristic of interest (e.g. generalized linear or additive models; Nelder and Wedderburn, 1972; Hastie and Tibshirani, 1990). One way to abolish this often unwarranted simplification are generalized additive models for location scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005) allowing for flexible relationships between the covariates and all parameters of the response distribution via flexible additive predictors. In this framework, the researcher can select from a diverse set of parametric distributions for discrete, continuous, mixed and multivariate response distributions (Klein, Kneib, Lang, Sohn et al., 2015; Klein, Kneib, Klasen and Lang, 2015). However, deciding on a parametric response distribution can quickly become a burden as it imposes strong assumptions on the model if not done with great care. One approach that does not entail a fixed parametric form is quantile regression introduced by Koenker and Bassett (1978). Since a distribution is satisfyingly captured by a dense grid of quantiles, each of these quantiles is modelled linearly or additively through covariates (Horowitz and Lee, 2005). Bayesian versions were developed e.g. by Yu and Moyeed (2001); Waldmann et al. (2013).

In contrast to all approaches mentioned so far, transformation models aim to infer the conditional distribution function directly. In an attempt to draw a bigger picture, we recapitulate a brief history of transformation models, while slightly focusing on Bayesian implementations. For a tour de force that concentrates on the frequentist perspective, see e.g. Hothorn et al. (2014); Möst (2015). Every transformation model entails a monotonically increasing transformation function $h$ that acts on the response and is designed to reframe an unknown distribution $\mathbb{P}(Y \leq y)$ in terms of a transformation $h$, s.t. $\mathbb{P}(h(Y) \leq h(y))$. The advent of parametric transformation models goes back to the Box-Cox model (Box and Cox, 1964) which ignited an area of active research that is still lit to this day. One approach that avoids strong assumptions on the parametric form of the transformation function was introduced by Cheng et al. (1995). It inspired plenty of models that share the estimation of a linear transformation function $h(y|\boldsymbol{x}) = h_Y(y) - \boldsymbol{x}^\top \boldsymbol{\beta}$ where the baseline transformation $h_Y(y)$ is estimated semiparametrically in conjunction with a linear, covariate-dependent shift $\boldsymbol{x}^\top \boldsymbol{\beta}$. Prominent representatives are the proportional odds or the proportional hazards model. One of the first Bayesian transformation models was proposed by Pericchi (1981). Mallick and Walker (2003) model the transformation function $h$ semiparametrically using (Bayesian) Bernstein polynomials and Pólya trees for the estimation of accelerated failure time models among others. Song and Lu (2012) use Bayesian P-splines (Lang and Brezger, 2004) for transformation models with additive shift effects and a Gaussian reference distribution. James et al. (2021) allow for discrete ordered and mixed discrete/continuous outcomes in conjunction with linear covariate effects.

Although very powerful in a lot of applications, transformation models of this type are considerably hindered by the additivity assumption on the scale of the transformation function $h(y|\boldsymbol{x}) = h_Y(y) + h_{\boldsymbol{x}}(\boldsymbol{x})$ where the explanatory variable can only contribute a shift of the baseline transformation $h_Y$ and can therefore influence the conditional location parameter only. One modern example that includes linear interactions of covariates and gained a lot of attention is distribution regression (Chernozhukov et al., 2013). Here, in the context of counterfactuals, the conditional transformation function $h(y|\boldsymbol{x}) = h_Y(y) - \boldsymbol{x}^\top \boldsymbol{\beta}(y)$ is supplemented with varying-coefficient type interactions $\boldsymbol{x}^\top \boldsymbol{\beta}(y)$ where the varying coefficients $\boldsymbol{\beta}(y)$ are estimated on basis of $\mathbb{P}(Y \leq y|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(\mathbb{1}(Y \leq y)|\boldsymbol{X} = \boldsymbol{x})$. This connection allows to account for heteroskedasticity or other patterns that vary with the covariates. An even more flexible variant comes with conditional transformation models (CTMs) as introduced by Hothorn et al. (2014) which share the same goal and aim to obtain an estimator for the whole conditional distribution function.

The main advantage of CTMs over the most popular competitors in distributional regression (GAMLSS and quantile regression) arise from estimating the conditional distribution rather than focusing on a specific property of this distribution (as in quantile regression) while avoiding strong assumptions on a parametric class of response distributions (as in GAMLSS). Furthermore, all properties of the response distribution can be consistently deduced from an estimated CTM. This is also the case for GAMLSS, but requires additional steps in quantile regression where an estimate for the conditional distribution function can only indirectly be determined from a sequence of quantile regressions. In the simulations and applications, we demonstrate these advantages of CTMs in more detail and provide detailed comparisons with competitors (also beyond GAMLSS and quantile regression). For example, we will study the conditional distribution of cholesterol levels depending, among others, on gender and age of patients. A distributional assessment offers additional insights beyond a mean-based analysis which, for example, allows to study typical ranges of cholesterol levels or the risk of exceeding a certain threshold. For a comprehensive comparison and motivation of distributional regression, see Kneib et al. (2022).

In this article, we propose the class of Bayesian conditional transformation models (BCTMs). BCTMs can be understood as a Bayesian interpretation of the likelihood-based CTMs of Hothorn et al. (2018) via the most likely transformation (MLT) model. Both models have in common that they target the direct estimation of the distribution function of a response $Y$ conditional on a set of covariates $\boldsymbol{X} = \boldsymbol{x}$ by means of estimating the conditional transformation function $h(y|\boldsymbol{x})$. Yet, Bayesian inference based on Markov chain Monte Carlo (MCMC) simulations additionally allows us to obtain exact inferences on all quantities of interest without relying on large sample approximations or bootstrap procedures. This can be of particular value in scenarios with smaller samples where the parameters themselves are of secondary interest compared to complex transformations thereof. As the MLT model, the BCTM can be applied to discrete and continuous responses in the presence of random censoring and furthermore includes smoothness penalties for high-dimensional effects induced by the prior supporting stable function estimates. Bayesian principles in connection with the modularity of implementation make it straightforward to expand BTCMs towards more complex prior structures enabling effect selection or different shrinkage properties for example. The idea of using monotonic P-splines for parametrizing transformation functions has been explored before (see e.g., Song and Lu, 2012; Tang et al., 2018), but our approach is innovative in a Bayesian setting with higher-dimensional interactions involving $y$, where curvatures are often complex and control over the penalization mechanism can contribute to a better model understanding. To summarize, we

- introduce BCTMs as a new model class;
- apply a B-spline basis in conjunction with reparametrized basis coefficients to impose monotonicity on the conditional transformation function in the $y$ direction (opposed to the standard of simple Bernstein polynomials in the MLT model);
- supplement the unreparameterized vector of basis coefficients with a partially improper multivariate Gaussian prior that enforces smoothness towards a straight line both for monotonic and for unrestricted nonlinear effects;
- develop Bayesian posterior estimation based on Hamiltonian Monte Carlo (HMC;

Neal et al., 2011; Betancourt, 2017) using the highly-efficient No-U-Turn Sampler (NUTS, Hoffman and Gelman, 2014) for the vector of basis coefficients;

- implement Bayesian model selection;
- evaluate the distribution recovery ability and validity of credible intervals for BCTMs in different simulations and compare them to its main competitors; and
- demonstrate different aspects and practical relevance of BCTMs in applications on cholesterol levels from the Framingham heart study and on leukemia survival times.

The rest of the paper is structured as follows: Sec. 2 introduces BCTMs as a model class consisting of several building blocks including prior assumptions and theoretical properties. Sec. 3 describes posterior estimation including Bayesian model selection. Sections 5 and 4 contain simulations and applications, respectively. Sec. 6 provides a brief review of our findings and proposes several directions for future research. The Supplement contains proofs of theoretical results as well as additional results for the simulations, an additional application on lung cancer survival times and further details.

## 2 Bayesian Conditional Transformation Models

In a CTM, the cumulative distribution function (CDF) of a response $Y \in \mathcal{S} \subset \mathbb{R}$ conditional on a set of covariates $\boldsymbol{X}$ is specified via

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = \mathbb{P}(Y \leq y|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(h(Y|\boldsymbol{x}) \leq h(y|\boldsymbol{x})) = F_Z(h(y|\boldsymbol{x})), \qquad (1)$$

where the covariate-dependent function $h(y|\boldsymbol{x}) : \mathcal{S} \to \mathbb{R}$ is assumed to be monotonically increasing in $y$. The main idea is to transform the response such that it follows a pre-specified reference distribution with continuous distribution function $F_Z : \mathbb{R} \mapsto [0, 1]$, similar as with popular transformations applied to match the assumption of homoscedastic, additive, normally distributed error terms. However, in CTMs the reference distribution is independent of $\boldsymbol{x}$ and does not contain any unknown parameters to be estimated. In this way, a CTM is characterized by the choice of the reference distribution $F_Z$ and a suitable parameterisation of the transformation function $h(y|\boldsymbol{x})$ such that an estimate of the latter yields an estimate of a possibly complex conditional cumulative distribution function (cCDF) $\hat{F}_{Y|\boldsymbol{X}}$. We will discuss both ingredients in more detail below. Naturally, distinctive characteristics of the transformation function such as monotonicity and smoothness are mirrored in $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$, which is why $h(y|\boldsymbol{x})$ has to be modelled with great care.

Following Hothorn et al. (2014), we assume an additive decomposition on the scale of the transformation function into $J$ partial transformation functions, i.e.

$$h(y|\boldsymbol{x}) = \sum_{j=1}^{J} h_j(y|\boldsymbol{x}), \quad j = 1, \dots, J, \tag{2}$$

where $h_j(y|\boldsymbol{x})$, in the broadest sense, can be understood as response-covariate interactions that are monotone only in direction of $y$. This structure is similar to the regression structure of generalized additive models, but rather than modelling the conditional expectation of the response $h_j$ act on the transformed response scale. To ensure identifiability, the partial transformation functions involving nonlinear terms are centered around zero, resulting in the additive decomposition $h(y|\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{J} h_j(y|\boldsymbol{x})$ with overall intercept $\beta_0$, which we will notationally suppress for most of what follows.

In light of (2), it is important to stress that additivity of the transformation function is assumed on the transformed scale, i.e. there is no explicit differentiation between signal and noise as in Gaussian regression models with separable error term. Hence, CTMs come with the benefit of a straightforward entry point to modelling all moments of the response distribution implicitly as functions of $\boldsymbol{x}$. In the realm of CTMs, the flexibility of $h_j(y|\boldsymbol{x})$ constitutes the scope of the impact a covariate is admitted to have on the whole cCDF.

We assume that each of the $J$ partial transformation functions $h_j(\cdot|\boldsymbol{x})$ can be approximated by a linear combination of basis functions, s.t. $h_j(y|\boldsymbol{x}) = \boldsymbol{c}_j(y, \boldsymbol{x})^\top \boldsymbol{\gamma}_j$, where $\boldsymbol{\gamma}_j$ is a vector of basis coefficients. Later we assume monotonicity of each partial transformation function in $y$, i.e. $h'_j(y|\boldsymbol{x}) = \frac{\partial h_j(y|\boldsymbol{x})}{\partial y} \geq 0$, which is sufficient but not necessary for an overall monotonic transformation function $h(y|\boldsymbol{x})$. Monotonicity is much easier to verify and interpret on the level of the partial transformation functions despite imposing stronger assumptions (Hothorn et al., 2014). In less structured, e.g. tree-based models (Hothorn and Zeileis, 2021), assuming monotonicity on the level of the complete transformation function may be more appropriate.

