Statistical Modelling xxxx; **xx(x)**: 1–23

Bayesian discrete conditional transformation models

Manuel Carlan[1] **and Thomas Kneib**[1]
[1]Chair of Statistics, Faculty of Business and Economic Sciences, University of Göttingen, Germany

Abstract: We propose a novel Bayesian model framework for discrete ordinal and count data based on conditional transformations of the responses. The conditional transformation function is estimated from the data in conjunction with an a priori chosen reference distribution. For count responses, the resulting transformation model is novel in the sense that it is a Bayesian fully parametric yet distribution-free approach that can additionally account for excess zeros with additive transformation function specifications. For ordinal categoric responses, our cumulative link transformation model allows the inclusion of linear and non-linear covariate effects that can additionally be made category-specific, resulting in (non-)proportional odds or hazards models and more, depending on the choice of the reference distribution. Inference is conducted by a generic modular Markov chain Monte Carlo algorithm where multivariate Gaussian priors enforce specific properties such as smoothness on the functional effects. To illustrate the versatility of Bayesian discrete conditional transformation models, applications to counts of patent citations in the presence of excess zeros and on treating forest health categories in a discrete partial proportional odds model are presented.

Key words: discrete responses, Bayesian transformation models, penalised splines, overdispersion, zero-inflation, partial proportional odds

Received April 2021; **revised** March 2022; **accepted** August 2022

1 Introduction

Discrete data commonly occur in almost every scientific area. In this article, we focus on the two relevant cases of count data and ordinal data as special instances of discrete response structures. Before the advent of generalized linear models (GLM, Nelder and Wedderburn, 1972), the peculiarities of count data were either ignored or treated simply by log transformations (Sokal and Rohlf, 1981). Then, the standard modeling approach for count data $Y \in \{0, 1, 2, \ldots\}$ became Poisson regression, $Y|\boldsymbol{x} \sim \text{Po}(\lambda_{\boldsymbol{x}})$. Since the Poisson distribution often turned out to be too simplistic for many applications, more advanced regression models were introduced as described, for example, by Cameron and Trivedi (1998), Winkelmann (2008) and Hilbe (2011) for negative binomial regression, $Y|\boldsymbol{x} \sim \text{NB}(\lambda_{\boldsymbol{x}}, \nu)$ accounting for potential overdispersion. Generalized additive models (GAM, Hastie and Tibshirani, 1990) unify these model types into one framework and drop the

Address for correspondence: Thomas Kneib, Chair of Statistics, Faculty of Business and Economic Sciences, University of Göttingen, Humboldtallee 3, D–37073 Göttingen, Germany.
E-mail: tkneib@uni-goettingen.de

10.1177/1471082X221114177

linearity assumption for the regression predictor. They require a fixed response distribution that belongs to the exponential family.

Similar to counts, ordered categorical data $Y \in \{1, \ldots, c+1\}$ occur in a manifold of scientific disciplines such as medicine or the social sciences. A researcher in medicine, for example, may want to distinguish between different kinds of infection grades, while an ecologist could be interested in measuring forest health in terms of defoliation categories. Exploiting the natural ordering in these kinds of data is firmly established in the statistical community by cumulative link models as shown in McCullagh (1980). Prominent versions are the discrete proportional odds model and the discrete proportional hazards model (Tutz, 2011). In its simplest form, the cumulative link model is given by $\pi_r = P(Y = r) = F(\gamma_r - \boldsymbol{x}^T\boldsymbol{\beta}) - F(\gamma_{r-1} - \boldsymbol{x}^T\boldsymbol{\beta}), r = 1, \ldots, c+1$ with some pre-specified cumulative distribution function F or equivalently $P(Y \leq r) = F(\gamma_r - \boldsymbol{x}^T\boldsymbol{\beta}) = \pi_1 + \pi_2 + \ldots + \pi_r$, where $\sum_{r=1}^{c+1} \pi_r = 1$ is required and the ordering $-\infty \equiv \gamma_0 < \ldots, < \gamma_{c+1} \equiv \infty$ needs to be obliged. It is possible to include category-specific regression effects $\boldsymbol{x}^T\boldsymbol{\beta}_r$, resulting in a (linear) non-proportional odds (Peterson and Harrell, 1990) or the non-proportional hazards model, depending on the choice of F.