The complete transformation function and its derivative with respect to $y$ are given by

$$h(y|\boldsymbol{x}) = \boldsymbol{c}(y, \boldsymbol{x})^\top \boldsymbol{\gamma}, \quad h'(y|\boldsymbol{x}) = \boldsymbol{c}'(y, \boldsymbol{x})^\top \boldsymbol{\gamma}, \tag{3}$$

with bases $\boldsymbol{c}(y, \boldsymbol{x}) = (\boldsymbol{c}_1(y, \boldsymbol{x})^\top, \ldots, \boldsymbol{c}_J(y, \boldsymbol{x})^\top)^\top$ and $\boldsymbol{c}'(y, \boldsymbol{x}) = (\boldsymbol{c}'_1(y, \boldsymbol{x})^\top, \ldots, \boldsymbol{c}'_J(y, \boldsymbol{x})^\top)^\top$ for the transformation function and its derivative, respectively, and the stacked vector of all basis coefficients $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_J^\top)^\top$.

In the following subsection, we introduce a generic and flexible joint basis for $c_j(y, \boldsymbol{x})$ that does not entail strict assumptions about the relationship between the moments of the response distribution and the respective covariates. Prior distributions, specific bases for the covariate effects, the choice of the references distribution $F_Z$ and a formal definition of our BCTM are covered in Sec. 2.2 to 2.6.

## 2.1 Generic conditional transformation functions

Let $\boldsymbol{a}_j(y)$ and $\boldsymbol{b}_j(\boldsymbol{x})$ denote vectors containing basis function evaluations $B_{j1d_1}(y), d_1 = 1, \ldots D_1$ and $B_{j2d_2}(\boldsymbol{x}), d_2 = 1, \ldots, D_2$ for the response and the covariates, respectively, such that $\boldsymbol{a}_j(y)^\top = (B_{j11}(y), \ldots, B_{j1D_1}(y))$, and $\boldsymbol{b}_j(\boldsymbol{x})^\top = (B_{j21}(y), \ldots, B_{j2D_2}(y))$. Denoting by $\otimes$ the usual Kronecker product, we then obtain the most general form of partial transformation function in a BCTM as $\boldsymbol{c}_j(y, \boldsymbol{x})^\top = (\boldsymbol{a}_j(y)^\top \otimes \boldsymbol{b}_j(\boldsymbol{x})^\top)^\top$, leading to

$$\begin{aligned} h_j(y|\boldsymbol{x}) &= \boldsymbol{c}_j(y, \boldsymbol{x})^\top \boldsymbol{\gamma}_j = (\boldsymbol{a}_j(y)^\top \otimes \boldsymbol{b}_j(\boldsymbol{x})^\top)^\top \boldsymbol{\gamma}_j = \sum_{d_1=1}^{D_1} \sum_{d_2=1}^{D_2} \gamma_{jd_1d_2} B_{j1d_1}(y) B_{j2d_2}(\boldsymbol{x}) \\ h_j'(y|\boldsymbol{x}) &= \boldsymbol{c}_j'(y, \boldsymbol{x})^\top \boldsymbol{\gamma}_j = (\boldsymbol{a}_j'(y)^\top \otimes \boldsymbol{b}_j(\boldsymbol{x})^\top)^\top \boldsymbol{\gamma}_j = \sum_{d_1=1}^{D_1} \sum_{d_2=1}^{D_2} \gamma_{jd_1d_2} B_{j1d_1}'(y) B_{j2d_2}(\boldsymbol{x}). \end{aligned} \tag{4}$$

Essentially, the Kronecker product establishes a parametric interaction by forming pairwise products of the basis functions $B_{j1d_1}(y)$ and $B_{j2d_2}(\boldsymbol{x})$. The derivative with respect to $y$ is therefore also a tensor product involving the differentiated basis functions $B_{j1d_1}'(y) = \partial B_{j1d_1}(y)/\partial y$. Specific restrictions on the two components of the tensor product lead to interesting special cases of the partial transformation function:

- Setting $\boldsymbol{a}_j(y) \equiv 1$ leads to simple shift effects that only depend on the covariates.
- Setting $\boldsymbol{b}_j(\boldsymbol{x}) \equiv 1$ yields an effect of only $y$ that induces changes of the distributional shape (up to other effects).
- Linear effects $\boldsymbol{a}_j(y) = (1, y)$ induce varying coefficient type effects where covariate effects linearly interact with the responses and the response takes the role of the interaction variable while the covariates are the effect modifiers.

Restricting our generic model to

$$h(y|\boldsymbol{x}) = h_Y(y) + h_{\boldsymbol{x}}(\boldsymbol{x}), \tag{5}$$

leads to a location-shift transformation model that comprises various earlier transformation models as special cases. In this case, only the location of the transformed response depends on the covariates via $h_{\boldsymbol{x}}(\boldsymbol{x})$ and higher moments are captured unconditionally by the monotonic transformation $h_Y(y)$. This model type is parametrized by restricting the joint basis to $\boldsymbol{c}(y, \boldsymbol{x})^\top = ((\boldsymbol{a}(y)^\top \otimes 1)^\top, (1 \otimes \boldsymbol{b}(\boldsymbol{x})^\top)^\top) = (\boldsymbol{a}(y)^\top, \boldsymbol{b}(\boldsymbol{x})^\top)^\top$ resulting in the shift transformation model $\boldsymbol{c}(y, \boldsymbol{x})^\top \boldsymbol{\gamma} = \boldsymbol{a}(y)^\top \boldsymbol{\gamma}_1 + \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\gamma}_2$.

We are relying on B-splines for the response dimension while various alternatives are available for the covariate dimension (see Sec. 2.3 for details). The choice of B-splines for representing $\boldsymbol{a}_j(y)$ is mainly determined by the availability of suitable reparameterisations of the corresponding basis coefficients that ensure monotonicity along $y$ in the tensor product for the partial response transformations and well-studied smoothness properties. More precisely, we follow Pya and Wood (2015) and reparameterize the $D = D_1 \cdot D_2$ dimensional basis vector $\boldsymbol{\gamma}_j = (\gamma_{j11}, \dots, \gamma_{j1D_2}, \gamma_{j21}, \dots, \gamma_{jD_1D_2})^\top$ in two steps. First, we set $\boldsymbol{\gamma}_j = \boldsymbol{\Sigma}_j \tilde{\boldsymbol{\beta}}_j$, where $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_{D_1} \otimes \boldsymbol{I}_{D_2}$, $\boldsymbol{I}_{D_2}$ is an identity matrix of size $D_2$, $\boldsymbol{\Sigma}_{D_1}$ is a lower triangular matrix of size $D_1$ with $\Sigma_{D_1,kl} = 0$ if $k < l$ and $\Sigma_{D_1,kl} = 1$ if $k \geq l$, and the vector $\tilde{\boldsymbol{\beta}}_j$ is

$$\tilde{\boldsymbol{\beta}}_j = (\beta_{j11}, \dots, \beta_{j1D_2}, \exp(\beta_{j21}), \dots \exp(\beta_{j2D_2}), \dots, \exp(\beta_{jD_11}), \dots, \exp(\beta_{jD_1D_2}))^\top. \quad (6)$$

Starting with a vector $\boldsymbol{\beta}_j = (\beta_{j11}, \dots, \beta_{j1D_2}, \beta_{j21}, \dots \beta_{j2D_2}, \dots, \beta_{D_11}, \dots, \beta_{jD_1D_2})^\top \in \mathbb{R}^D$ of unconstrained parameters, these choices ensure that the vector of basis coefficients $\boldsymbol{\gamma}_j$ is strictly increasing along the response dimension $y$ which, in turn, implies a tensor product effect that is monotonically increasing along $y$. The complete model vectors of basis coefficients are then given by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_J^\top)^\top$ and $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^\top, \dots, \tilde{\boldsymbol{\beta}}_J^\top)^\top$, while the overall model matrix $\boldsymbol{\Sigma}$ is block diagonal with $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_J$ as diagonal elements. We formalize the monotonicity of $h(y|\boldsymbol{x})$ along the $y$ dimension in the following theorem.

**Theorem 2.1** (Monotonically increasing transformation function along $y$)**.** *Let $h(\cdot|\boldsymbol{x}) : \mathcal{S} \to \mathbb{R}$ be the transformation function* (2) *with basis representation* (3) *and partial transformation functions $h_j$ as in* (4)*. Let furthermore $\boldsymbol{\gamma}_j = \boldsymbol{\Sigma}_j \tilde{\boldsymbol{\beta}}_j$ with $\tilde{\boldsymbol{\beta}}_j$ as in* (6) *and $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_{D_1} \otimes \boldsymbol{I}_{D_2}$ as defined above. Then, $h(\cdot|\boldsymbol{x})$ is monotonically increasing, that is for all $y_1, y_2 \in \mathcal{S}$ with $y_1 < y_2$ we have $h(y_1|\boldsymbol{x}) \leq h(y_2|\boldsymbol{x})$.*

A proof of Theorem 2.1 can be found in the Supp. Part A.

In contrast to the original MLT model introduced by Hothorn et al. (2018) that uses

Bernstein polynomials as a basis for nonlinear effects, the BCTM is supplemented with a smoothness-inducing penalty through its prior and is therefore less restrained regarding the number of model terms $J$ and functional complexity in direction of the covariates. Recently, the MLT model has also been extended in this direction, for shift effects, in the R package tramME (Tamási and Hothorn, 2022).

Of course, other basis functions than B-splines are immediately conceivable for $\boldsymbol{a}_j(y)$. The main requirements for a suitable specification include the ability to incorporate monotonicity constraints, the analytical availability of the basis functions and their derivatives, and the numerically stable evaluation of these. While B-splines fulfill these requirements, investigating other choices and their properties is a promising avenue for future research.

## 2.2 Prior specifications

Overfitting of unregularized splines can be avoided in our Bayesian framework by enforcing smoothness and regularization through shrinkage priors. For the special case of B-splines, Bayesian P-splines assign multivariate Gaussian priors to the regression coefficient vectors. We follow Kneib et al. (2019) and adopt this principle to tensor product terms such that the prior for the coefficient vector $\boldsymbol{\beta}_j$ associated with one of the partial transformation functions $h_j$ in (4) is multivariate Gaussian with expectation zero and precision matrix

$$\boldsymbol{K}_j \equiv \boldsymbol{K}_j(\tau_j^2, \omega_j) = \frac{1}{\tau_j^2}\Big[\omega_j(\boldsymbol{K}_{1j} \otimes \boldsymbol{I}_{D_2}) + (1-\omega_j)(\boldsymbol{I}_{D_1} \otimes \boldsymbol{K}_{2j})\Big], \tag{7}$$

where $\boldsymbol{K}_{j1}$ and $\boldsymbol{K}_{j2}$ are potentially rank deficient prior precision matrices of dimensions $(D_1 \times D_1)$ and $(D_2 \times D_2)$, respectively, controlling the type of smoothness required along the response and the covariate dimension, respectively. For the response dimension, we set $\boldsymbol{K}_{1j} = \boldsymbol{D}_{1j}^\top \boldsymbol{D}_{1j}$ where $\boldsymbol{D}_{1j}$ is a $(D_1-2) \times D_1$ partial first difference matrix consisting only of zeros except that $\boldsymbol{D}_j[d_1, d_1+1] = -\boldsymbol{D}_j[d_1, d_1+2] = 1$ for $d_1 = 1, \ldots, D_1-2$ (Pya and Wood, 2015; Pya, 2010). The prior precision matrix $\boldsymbol{K}_{2j}$ shrinks in the direction of the respective covariate and the specific choice depends on the considered covariate effect of interect (see Sec. 2.3 for some examples). For monotonic nonlinear effects, the resulting penalty is quadratic in the (non-exponentiated) parameters $\boldsymbol{\beta}_j$. This corresponds to log differences in $\gamma_{jd_1d_2}$ for $d_1 > 2$, such that a first order random walk prior penalizes the squared differences between adjacent $\beta_{jd_1d_2}$, resulting in shrinkage towards a straight line, similar to second order random walk penalties for univariate P-splines (Pya and Wood, 2015). The complete prior precision matrix $\boldsymbol{K} \equiv \boldsymbol{K}(\boldsymbol{\tau}^2, \boldsymbol{\omega})$ is given as the block diagonal matrix with matrices $\boldsymbol{K}_j$ as diagonal elements. We formalize the prior for $\boldsymbol{\gamma}_j$ in the following proposition.

**Proposition 2.2** (Prior for $\boldsymbol{\gamma}_j$)**.** *Let* $p_\beta(\boldsymbol{\beta}_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j^\top \boldsymbol{K}_j^- \boldsymbol{\beta}_j\right)$ *be the partially improper multivariate Gaussian prior with generalized inverse* $\boldsymbol{K}_j^-$ *of the prior precision matrix in* (7)*. Assume furthermore for notational simplicity that* $D_2 = 1$ *such that* $\gamma_{jd1d2} \equiv \gamma_{jd1}$*. Then, the prior for* $\boldsymbol{\gamma}_j$ *is given by*

$$p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_j|\tau^2) \propto p_{\boldsymbol{\beta}}\left(\gamma_{j1}, \log(\gamma_{j2}-\gamma_{j1}), \dots, \log(\gamma_{jD} - \sum_{i=1}^{D-1}(D-i)\gamma_{ji})\right) \prod_{k=1}^{D} \frac{1}{\gamma_{jk} - \sum_{i=1}^{k-1}(k-i)\gamma_{ji}}$$

*Proof.* The proof follows directly by applying the multivariate change of variable theorem twice to the transformation $g_2 \circ g_1 : \mathbb{R}^D \to \mathbb{R}^D$ with $\tilde{\beta}_1 = g_1(\beta_{j1}) = \tilde{\beta}_{j1}$, $\tilde{\beta}_{jk} = g_1(\beta_{jk}) = \exp(\beta_{jk})$, $k = 2, \dots, D$ and $\gamma_{jk} = g_2(\tilde{\beta}_{jk}) = \sum_{i=1}^{k} \tilde{\beta}_{ji}$, $k = 1, \dots, D$. □

The amount of smoothness induced by the precision matrix (7) is controlled by the overall smoothing variance $\tau_j^2 > 0$ and the weight parameter $\omega_j \in [0, 1]$. Following Kneib et al. (2019), we assume a discrete prior for the latter which has the advantage that generalized determinants of $\boldsymbol{K}_j$ can be pre-computed which considerably facilitates the numerically efficient implementation while still enabling anisotropic amounts of smoothness along the response and the covariate dimension. A uniform prior on a moderate number of equispaced values is used as a default for $\omega_j$. For the smoothing variance $\tau_j^2$, we consider two alternatives: Standard inverse gamma (IG) priors $\tau_j^2 \sim \mathrm{IG}(a_j, b_j)$ where the hyperparameters are chosen among popular combinations such as $a_j = 1$, $b_j = 0.001$ to mimic a weakly informative setting, and scale-dependent (SD) hyperpriors as suggested in Klein and Kneib (2016). The latter results in a Weibull prior for $\tau_j^2$ with shape parameter 0.5 and scale parameter $\theta$ determined from a scaling criterion on expected effect sizes. We transfer this concept to partially monotonic tensor product effects where, to achieve numerical stability, it is important to control the variation of those parameters exponentiated in (6). More precisely, we consider the scaling criterion $\mathbb{P}\left(\max_{d_1=2,\dots,D_1, d_2=1,\dots,D_2} |\beta_{jd_1,d_2}| \leq c\right) = 1 - \alpha$ with user-specified values for $c$ and $\alpha$. To determine the marginal prior distribution of $\tau_j^2$ required to evaluate the scaling criterion, we follow a simulation-based approach to marginalize out any additional hyperparameters. From the support of the exponential function, $c = 3$ and $\alpha = 0.01$ are useful standards also used later in our empirical studies.

In a last step, we collect all model parameters in $\boldsymbol{\vartheta} = (\beta_0, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J, \tau_1^2, \dots, \tau_J^2, \omega_1, \dots, \omega_J)^\top = (\boldsymbol{\beta}^\top, (\boldsymbol{\tau}^2)^\top, \boldsymbol{\omega}^\top)^\top$ with prior $\pi_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})$ which can be factorized into products of the individual priors. The coefficient $\beta_0$ denotes the intercept, while $\boldsymbol{\beta}$, $\boldsymbol{\tau}^2$, $\boldsymbol{\omega}$ are used to denote all basis coefficients, smoothing variances and anisotropy weights, respectively.

## 2.3 Bases for the covariate effects

We highlight special cases of bases $\boldsymbol{b}(\boldsymbol{x})$ relevant for our applications:

- *Linear effects.* The basis for linear effects of covariates $x_1, \dots, x_p$ collected in $\boldsymbol{x}$ is $\boldsymbol{b}_j(\boldsymbol{x})^\top = (x_1, \dots, x_p)^\top$ and we use a non-informative prior with $\boldsymbol{K}_{2j} = \boldsymbol{0}$. This also applies to the overall intercept $\beta_0$ when centering the partial transformation functions.
- *Random effects.* Random effects (or frailties) are based on a grouping indicator $g \in \{1, \dots, G\}$. The resulting $G$-dimensional basis vector $\boldsymbol{b}_j(g)$ has entry one if $y$

belongs to group $g$ and zero otherwise and we set $\boldsymbol{K}_{2j} = \boldsymbol{I}_G$ for i.i.d. random effects.

- *Discrete spatial effects.* Similar to random effects, a spatial effect of a discrete spatial variable $s \in \{1, \dots, S\}$ is constructed as an indicator with entries in the $S$-dimensional basis vector $\boldsymbol{b}_j(s)$ set to one if $y$ belongs to region $s$ and zero otherwise. We induce spatial smoothing in form of a Gaussian Markov random field (GMRF Rue and Held, 2005) prior. The precision matrix $\boldsymbol{K}_{2j}$ reflects the spatial orientation of the data, i.e. we define two regions as neighbours if they share a common border.

## 2.4 Choice of the reference distribution

As already stated, it is the task of the conditional transformation function $h(y|\boldsymbol{x})$ to transform the response values conditionally on the explanatory variables $\boldsymbol{x}$ such that they follow the reference distribution $F_Z$. In that light, $F_Z$ plays a similar role as the inverse known link function in prominent model classes such as generalized linear models, but is less restrictive in the sense that the resulting conditional distribution does not have to be of known type. From a modelling perspective, it guarantees that the resulting estimated conditional density function integrates to one without requiring complex constraints. Note that no unknown parameters are included in $F_Z$ and no restrictions besides continuity and log-concavity of the reference density $f_Z(y) = \frac{\partial F_Z(y)}{\partial y}$ are required. In theory, any cCDF can be represented as a BCTM when the transformation function is chosen flexible enough. However, in practice the actual ability to represent various types of cCDFs is limited by the choices made for parameterizing the transformation function. For example, when the reference distribution has light tails, one requires considerable flexibility in the transformation function to enable the representation of heavy-tailed distributions. Similarly, restricting the shape of the influence that the covariates can have on the transformation function also imposes restrictions on the cCDFs that can be generated via a BCTM.

Other relevant aspects for the choice of the reference distribution entail (i) interpretation, (ii) convenience, and (iii) theoretical properties. For the sake of interpretation, it is advised to consider further characteristics such as skewness or positivity of $y$ when choosing $F_Z$. Prominent options also used in the applications in Sec. 5 are the standard normal CDF, $F_Z(z) = \Phi(z)$, the standard logistic CDF $F_Z(z) = F_{\mathrm{SL}}(z) = (1 + \exp(-z))^{-1}$ (leading to (non-)proportional odds models) and the minimum extreme value distribution, $F_Z(z) = F_{\mathrm{MEV}}(z) = 1 - \exp(-\exp(z))$ (leading to (non-)proportional hazards models). Note that simple transformation models of type (5) are interpretative in the sense that the term $h_{\boldsymbol{x}}(\boldsymbol{x})$ constitutes the log odds ratio if $F_Z(z) = F_{\mathrm{SL}}(z)$ and the log hazards ratio if $F_Z(z) = F_{\mathrm{MEV}}(z)$, a previous result we use in Sec. 5.2.

The convenience argument favours distributions that are both well studied and numerically easy. Finally, certain properties of the resulting estimates also depend on the choice of the reference distribution. For example, restricting the reference distribution to have log-concave densities ensures that (under standard regularity conditions) that the MLE is unique and consistent which, in turn, often implies unimodal posteriors that are easier to explore with MCMC schemes. For an overview of the numerous possibilities of reference distributions that come with CTMs, see Hothorn et al. (2018).

## 2.5 Transformation densities

In this section, we introduce the conditional transformation densities $f_Y(y|\boldsymbol{\beta})$ given the vector of basis coefficients $\boldsymbol{\beta}$ (before the reparameterization). To emphasize that $\boldsymbol{\gamma}$ is a partially nonlinear reparameterization of $\boldsymbol{\beta}$, we write $\boldsymbol{\gamma}(\boldsymbol{\beta})$.

**Continuous responses** The density and log-density can easily be derived from equation

(1) together with the parametrization of $h$ and $h'$ in equation (4) such that

$$f_Y(y|\boldsymbol{\beta}) = \frac{\partial F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)}{\partial y} = f_Z\left(h(y|\boldsymbol{x})\right) h'(y|\boldsymbol{x}) = f_Z(\boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta}))\boldsymbol{c}'(y,\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta}), \qquad (8)$$

where $f_Z$ denotes the density of the chosen reference distribution. In theory, for any absolute continuous response distribution $F_Y$ and reference distribution $F_Z$ with log-concave density $f_Z$, there exists a unique, monotonically increasing transformation function $h$, such that $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(\cdot) = F_Z(h(\cdot|\boldsymbol{x}))$ (see Cor. 1 of Hothorn et al., 2018).