The dissemination of Markov chain Monte Carlo (MCMC) simulation techniques led to the development of Bayesian analogues for established models in the form of Bayesian GLMs (Dey et al., 2000) with many extensions, for example, by Frühwirth-Schnatter and Wagner (2006), Frühwirth-Schnatter et al. (2009) , Rodrigues (2003) and the Bayesian GAM (Brezger and Lang, 2006). Ghosh et al. (2006) describe a Bayesian treatment of zero-inflated regression models, and Klein et al. (2015a) introduce zero-inflated and overdispersed count data to the framework of Bayesian structured additive distributional regression (Klein et al., 2015b). In a non-transformation environment, Lavine and Mockus (1995) and Dunson (2005), among others, apply a (strictly) isotonic regression function for count responses on the basis of a Dirichlet process mixture prior.

To bridge the gap between discrete ordinal and count regression models, we consider count data as ordinal categorical data with a very high number of intercept thresholds that, however, are not estimated but rather are fixed by design at all non-negative integers. Methodologically, both approaches are unified by the idea of a direct parametrization of the transformation function. Similar to Siegfried and Hothorn (2020), we treat the smooth parametrization of the thresholds as the defining element of the count transformation approach used in this article. While overdispersion is absorbed by the smooth transformation of the counts, we supplement the model with a second component that explicitly accounts for eventual zero inflation. For a discussion of the connection between (binary) regression and transformation models, see Doksum and Gasko (1990).

To summarize, in this article, we aim to do the following:

- Propose a Bayesian approach for count transformation models based on flexible transformation functions that are inferred from the data, which–in its simplest form with linear covariate shift effects–results in a distribution-free yet interpretable model framework for count data that automatically accounts for over- and underdispersion in the response distribution,
- Account for excess zeros in two-component mixtures models,
- Propose a Bayesian approach for cumulative link transformation models with Bayesian proportional odds and proportional hazards models as special cases,
- Allow for the inclusion of category-specific effects, resulting in non-proportional transformation model types,

- Combine both model types into the class of Bayesian discrete conditional transformation models (BDCTM) and establish it as an extension of Bayesian conditional transformation models (BCTM) for continuous responses,
- Supplement all models with non-linear, possibly high-dimensional covariate effects and interactions, and
- Illustrate BDCTM's capability in the presence of count and categoric data in two applications.

The rest of this article is structured as follows: Section 2 introduces the model class we refer to as BDCTM with a preliminary discussion of its building blocks. Section 3 contains a description of posterior estimation. A simulation study evaluating BDCTM's performance in a count data setting is presented in Section 4. Section 5 features an application on patent citation counts and an application on forest health categories. We conclude in Section 6.

2 Bayesian discrete conditional transformation models

In what follows, we introduce BDCTM as a model class that represents a novel approach to the direct estimation of the conditional distribution function $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|x)$ based on an independent sample of discrete responses $Y_1, \ldots, Y_n$ conditional on covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. We broadly distinguish between cases of count data and ordered categorical data with a finite sample space, which have to be addressed by different assumptions on the sampling distribution and different basis functions.

Let y be an observation of a count or ordered categorical response variable Y and let $\boldsymbol{x}^T = (x_1, \ldots, x_q)$ be a vector of observed explanatory variables. Moreover, let F_Z be the cumulative distribution function of an a priori chosen reference distribution, linking a discrete and monotonically increasing transformation function $h(y|\boldsymbol{x})$ to the conditional distribution function $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|\boldsymbol{x})$ via the connection

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|\boldsymbol{x}) = P(Y \leq y|\boldsymbol{x}) = F_Z(h(y|\boldsymbol{x})). \tag{2.1}$$

The responses are transformed towards the reference distribution conditionally on $\boldsymbol{x}$ by means of the transformation function $h(y|\boldsymbol{x})$. Through allowing different complexities of the transformation function $h(y|\boldsymbol{x})$, BDCTM is able to resemble and expand on established models for count and ordinal data without requiring a fixed response distribution. The encompassing goal of all models described in this article is to obtain an estimate of the distribution function $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ by means of estimating $h(y|\boldsymbol{x})$. In contrast to Bayesian CTMs for continuous responses, the transformation function will no longer be bijective since a continuous reference distribution is linked to the CDF of a discrete response variable.