It is important to note that both for univariate and bivariate effects involving $y$, the part of the effect that belongs to the null space of $\boldsymbol{K}_j$ consists of all location shifts and linear effects in $y$. In the context of BCTMs with the popular choice $F_Z = \Phi$, this means that the penalty shrinks towards the Gaussian location-scale family. In other words, the null space of the rank-deficient precision matrix consists of all Gaussian conditional distribution functions. This observation can be put to use when considered from the perspective of SD priors for the variances as described in Klein and Kneib (2016).

**Discrete ordinal responses** In case of discrete ordinal responses with a finite sample space where $Y \in \{y_1, \ldots, y_K\}$, the corresponding conditional density function is given by

$$f_Y(y_k|\boldsymbol{\beta}) = \begin{cases} F_Z(\boldsymbol{c}(y_1,\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) & k = 1 \\ F_Z(\boldsymbol{c}(y_k,\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) - F_Z(\boldsymbol{c}(y_{k-1},\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) & k = 2, \ldots, K-1 \\ 1 - F_Z(\boldsymbol{c}(y_{K-1},\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) & k = K. \end{cases} \qquad (9)$$

For countably infinite sample spaces (as e.g. for count data) with $Y \in \{y_1, y_2, y_3, \ldots\}$, the density is given by

$$f_Y(y_k|\boldsymbol{\beta}) = \begin{cases} F_Z(\boldsymbol{c}(y_1,\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) & k = 1 \\ F_Z(\boldsymbol{c}(y_k,\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) - F_Z(\boldsymbol{c}(y_{k-1},\boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) & k > 1. \end{cases} \qquad (10)$$

**Censored responses** The Bayesian conditional transformation model incorporates all forms of random censoring. In the presence of censored observations, only the likelihood has to be adapted while the transformation function remains the same. The likelihood contributions for right-, left, and interval-censored continuous or discrete observations respectively are then given by

$$\begin{aligned} 1 - F_Z(\boldsymbol{c}(\underline{y}, \boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) \text{ for } y \in (\underline{y}, \infty) & \qquad \text{“right censored”} \\ F_Z(\boldsymbol{c}(\overline{y}, \boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) \text{ for } y \in (-\infty, \overline{y}) & \qquad \text{“left censored”} \\ F_Z(\boldsymbol{c}(\overline{y}, \boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) - F_Z(\boldsymbol{c}(\underline{y}, \boldsymbol{x})^\top \boldsymbol{\gamma}(\boldsymbol{\beta})) \text{ for } y \in (\underline{y}, \overline{y}] & \qquad \text{“interval censored”}. \end{aligned} \tag{11}$$

It is also possible to adapt densities for truncated observations (Hothorn et al., 2018).

## 2.6 Formal definition of BCTMs

**Definition 2.3** (BCTM)**.** *The quadruple* $\big(\boldsymbol{\vartheta}, F_Z, \boldsymbol{c}, \pi_{\vartheta}(\cdot)\big)$ *of unknown model parameters* $\boldsymbol{\vartheta}$*, a choice for the basis* $\boldsymbol{c}$*, the reference distribution* $F_Z$ *and joint prior* $\pi_{\vartheta}$ *is called Bayesian conditional transformation model (BCTM).*

# 3 Posterior Inference

## 3.1 Posterior and estimation via MCMC

Assuming conditional independence the joint posterior is given by

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \boldsymbol{\omega} | \boldsymbol{y}) \propto \prod_{i=1}^{n} f_Y(y_i | \boldsymbol{\beta}) \left[ \pi(\beta_0) \prod_{j=1}^{J} [\pi(\boldsymbol{\beta}_j | \tau_j^2, \omega_j) \pi(\tau_j^2), \pi(\omega_j)] \right]. \tag{12}$$

To obtain samples from (12) we use an MCMC sampler with three alternating steps:

**Step 1.**: Sample from $p(\boldsymbol{\beta} | \boldsymbol{\tau}^2, \boldsymbol{\omega}, \boldsymbol{y})$ using the NUTS.

**Step 2.**: For $j = 1, \ldots, J$, sample from $p(\tau_j^2 | \boldsymbol{\beta}_j, \boldsymbol{y})$ using a Gibbs sampler in case of an IG prior or iteratively weighted least squares (IWLS) proposals in case of SD priors.

**Step 3.**: For $j = 1, \dots, J$, sample $\omega_j$ with a Gibbs step from its discrete full conditional.

The resulting MCMC samples can then be used to estimate various model properties based on their posterior, relying on the law of large numbers. For example, the conditional distribution $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$ as $\hat{F}_{Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{y}}(y) = F_Z(\hat{h}(y|\boldsymbol{x}))$ where $\hat{h}(y|\boldsymbol{x})$ is the posterior mean estimate $\hat{h}(y|\boldsymbol{x}) = \boldsymbol{c}(y, \boldsymbol{x})^\top \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\gamma}^{[s]}$ with posterior samples $\boldsymbol{\gamma}^{[1]}, \dots, \boldsymbol{\gamma}^{[S]}$. Similarly, a posterior mean estimate for $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$ can be determined as $\hat{F}_{Y|\boldsymbol{X}=\boldsymbol{x}} = \frac{1}{S}\sum_{s=1}^{S} F_Z(\boldsymbol{c}(y, \boldsymbol{x})^\top \boldsymbol{\gamma}^{[s]})$. The posterior samples also provide us with the basis of deriving the complete posterior distribution of $h(y|\boldsymbol{x})$, $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$, and any transformation thereof.

**Updating the basis coefficients** Basis coefficients are updated jointly by sampling from the log full conditional $\log(p(\boldsymbol{\beta}|\boldsymbol{\tau}^2, \boldsymbol{\omega}, \boldsymbol{y})) \propto \sum_{i=1}^{n} f_Y(y_i|\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{K}(\boldsymbol{\tau}^2, \boldsymbol{\omega})\boldsymbol{\beta}$, where the first term arises from one of the likelihoods described in Sec. 2.5 and the second term arises from the Gaussian prior). High dimensionality and strong dependencies among coefficients (stemming partly from the monotonicity constraints) aggravate sampling from the posterior distribution. This is further exacerbated by the mixed linear-nonlinear dependence of the transformation function on $\tilde{\boldsymbol{\beta}}$, rendering e.g. random-walk Metropolis algorithms slow and inefficient. One possible remedy lies in including gradient information as done by HMC. This, however, comes with the drawback that two additional tuning parameters (step size $\epsilon$ and number of leapfrog steps $L$) have to be set manually. To avoid this tricky task, we implement NUTS with dual averaging (Nesterov, 2009) that uses Hamiltonian principles for efficient exploration of the target distribution of $\boldsymbol{\beta}$ in an adaptive fashion The adaptive nature of NUTS enables a streamlined estimation process, effectively abolishing the need for costly preliminary tuning runs at the expense of some additional computation time per iteration which is owed mainly to the more sophisticated proposals.

The required gradient of the unnormalized log-posterior of the basis coefficients vector

$\boldsymbol{\beta}$ for continuous responses is given by

$$s(\boldsymbol{\beta}) \equiv \frac{\partial \log(p(\boldsymbol{\beta}|\boldsymbol{\tau}^2, \boldsymbol{y}))}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[ \boldsymbol{c}(y_i, \boldsymbol{x}_i)^\top \boldsymbol{\Sigma} \boldsymbol{C} \frac{f'_Y(y_i|\boldsymbol{\beta})}{f_Y(y_i|\boldsymbol{\beta})} + \frac{\boldsymbol{c}'(y_i, \boldsymbol{x}_i)^\top \boldsymbol{\Sigma} \boldsymbol{C}}{\boldsymbol{c}'(y_i, \boldsymbol{x}_i)^\top \boldsymbol{\Sigma} \tilde{\boldsymbol{\beta}}} \right] - \boldsymbol{K}\boldsymbol{\beta},$$

where $\boldsymbol{C}$ is a diagonal matrix with entries $C_{dd} = 1$ if $\tilde{\beta}_d = \beta_d$, $C_{dd} = \exp(\beta_d)$ otherwise, and similar expressions can straightforwardly be derived for discrete or censored responses.

Potentially flat parts of a fitted transformation function based on the reparameterization in Sec. 2.1 demand some parameter components to be close to zero and thus the corresponding logarithms to approach minus infinity. For NUTS, this does not result in overflow errors, but can lead to divergent transitions in the sampling path and NUTS trees with large tree depth as the different curvatures demand very different step sizes. Using a non-centered parametrization (Papaspiliopoulos et al., 2007) as a remedy is not feasible in a straightforward manner, because of the nonlinear transformation in the coefficient vectors. Instead, we found it helpful to increase the goal acceptance rate, forcing the sampler to take smaller steps, which is a small price to pay for the non-occurrence of divergencies. If the problem persists it is possible to drop unidentified (i.e. reparameterized coefficients that should be close to zero) in each iteration judging by the eigenvalues of the matrix square root of the Hessian of the posterior at (12) in an efficient way (Pya and Wood, 2015).

Furthermore, we resort to augmented precision matrices, e.g. $\boldsymbol{K}_j = \frac{1}{\tau_j^2}[\omega_j \boldsymbol{K}_{1j} + (1 - \omega_j)\boldsymbol{K}_{2j}] + 10^{-6}\boldsymbol{I}$ to ensure positive definiteness and therefore a soft threshold for coefficient variances (Andrinopoulou et al., 2018). The NUTS warm-up phase can often be supported by standardizing each covariate or by rescaling them to (0,1). Both measures can facilitate mass matrix adaption. Regarding sampling efficiency, we found that using SD priors for the smoothing variances can decrease run times and improve the effective sample size.

**Updating the smoothing variances** When using an IG prior for the smoothing variances, they can be updated directly with a Gibbs step from the full conditional $\tau_j^2|\cdot \sim$

$\mathrm{IG}\left(a_j + \frac{\mathrm{rk}(\boldsymbol{K}_j)}{2}, b_j + \frac{1}{2}\boldsymbol{\beta}_j^\top \boldsymbol{K}_j \boldsymbol{\beta}_j\right)$. For the SD prior, updates can be implemented via IWLS proposals of log-variances following Klein and Kneib (2016).

**Updating the weights** The updates of the weights are straightforward using Gibbs sampling due to their discrete prior structure (Kneib et al., 2019).

**Computational details** While BCTMs are pretty robust regarding the choice of hyperparameters, varying them can improve computational speed and stabilize estimates that involve a monotonicity constraint. All results shown in Secs. 4, 5 were obtained with $4,000$ MCMC iterations with a NUTS warm-up phase of $2,000$ and a burn-in of $2,000$. Computations were carried out in R version 4.1.0 (R Core Team, 2020). To improve computing time, parts of the sampler were programmed using `Rcpp` (Eddelbuettel and Balamuta, 2017). The `MASS` matrix adaption scheme was adopted from `adnuts` (Monnahan and Kristensen, 2018).

## 3.2 Model choice and variable selection

For model selection, we use the Watanabe-Akaike information criterion (WAIC Watanabe, 2010). It can be seen as approximation to computationally expensive cross validation (CV) and is conveniently computed from $s = 1, \dots, S$ posterior samples. We validated WAIC against CV in some of our applications and found good agreements that support using information criteria as the basis for model choice and variable selection.

The WAIC overcomes certain limitations of the DIC (DIC; Spiegelhalter et al., 2002) such as its dependence on the posterior mean as a specific point estimate or the potential of observing negative effective parameter counts. It is given by $\mathrm{WAIC} = (-2l_{\mathrm{WAIC}} + 2p_{\mathrm{WAIC}})$, where $l_{\mathrm{WAIC}} = \sum_{i=1}^n \left(\frac{1}{S}\sum_{s=1}^S f_Y(y_i|\boldsymbol{\beta}^{[s]})\right)$ and $p_{\mathrm{WAIC}} = \sum_{i=1}^n \mathrm{Var}(\log(f_Y(y_i|\boldsymbol{\beta})))$. In the regression literature, information criteria like the DIC and the WAIC are primarily used to discriminate between different types of response distributions and predictor specifications (e.g. Klein, Kneib, Lang, Sohn et al., 2015). In the holistic approach of BCTMs, the transformation function determines both the response distribution and the “predictor” which is why it is sufficient to use information criteria to compare different (partial) transformation function specifications that differ in flexibility and interaction structure. We also considered the deviance information criterion (DIC) as introduced by Spiegelhalter et al. (2002) which yielded similar results and is therefore omitted in the following.

# 4 Simulations

We conducted simulations to evaluate the empirical performance of BCTMs to recover the true data generating process compared to several competing methods from the literature (Sec. 4.1). Here, we also confirm the comparable ability of MLTs and BCTMs of modelling cCDFs in general. However, in Sec. 4.2 (and the applications and discussion in Secs. 5,6) we demonstrate the advantages that BCTMs may offer, such as to providing valid uncertainty estimates by means of coverage rates.

## 4.1 Recovering the conditional distribution

In this section, we mimic the simulation design of Hothorn et al. (2014) to benchmark our BCTM against its frequentist counterpart, the MLT as implemented in the R-package (**mlt**, Hothorn, 2017), Bayesian GAMLSS (Klein, Kneib, Lang, Sohn et al., 2015), and Bayesian semiparametric quantile regression (Waldmann et al., 2013). For both Bayesian benchmarks, we use the R package **bamlss** (Umlauf et al., 2018).

**Simulation design** For datasets of size $n = 200$, we generate two covariates as i.i.d. realizations via $x_1 \sim U[0, 1]$ as well as $x_2 \sim U[-2, 2]$. The response $y$ is assumed to follow a heteroscedastic varying coefficient model (VCM)

$$Y = \frac{1}{x_1 + 0.5} x_2 + \frac{1}{x_1 + 0.5}\epsilon, \quad \epsilon \sim \mathrm{N}(0, 1), \tag{13}$$

such that an appropriate CTM has to emulate a Gaussian location-scale model under the premises that the mean depends on the nonlinear varying coefficient $(x_1 + 0.5)^{-1}$ for $x_2$ and that the variance is a nonlinear function of $x_1$. To analyse the stability in the presence of noise variables, we consider six scenarios, where $p = 0, \dots, 5$ i.i.d. realizations from the standard uniform $U[0, 1]$ with zero influence on the response are added. The complete vector of covariates is denoted by $\boldsymbol{x}_p = (x_1, x_2, x_3, \dots, x_{p+2})^\top$.

**Benchmark methods** For each of the resulting six scenarios, we fit

- Lin. BCTM $\big(\boldsymbol{\vartheta}, \Phi, ((1, y) \otimes (1, \boldsymbol{x}_p^\top))^\top, \pi_\vartheta(\boldsymbol{\vartheta})\big)$: a restricted BCTM consisting of simple linear interactions
- Lin. MLT: a linear MLT of the same type
- Full BCTM $\big(\boldsymbol{\vartheta}, \Phi, (\boldsymbol{a}(y)^\top \otimes (\boldsymbol{b}(x_1)^\top, \dots, \boldsymbol{b}(x_{p+2})^\top))^\top, \pi_\vartheta(\boldsymbol{\vartheta})\big)$: a nonlinear BCTM consisting of nonlinear interactions with basis dimension of 10 in $\boldsymbol{a}$ and $\boldsymbol{b}$
- Full MLT: a nonlinear MLT of the same type with Bernstein polynomials of order 10, i.e. with joint basis $(\boldsymbol{a}_{\mathrm{Bs}}(y)^\top \otimes (\boldsymbol{b}_{\mathrm{Bs}}(x_1)^\top, \dots, \boldsymbol{b}_{\mathrm{Bs},10}(x_{p+2})^\top))^\top$
- Oracle BAMLSS: a Gaussian location-scale BAMLSS based on model (13), i.e. $\eta_\mu = \beta_0 + x_2 \cdot f(x_1) + \sum_{k=0}^{p} f(x_{2+k})$ and $\eta_{\sigma^2} = \beta_0 + f(x_1)$ and
- BAMLSS QR: a Bayesian semiparametric quantile regression specification with nonlinear effects of all explanatory variables

Note that it is not straightforward to compare the potential complexity of function estimates achieved with the spline-based specifications in BCTMs with the one of polynomial specifications in MLTs. We used a relatively large degree to avoid trivially better performance of BCTMs. Further details on the model specifications are given in Supp. Tab. C.7.

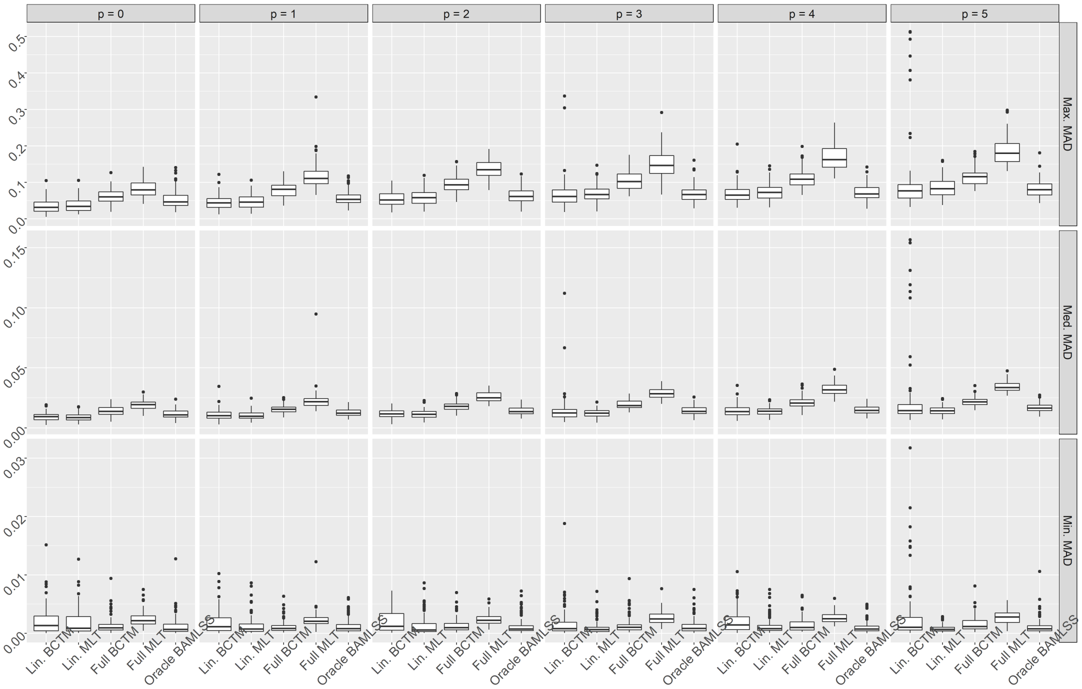

Figure 1: Simulation 1. Minimum, median and maximum of the mean absolute deviation (MAD) between true and estimated predictive probabilities for the Lin. BCTM/MLT, the Full BCTM/MLT and the BAMLSS for the six scenarios with $p = 0, \ldots, 5$ noise covariates based on 100 replications each.

It is important to stress that Lin. BCTM/MLT and Oracle BAMLSS have in common that they are restricted by design to the true (Gaussian) distribution. In addition, the Oracle BAMLSS is the only model that is supplemented with the true predictor for the variance in all scenarios. Yet, despite being linear in the covariates on the scale of the transformation function, the Lin. BCTM/MLT are nonlinear on the scale of the response. Since this information is in general not available, we also include the Full BCTM/MLT.

**Performance measures** As a first measure of performance, we computed the mean absolute deviation (MAD) of the estimates of $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$ from the true probabilities over a grid of $y$, $x_1$ and $x_2$ $\mathrm{MAD}(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} |F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_i) - \hat{F}_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_i)|$ based on 100 replications. Fig. 1 summarizes the empirical distributions of the minimum, median and maximum MAD for all models that provide estimates for the complete cCDF, i.e. all but the BAMLSS QR. As a second performance measure, we computed conditional quantiles of the fitted response

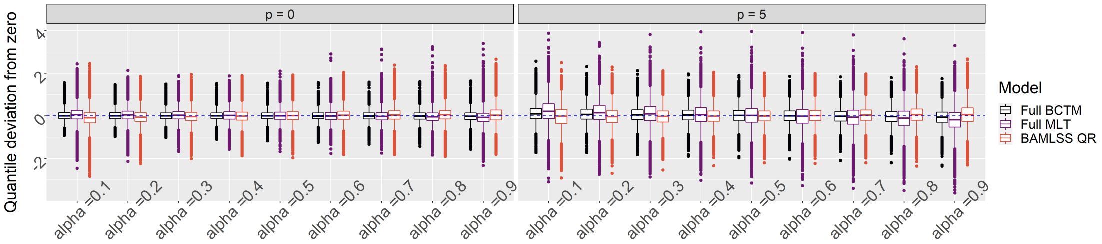

Figure 2: Simulation 1. Bias in estimated quantiles for various quantile levels $\alpha$ for the six scenarios with $p = 0, \ldots, 5$ noise covariates based on 100 replications each.