We proceed with discussing each of the components of a BDCTM in more detail. Section 2.1 introduces the basic structure assumed for the transformation functions. Sections 2.2 and 2.3 present model variants for count data and ordinal responses, respectively, while Section 2.4 discusses a generic basis function representation for the transformation functions. Section 2.5 introduces the corresponding prior assumptions, Section 2.6 discusses partial contributions to the transformation function, and Section 2.7 contemplates on the relevance of the choice of the reference distribution.

2.1 Transformation functions

Similar to Hothorn et al. (2014), we assume an additive decomposition on the scale of the transformation function into J partial transformation functions

$$h(y|\boldsymbol{x}) = \sum_{j=1}^{J} h_j(y|\boldsymbol{x}), \tag{2.2}$$

where $h_j(y|\boldsymbol{x})$ are response-covariate interactions that are monotone only in direction of y. We denote partial transformation functions that depend only on the covariates simply by $h(\boldsymbol{x})$. A simple transformation model, for example, is obtained by setting $h_1(y|\boldsymbol{x}) = h_Y(y)$ and $h_2(y|\boldsymbol{x}) = h(\boldsymbol{x})$. We explicitly allow the inclusion of linear and non-linear covariate effects, that is,

$$h(\boldsymbol{x}) = \boldsymbol{z}^T\boldsymbol{\beta} + f_1(\boldsymbol{\nu}) + \ldots + f_L(\boldsymbol{\nu}), \tag{2.3}$$

where in $\boldsymbol{x} = (\boldsymbol{z}^T, \boldsymbol{\nu}^T)^T$, $\boldsymbol{z}$ contains all covariates associated with linear effects and $\boldsymbol{\nu}$ contains covariates with assumed non-linear effects $f_1, \ldots, f_L$.

2.2 Count transformation models

We distinguish between two related model types for count data: simple shift count transformation models that are able to deal with overdispersion and two-component mixture transformation models that can additionally deal with excess zeros.

Mean-shift count transformation models: Regular count transformation models are defined by shifts of the non-linear baseline transformation function h_Y:

$$F_{Y|X=\boldsymbol{x}}(y|\boldsymbol{x}) = F_Z(h_Y(\lfloor y \rfloor) - h(\boldsymbol{x})) \tag{2.4}$$

where $\lfloor y \rfloor$ denotes the floor function returning the greatest integer less than or equal to y. Since all moments besides the conditional mean (which is shifted by $h(\boldsymbol{x})$) are captured solely by $h_Y(\lfloor y \rfloor)$, independently of the covariates, the resulting model is not affected by over- or underdispersion. Model (2.4) is similar to a regular linear transformation model, but the application of the floor function leads to jumps at the respective integers, such that the transformation function $h_Y(y)$ is only evaluated at the distinctive response values $y \in \{0, 1, 2, \ldots\}$ and, as a consequence, the overall transformation is no longer invertible. The likelihood-based version of this model type restricted to linear covariate shifts was discussed in detail in Siegfried and Hothorn (2020).

Two-component mixture count transformation models: Besides over- and underdispersion, count data often come with an excess number of zeros, which needs to be accomodated in the model. One possibility is to add a second component to the linear transformation function that captures zeros

(Hothorn et al., 2018). A transformation function in that vein can be depicted as:

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|\boldsymbol{x}) = F_Z(h_Y(\lfloor y \rfloor) - h(\boldsymbol{x}) + \mathbb{1}(y=0)(\beta_0 - h_0(\tilde{\boldsymbol{x}}))), \tag{2.5}$$

where $h_0(\tilde{\boldsymbol{x}})$ and $h(\boldsymbol{x})$ can consist of different linear and non-linear effects of different sets of covariates.–This two component mixture transformation model resembles a hurdle model with hurdle at zero, where the probability of an excess zero is perceived as the mean-shifted deviation from a regular count transformation model at $y = 0$:

$$P(Y = 0|\boldsymbol{X} = \boldsymbol{x}) = F_Z(h_Y(0) - h(\boldsymbol{x}) + (\beta_0 - h_0(\tilde{\boldsymbol{x}}))). \tag{2.6}$$

The process generating non-zeros in this case is not explicitly truncated but stems from a transformation function that excludes the zeros.