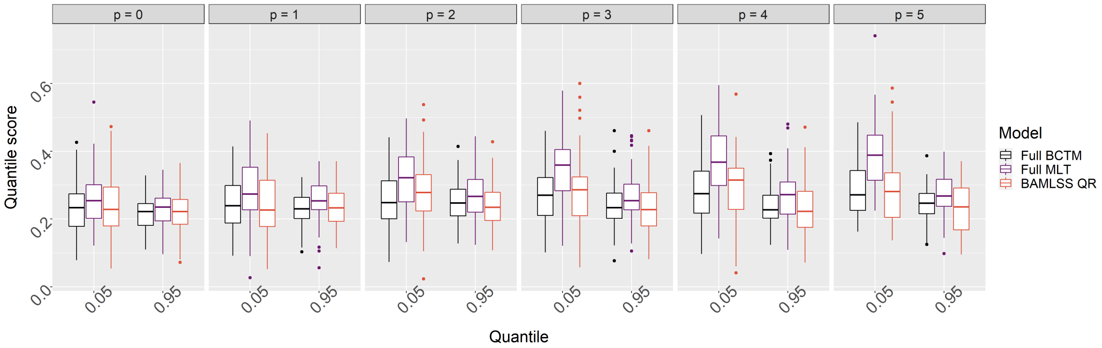

Figure 3: Simulation 1. Estimated quantile scores $\text{QS}(F^{-1}(\alpha), y)$ for $\alpha = 0.05$ and 0.95 for the six scenarios with $p = 0, \ldots, 5$ noise covariates based on one test dataset each.

distribution corresponding to a sequence of probabilities $\alpha$ via numerical inversion. Fig. 2 shows the deviations of these from their true counterparts together with similar results obtained via QR BAMLSS which was used as a benchmark. Third, as a measure of accuracy that concentrates on tail features of the distribution, Fig. 3 shows the quantile score function $\text{QS}_\alpha(F^{-1}(\alpha), y) = 2(\mathcal{I}(y < F^{-1}(\alpha)) - \alpha)(\alpha - y)$ at $\alpha = 0.05$ and $\alpha = 0.95$, where $\mathcal{I}(A)$ $= 1$ if $A$ is true, and zero otherwise (Gneiting, 2011). Last, to measure the overall forecast accuracy, we plot the decomposition of the continuous ranked probability score (CRPS; Laio and Tamea, 2007) which can be written as $\text{CRPS}(F, y) = \int_0^1 \text{QS}_\alpha(F{-}1(\alpha), y) d\alpha$ in Fig. 4. Both, the QS and CRPS are based on the prediction grids used for the MAD and lower values suggest greater accuracy.

**Results** Lin. BCTM yields MADs that are very close to those of the Oracle BAMLSS and also performs better than Lin. MLT for all $p$. The Full BCTM/MLT have somewhat higher

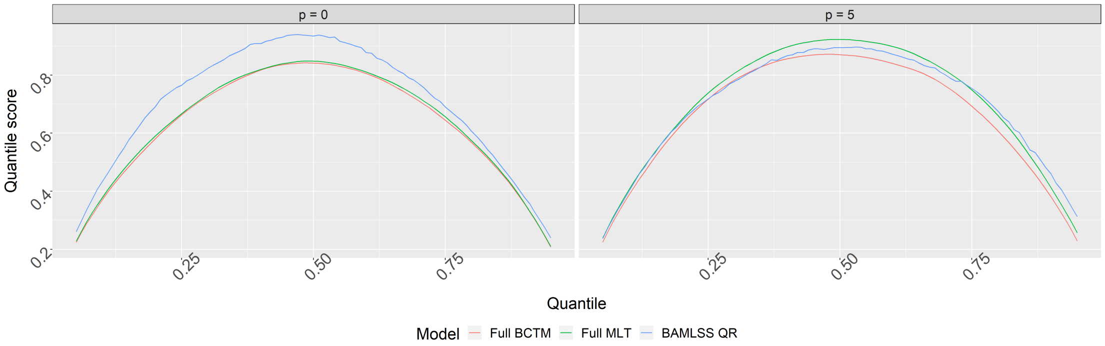

Figure 4: Simulation 1. Estimation CRPS decomposition for two scenarios with $p = 0, 5$ noise covariates based on one test dataset each.

MADs, and MLT is again worse than BCTM when the number of noise variables is large. All methods do recover the true conditional quantiles well. Full BCTM/MLT are on par with BAMLSS and in particular with the BAMLSS QR which is specifically tailored to estimate conditional quantiles. Full BCTM performs best in terms of QS and CRPS. Specifically, Full BCTM and BAMLSS QR are similar and outperform Full MLT in terms of QS, while full BCTM is slightly better than Full MLT and Bayes QR worst according to the CRPS. In summary, BCTMs (similar as their frequentist counterpart) enable proper and reliable modelling of the complete cCDF and its quantiles, while avoiding restrictive assumptions on the shape of the distribution. In the following simulations on coverage rates and in the applications, we provide details on situations where BCTMs offer particular advantages over MLTs, see also the summary in the discussion section.

## 4.2 Coverage rates

To compare BCTM and MLT from a different perspective, we consider empirical coverage rates of pointwise 95% credible/confidence intervals in a simulation setting that concentrates on the estimation of nonlinear covariate effects.

**Simulation design** For datasets of size $n = 100$ (for $n = 500$, see Supp. Part C), we generate four i.i.d. covariates via $x_p \sim U[-2, 2]$, $p = 1, \dots, 4$ and assume four test functions

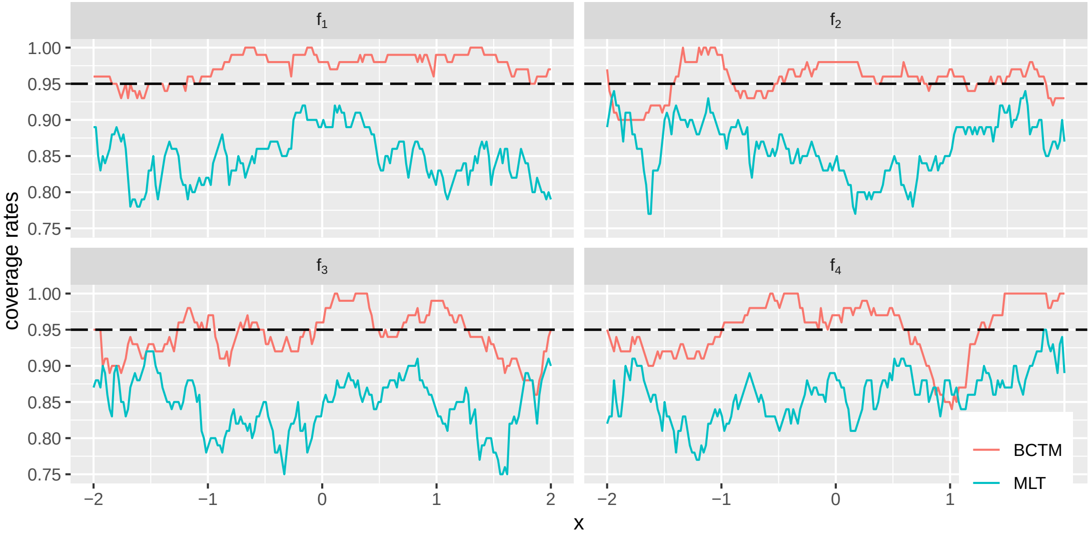

Figure 5: Simulation 2. Coverage rates of pointwise 95% credible/confidence intervals of BCTM (red) and MLT (blue) for $f(x_1), \dots, f(x_4)$ evaluated on equally spaced grids within the range of $x_1, \dots, x_4$. The nominal 95% level is shown as a dashed line.

$f_1(x) = x$, $f_2(x) = x + \frac{(2x-2)^2}{5.5}$, $f_3(x) = -x + \pi\sin(\pi x)$ and $f_4(x) = 0.5x + 15\phi(2(x - 0.2)) - \phi(x + 0.4)$, where $\phi(\cdot)$ denotes the density of the standard normal distribution. The responses are then generated as $y = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + \epsilon$, $\epsilon \sim \mathrm{N}(0, 1)$.

**Benchmark methods** We fit linear Gaussian CTMS, i.e.

- $\big(\boldsymbol{\vartheta}, \Phi, ((1, y)^\top, (\boldsymbol{b}(x_1)^\top, \dots, \boldsymbol{b}(x_4)^\top))^\top, \pi_{\vartheta}(\cdot)\big)$: a linear BCTM with 20 B-spline basis functions in $\boldsymbol{b}$ and nonlinear shift effects
- a linear MLT with nonlinear shifts of the same type specified in terms of Bernstein polynomials of order 10, i.e. with joint basis $((1, y)^\top, (\boldsymbol{b}_{\mathrm{Bs},10}(x_1)^\top, \dots, \boldsymbol{b}_{\mathrm{Bs},10}(x_4)^\top))^\top$

We also experimented with the recent extension of MLT to penalized spline shift effects provided in the tramME package.

**Performance measure** Empirical coverage rates of pointwise 95% credible/confidence intervals based on 100 replications are shown in Fig. 5. For the BCTM, these can readily be computed from the MCMC output, while for the MLT an additional computationally costly parametric bootstrap (Hothorn, 2017) has to be run.

**Results** Fig. 5 confirms the validity of the credible intervals provided by the BCTM as the desired 95% level is mostly maintained which is not the case for MLT. MLTs underestimate the uncertainty such that the corresponding confidence intervals are too narrow and ultimately imply coverages that are way below the intended coverage probabilities. Corresponding effect estimates are shown in Supp. Part C.2, revealing that, in addition, MLT estimates show considerably more fluctuation around the true values, which adds a bias component that further deteriorates the performance of corresponding confidence intervals. On the other hand, effect estimates obtained by tramME are very stable, providing strong competitor to the BCTM in this regard, if the true effects are nonlinear while suffering difficulties for effects that are indeed linear (see Supp. Part C.2 for details).

# 5 Applications

We illustrate the versatility of BCTMs and potential advantages over CTMs and other distributional regression models in three applications that differ with respect to the chosen reference distribution and transformation effect types. The first one highlights the applicability of the BCTM in the presence of highly skewed data (Sec. 5.1). The remaining two are BCTMs for (right-censored) survival data in form of a (non-)proportional hazards (NPH) model with random or spatial frailties (Sec. 5.2) and a partial (non-)proportional odds (PO) model (Supp. Part B.3). While not shown here, it is straightforward to derive additional quantities of interest such as quantile curves or odds by transformations of $\hat{h}$. Throughout this section, we use cubic B-spline bases of dimension $D_1 = 20$ for $\boldsymbol{a}$ for univariate splines and dimension $D_1 = 10$ for bivariate splines. We adapt the number of basis functions for the covariate effects in $\boldsymbol{b}(\boldsymbol{x})$ according to subject-matter. As a default, we use IG priors for the smoothing variances unless explicitly stated otherwise.

## 5.1 Framingham heart study

The Framingham Heart Study dataset of (Zhang and Davidian, 2001) contains the cholesterol levels (*cholst*) of 200 patients at three to six different measurement points over the course of up to 10 years along the current *age* and *sex* of each individual with $n = 1044$ observations in total. We fitted various BCTM specifications that differ in their specific form of the transformation function and the chosen hyperprior. All of them are based on

- $\big(\boldsymbol{\vartheta}, \Phi, (\boldsymbol{a}(y)^\top \otimes (1, age), (year, sex))^\top, \pi_\vartheta(\boldsymbol{\vartheta})\big)$: a response-varying VCM for *age* and intercept in $\boldsymbol{a}(y)$, leading to

$$\begin{aligned}\mathbb{P}(cholst \leq y|\boldsymbol{x}) &= \Phi\left(h_1(y) + h_2(y|age) + h_3(y|year) + h_4(y|sex)\right)\\ &= \Phi\left(\boldsymbol{a}(y)^\top \boldsymbol{\gamma}_1 + age \cdot \boldsymbol{a}(y)^\top \boldsymbol{\gamma}_2 + year \cdot \gamma_3 + sex \cdot \gamma_4\right).\end{aligned}$$

- $\big(\boldsymbol{\vartheta}, \Phi, (\boldsymbol{a}(y)^\top \otimes \boldsymbol{b}(\text{age})^\top, (year, sex))^\top, \pi_\vartheta(\boldsymbol{\vartheta})\big)$: a full BCTM for *age*, where $\boldsymbol{a}(y)$ and $\boldsymbol{b}(age)$ contain an intercept, the tensor product is centered around zero and $\boldsymbol{b}(age)$ consists of a 10-dimensional B-splines basis leading to

$$\begin{aligned}\mathbb{P}(cholst \leq y|\boldsymbol{x}) &= \Phi\left(h_1(y|age) + h_2(y|year) + h_3(y|sex)\right)\\ &= \Phi\left((\boldsymbol{a}(y)^\top \otimes \boldsymbol{b}_1(age)^\top)^\top \boldsymbol{\gamma}_1 + year \cdot \gamma_2 + sex \cdot \gamma_3\right).\end{aligned}$$

The default uses IG priors for both models but variants also employ the SD priors as competitors. Furthermore, we considered both models augmented by patient-specific i.i.d. random effects. We benchmark the BCTMs against the Bayesian GAMLSS of Michaelis et al. (2018) based on a skew-t distribution for the responses and predictors $\eta_k$ for all $K = 4$ distributional parameters (location, scale, degrees of freedom, and skewness) given by

$$\eta_k = \beta_{k,0} + x_{\text{sex}}\beta_{k,\text{sex}} + x_{\text{age}}\beta_{k,\text{age}} + x_{\text{year}}\beta_{k,\text{year}}, \quad k = 1, \ldots, K.$$

A variant thereof also contains the patient-specific i.i.d. random effects, see Supp. Tab. B.1

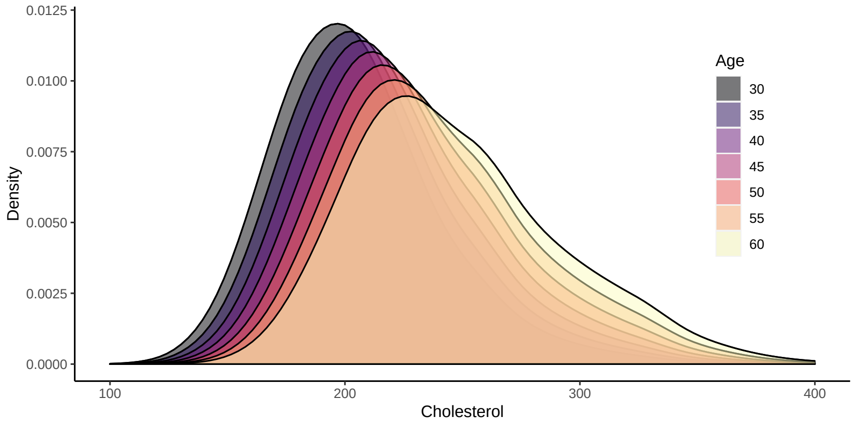

Figure 6: Framingham heart study. VCM with random effects set to zero. Shown are the estimated conditional cholesterol densities for different patient ages in the middle of the study.

for full details on all model specifications.

**Model selection** All models are compared to each other using the DIC, WAIC and log-scores based on 10-fold CV in Supp. Tab. B.2. All criteria favour the tensor product spline BCTM with random effect over the GAMLSS specifically tailored to skewed responses. The inclusion of random effects seems essential for obtaining realistic models while only smaller improvements result from the consideration of tensor products rather than VCMs. Replacing IG priors with SD priors does only yield a small performance improvement for the models without random effects. However, applying the SD prior results in noticeable improvements in effectiveness and stability of the sampler, see Supp. Tab. B.3, B.4.

**Results** Fig. 6 shows estimated conditional densities for different patient ages in the middle of the study ($year = 4$) for the VCM with random effects set to zero. With increasing age, the mode of the conditional distribution is shifted towards higher cholesterol values. Moreover, the estimated conditional densities become more and more right-skewed, indicating the presence of more extreme cholesterol values. On the other hand, the left tail does not change as much. Fig. 7 shows an estimated heat map that was obtained from the tensor product model assuming nonlinear covariate effects. While the general result is similar to the one in Fig. 6, we see a reversal of the trend towards right-skewness at $age \approx 55$.

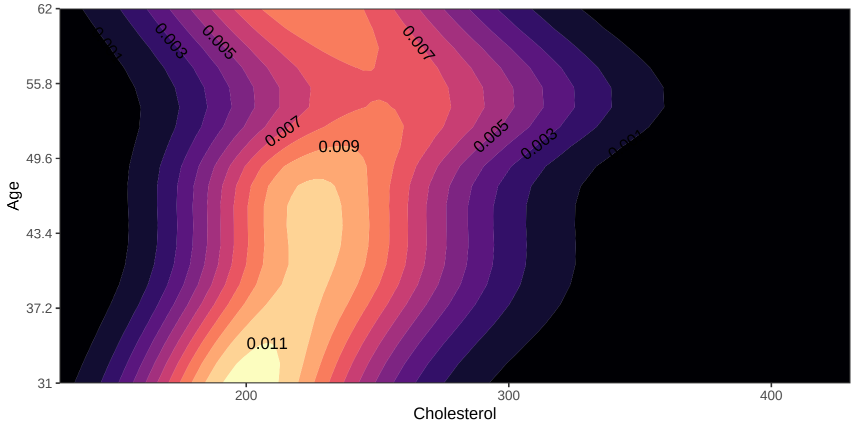

Figure 7: Framingham heart study. Tensor product BCTM with random effects set to zero. Shown are the estimated predictive densities of cholesterol for different patient ages at $year = 4$.

## 5.2 Leukemia survival

The second analysis considers acute myeloid leukemia survival of $n = 1043$ patients (Henderson et al., 2002) with 184 survival times being right-censored. In addition to the impact of the prognostic factors *age*, *sex*, white blood cell count (*wbc*) and the Townsend score (*tpi*), indicating less affluent residential areas for higher values, we investigate spatial patterns in form of the indicator $\boldsymbol{b}(s)$ for 24 administrative regions in North West England.

In a first step, we fitted linear PH models $\big(\boldsymbol{\vartheta}, F_{\text{MEV}}, (\boldsymbol{a}(t)^\top \otimes 1)^\top, \boldsymbol{x}^\top))^\top, \pi_\vartheta(\boldsymbol{\vartheta})\big)$ both without and with random effect for the administrative districts, and where $F_{\text{MEV}}$ denotes the CDF of the minimum extreme value distribution. Next, to account for spatial dependencies through a GMRF, we fit a $\big(\boldsymbol{\vartheta}, F_{\text{MEV}}, (\boldsymbol{a}(t)^\top \otimes 1)^\top, (1 \otimes (\boldsymbol{b}(s)^\top, \boldsymbol{x}^\top))^\top, \pi_\vartheta(\boldsymbol{\vartheta})\big)$ resulting in the spatial PH model

$$\mathbb{P}(Time \leq t|\boldsymbol{x}) = F_{\text{MEV}}(h_1(t) + h_2(s) + h_3(\boldsymbol{x})) = F_{\text{MEV}}(\boldsymbol{a}(t)^\top \boldsymbol{\gamma}_1 + \boldsymbol{x}^\top \boldsymbol{\gamma}_2 + \boldsymbol{b}(s)^\top \boldsymbol{\gamma}_3).$$

As a last expansion, we fitted non-spatial and spatial NPH models for age, i.e. $\big(\boldsymbol{\vartheta}, F_{\text{MEV}}, (\boldsymbol{a}(t)^\top \otimes \boldsymbol{b}(age))^\top, \boldsymbol{x}^\top))^\top, \pi_\vartheta(\boldsymbol{\vartheta})\big)$ and $\big(\boldsymbol{\vartheta}, F_{\text{MEV}}, (\boldsymbol{a}(t)^\top \otimes \boldsymbol{b}(age))^\top, (1 \otimes (\boldsymbol{b}(s)^\top, \boldsymbol{x}^\top))^\top, \pi_\vartheta(\boldsymbol{\vartheta})\big)$, respectively. All model specifications are compared via the WAIC in Tab. 1 and respective posterior mean estimates of the log-negative hazard ratios are presented in Tab. 2. As a

Table 1: Leukemia survial. Shown are the WAIC for all BCTM models, i.e. PH model without random effects (bctm/mlt), with random effects (bctm_re/mlt_re), the spatial PH model (bctm_spat) and the NPH models without and with spatial effect (bctm_nph/bctm_nph_spat).

| Model | bctm | bctm_re | bctm_spat | bctm_nph | bctm_nph_re | bctm_nph_spat |
|---|---|---|---|---|---|---|
| WAIC | 12435 | 12426 | **12420** | 12743 | 12768 | 12766 |

Table 2: Leukemia survial. Estimated posterior of the log negative hazard ratios. The rows correspond to the PH model without random effects (bctm/mlt), with random effects (bctm_re/mlt_re), the spatial PH model (bctm_spat) and the NPH models without and with random/spatial effect (bctm_nph/mlt_nph/bctm_nph_re/bctm_nph_spat).

| Model | *tpi* | *age* | *sex* | *wbc* |
|---|---|---|---|---|
| bctm | 0.112 | 0.556 | 0.027 | 0.203 |
| mlt | 0.102 | 0.552 | 0.035 | 0.204 |
| bctm_re | 0.115 | 0.577 | 0.029 | 0.206 |
| mlt_re | 0.120 | 0.605 | 0.033 | 0.207 |
| bctm_spat | 0.114 | 0.590 | 0.035 | 0.208 |
| bctm_nph | 0.111 | - | 0.028 | 0.198 |
| mlt_nph | 0.141 | - | 0.005 | 0.190 |
| bctm_nph_re | 0.085 | - | 0.043 | 0.202 |
| bctm_nph_spat | 0.110 | - | 0.036 | 0.201 |

baseline check, Tab. 2 also includes esimates from the MLT for which however only the PH model with and without random effects for the districts and the NPH model (without random effects) can be estimated using the R packages **tram** (Hothorn, 2022) and **tramME** (Tamási and Hothorn, 2022). Details on the specifications can be found in Supp. Part B.2. Estimates of these models are similar to the ones of the corresponding BCTM.

Since overall the WAIC favours the spatial PH model (bctm_spat), Fig. 8 shows the resulting estimated conditional survivor functions defined as $S(t|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Time > t|\boldsymbol{X} = \boldsymbol{x}) = 1 - \mathbb{P}(Time \leq t|\boldsymbol{X} = \boldsymbol{x})$ for different Townsend scores (Panel A) accompanied by a depiction of the estimated spatial effect (Panel B). It confirms the findings of Tab. 2, indicating that affluency (lower *tpi*) is associated with higher survival at all times. The spatial effect of association to a district is associated with lower survival for higher values, and is therefore hinting on a lower mortality cluster in the northwest and on a high-risk "belt" running from northeast to southwest. Finally, Tab. B.5 of the Supplement shows the estimated posterior means of the log-negative hazard ratios (collected in $\boldsymbol{\gamma}_2$), medians

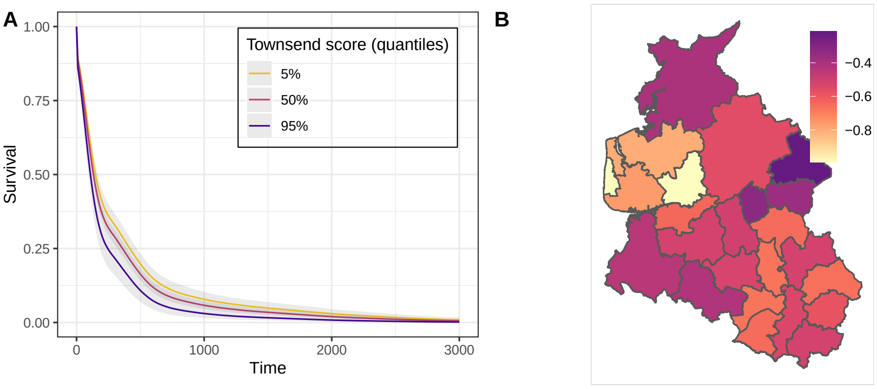

Figure 8: Leukemia survival. (A) Estimated survivor functions with 95% credible interval bands for different quantiles of Townsend score (the lower, the more affluent) for females with $age = 49$ and $wbc = 38.6$(mean) for the first region. (B) Posterior mean spatial frailties (the larger, the higher mean mortality). Results are based on the spatial PH model $\big(\boldsymbol{\vartheta}, F_{\text{MEV}}, (\boldsymbol{a}(t)^\top \otimes 1)^\top, (1 \otimes (\boldsymbol{b}(s)^\top, \boldsymbol{x}^\top))^\top, \pi_{\vartheta}(\boldsymbol{\vartheta})\big)$.

and credible intervals of the same model. Similar to the results in Zhou et al. (2020), we find that *tpi*, *age* and *wbc* are significant risk factors for surviving leukemia.

# 6 Summary and Discussion

Our Bayesian treatment of CTMs based on MCMC is attractive for an assortment of reasons. Sampling-based inference provides posterior samples for coefficients of the conditional transformation function which can be transferred to samples of the cCDF, but also to samples of any quantity of interest that relies on the cCDF. It is straightforward, for example, to obtain point estimates and credible intervals without having to dive into asymptotics. Furthermore, the Bayesian paradigm offers a natural way to impose smoothness penalties on the crucial nonlinear transformation functions which were not available in the original development of likelihood-based CTMs based on Bernstein polynomial bases. In addition, BCTMs are able to resemble and even expand upon models ranging from simple to complex in settings with continuous, discrete and censored data without requiring strong assumptions. Finally, the Bayesian approach will prove beneficial for future model developments where more complex hyperprior structures can be used to enforce desirable model properties, e.g. using Bayesian effect selection priors.

In flat regions of the curve however, the reparameterization of the basis coefficients may lead to weakly identified untransformed $\boldsymbol{\beta}$, resulting in potentially inefficient sample runs. In a Gaussian response model, this issue is explicitlty tackled in McKay Curtis and Ghosh (2011) who use a spike and slab prior directly on the basis coefficients that resulted in zero coefficients for flat regions. In our approach, we considered scale-dependent hyperpriors to counter mixing problems, but expanding such investigations to a wider scope with interactions is certainly an interesting field for future research.

Instead of specifying $F_Z$ a priori, it could also be interesting to include it as an additional free parameter in the estimation process. Among others, Linton et al. (2008); Politis (2013) describe the situation of a "model free" paradigm where the reference distribution is estimated without invoking any predetermined model (but by fully parameterizing the transformation function). This restriction is alleviated by the fact that in theory, arbitrarily complex distributions can be transformed to a basic reference distribution as long as the transformation function $h$ is flexible enough. Abandoning the additivity assumption in $h(y|\boldsymbol{x})$ in favor of e.g. tensor spline interactions however, can become computationally costly and numerically unstable due to the high dimensionality of the resulting basis, but can also be tackled by estimating the reference distribution in conjunction with a simpler structure in the transformation function. The idea of a free $F_Z$ was investigated in a Bayesian setting by Walker and Mallick (1999); Mallick and Walker (2003) who use a Pólya tree prior for a series of (unconditional) semiparametric transformation models. Embedding it in the BCTM framework would result in a potentially very powerful addition to model flexibility.

# SUPPLEMENTARY MATERIAL

**supplement.pdf** This supplement contains the proof of Theorem 2.1 and further results for simulations and applications.

**Code** to reproduce the results from the applications is available on GitHub.

# References

Andrinopoulou, E.-R., Eilers, P. H., Takkenberg, J. J. and Rizopoulos, D. (2018). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines, *Biometrics* **74**(2): 685–693.

Betancourt, M. (2017). Conceptual intro to Hamiltonian Monte Carlo, *arXiv:1701.02434* .

Box, G. E. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society: Series B (statistical methodology)* **26**(2): 211–243.

Cheng, S., Wei, L. and Ying, Z. (1995). Analysis of transformation models with censored data, *Biometrika* **82**(4): 835–845.

Chernozhukov, V., Fernández-Val, I. and Melly, B. (2013). Inference on counterfactual distributions, *Econometrica* **81**(6): 2205–2268.

Eddelbuettel, D. and Balamuta, J. J. (2017). Extending *R* with *C++*: A Brief Introduction to *Rcpp*, *PeerJ Preprints* **5**: e3188v1.

Gneiting, T. (2011). Quantiles as optimal point forecasts, *International Journal of forecasting* **27**(2): 197–207.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, Vol. 43, CRC press.

Henderson, R., Shimakura, S. and Gorst, D. (2002). Modeling spatial variation in leukemia survival data, *Journal of the American Statistical Association* **97**(460): 965–972.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo., *Journal of Machine Learning Research* **15**(1): 1593–1623.

Horowitz, J. L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model, *Journal of the American Statistical Association* **100**(472): 1238–1249.

Hothorn, T. (2017). mlt: Most likely transformations. r package vignette version 0.2-0.

Hothorn, T. (2022). tram: Transformation models. R package vignette version 0.7-0.

Hothorn, T., Kneib, T. and Bühlmann, P. (2014). Conditional transformation models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1): 3–27.

Hothorn, T., Möst, L. and Bühlmann, P. (2018). Most likely transformations, *Scandinavian Journal of Statistics* **45**(1): 110–134.

Hothorn, T. and Zeileis, A. (2021). Predictive distribution modeling using transformation forests, *Journal of Computational and Graphical Statistics* **30**(4): 1181–1196.

James, N. T., Harrell, F. E. and Shepherd, B. E. (2021). Bayesian cumulative probability models for continuous and mixed outcomes, *arXiv:2102.00330* .

Klein, N. and Kneib, T. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression, *Bayesian Analysis* **11**(4): 1071–1106.

Klein, N., Kneib, T., Klasen, S. and Lang, S. (2015). Bayesian structured additive distributional regression for multivariate responses, *Journal of the Royal Statistical Society. Series C: Applied Statistics* **64**(4): 569–591.

Klein, N., Kneib, T., Lang, S., Sohn, A. et al. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany, *The Annals of Applied Statistics* **9**(2): 1024–1052.

Kneib, T., Klein, N., Lang, S. and Umlauf, N. (2019). Modular regression-a lego system for building structured additive distributional regression models with tensor product interactions, *Test* **28**(1): 1–39.

Kneib, T., Silbersdorff, A. and Säfken, B. (2022). Rage against the mean - A review of distributional regression approaches, *Econometrics and Statistics* .

Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econometrica* **46**: 33–50.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrology and Earth System Sciences* **11**(4): 1267–1277.

Lang, S. and Brezger, A. (2004). Bayesian P-splines, *Journal of Computational and Graphical Statistics* **13**(1): 183–212.

Linton, O., Sperlich, S., Van Keilegom, I. et al. (2008). Estimation of a semiparametric transformation model, *The Annals of Statistics* **36**(2): 686–718.

Mallick, B. K. and Walker, S. (2003). A Bayesian semiparametric transformation model incorporating frailties, *Journal of Statistical Planning and Inference* **112**(1-2): 159–174.

McKay Curtis, S. and Ghosh, S. K. (2011). A variable selection approach to monotonic regression with Bernstein polynomials, *Journal of Applied Statistics* **38**(5): 961–976.

Michaelis, P., Klein, N. and Kneib, T. (2018). Bayesian multivariate distributional regression with skewed responses and skewed random effects, *Journal of Computational and Graphical Statistics* **27**(3): 602–611.

Monnahan, C. C. and Kristensen, K. (2018). No-U-Turn Sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages., *PLoS ONE* **13**(5): e0197954.

Möst, L. (2015). *Conditional Transformation Models-Interpretable Parametrisations and Censoring*, Verlag Dr. Hut.

Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics, *in* S. Brooks, A. Gelman, G. Jones and X.-L. Meng (eds), *Handbook of Markov chain Monte Carlo*, 1st edn, Chapman & Hall/CRC, New York, chapter 5, pp. 133–162.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models, *Journal of the Royal Statistical Society: Series A (General)* **135**(3): 370–384.

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems, *Mathematical programming* **120**(1): 221–259.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007). A general framework for the parametrization of hierarchical models, *Statistical Science* **22**(1): 59–73.

Pericchi, L. (1981). A Bayesian approach to transformations to normality, *Biometrika* **68**: 35–43.

Politis, D. N. (2013). Model-free model-fitting and predictive distributions, *Test* **22**(2): 183–221.

Pya, N. (2010). *Additive models with shape constraints*, PhD thesis, University of Bath.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models, *Statistics and Computing* **25**(3): 543–559.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(3): 507–554.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, CRC.

Song, X.-Y. and Lu, Z.-H. (2012). Semiparametric transformation models with Bayesian P-splines, *Statistics and Computing* **22**(5): 1085–1098.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4): 583–639.

Tamási, B. and Hothorn, T. (2022). tramME: Mixed-effects transformation models using template model builder, *R Journal* pp. Epub–ahead.

Tang, N., Wu, Y. and Chen, D. (2018). Semiparametric Bayesian analysis of transformation linear mixed models, *Journal of Multivariate Analysis* **166**: 225–240.

Umlauf, N., Klein, N. and Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale, and shape (and beyond), *Journal of Computational and Graphical Statistics* **27**: 612–627.

Waldmann, E., Kneib, T., Yue, Y. R., Lang, S. and Flexeder, C. (2013). Bayesian semiparametric additive quantile regression, *Statistical Modelling* **13**(3): 223–252.

Walker, S. and Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model, *Biometrics* **55**(2): 477–483.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11**(Dec): 3571–3594.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression, *Statistics & Probability Letters* **54**(4): 437–447.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data, *Biometrics* **57**(3): 795–802.

Zhou, H., Hanson, T. and Zhang, J. (2020). spBayesSurv: Fitting Bayesian spatial survival models using R, *Journal of Statistical Software* **92**(9): 1–33.