All count transformation functions of this type have in common that they act on the floor function $\lfloor y \rfloor$, resulting in step functions in direction of y and thus the desired discrete distribution functions. Comparing this to the ordinal response models discussed in the next section, count data transformation models can also be considered as introducing a latent, continuous scale, implicitly determined by the transformation function, with a large number of pre-specified thresholds corresponding to the non-negative integers.

2.3 Cumulative link transformation models

For ordered categorical data, we distinguish between cumulative models with and without category-specific shifts. From a transformation perspective, the latter are modeled in terms of response-covariate interactions that can be linear or non-linear in direction of the respective covariate.

Proportional models: The simplest cumulative transformation model is:

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_r|\boldsymbol{x}) = F_Z(h_Y(y_r) - h(\boldsymbol{x})), \tag{2.7}$$

where the term $h(\boldsymbol{x})$, which is independent of the category r, constitutes the log-odds ratio to $h(0)$ or the log-hazard ratio in model types (2.4) and (2.7), depending on the choice of reference distribution.

Non-proportional models: Models of type (2.7) can be generalized by a category-specific shift resulting in the following model:

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_r|\boldsymbol{x}) = F_Z(h_Y(y_r) + h_r(\boldsymbol{x})), \tag{2.8}$$

where $h_r(\boldsymbol{x})$ induces the category-specific shifts, resulting in linear or non-linear non-proportional odds or hazards models depending on F_Z and on whether $h_r(\boldsymbol{x})$ consists of linear or non-linear effects. Partial proportional models as shown in the application in Section 5.2 consist of a mixture of proportional and non-proportional effects. The reparameterization illustrated in the following section guarantees that the implied probabilities $P(Y = r) = F_Z(\gamma_r - h_r(\boldsymbol{x})) - F_Z(\gamma_{r-1} - h_{r-1}(\boldsymbol{x}))$ are always positive.

2.4 A generic joint basis

We assume that each of the J partial transformation functions can be approximated by a linear combination of basis functions $\boldsymbol{c}_j$ such that

$$h_j(y|\boldsymbol{x}) = \boldsymbol{c}_j(y, \boldsymbol{x})^T \boldsymbol{\gamma}_j,$$

where $\boldsymbol{\gamma}_j$ is a vector of basis coefficients. Based on the additivity assumption in (2.2), the complete conditional transformation function can be denoted as

$$h(y|\boldsymbol{x}) = \boldsymbol{c}(y, \boldsymbol{x})^T \boldsymbol{\gamma} \tag{2.9}$$

with joint basis

$$\boldsymbol{c}(y, \boldsymbol{x}) = (\boldsymbol{c}_1(y, \boldsymbol{x})^T, \dots, \boldsymbol{c}_J(y, \boldsymbol{x})^T)^T$$

and $\boldsymbol{\gamma}$ contains all partial basis coefficient vectors,

$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_J^T)^T. \tag{2.10}$$

This allows us to write all discrete conditional transformation models treated in this article in the general form:

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(\boldsymbol{c}(y, \boldsymbol{x})^T \boldsymbol{\gamma}). \tag{2.11}$$

We call models of type (2.11) Bayesian discrete transformation models (BDCTM). They can be conceived as extensions of the versatile model class of BCTM for continuous responses that were introduced by Carlan et al. (2020), taking the additional challenges arising from discrete responses into account. In this tradition, a BDCTM is fully specified by a reference distribution F_Z, the joint basis $\boldsymbol{c}(y, \boldsymbol{x})$ and a vector of basis coefficients $\boldsymbol{\gamma}$ together with suitable priors, which are introduced in the next section. The rest of this section discusses the generic basis that is used by the BDCTM in greater detail.

Let $\boldsymbol{a}_j$ denote a basis transformation of y with dimension D_1, collecting evaluated basis functions $B_{j1d_1}(y)$, $d_1 = 1, \dots, D_1$, and let $\boldsymbol{b}_j$ denote a basis transformation of x with dimension D_2 collecting evaluated basis functions $B_{j2d_2}(x)$, $d_2 = 1, \dots, D_2$. The resulting effects are approximated by the following linear combinations:

$$h_j(y) = \sum_{d_1=1}^{D_1} \gamma_{j1d_1} B_{j1d_1}(y) = \boldsymbol{a}(y)^T \boldsymbol{\gamma}_{j1}, \quad h_j(x) = \sum_{d_2=1}^{D_2} \gamma_{j2d_2} B_{j2d_2}(x) = \boldsymbol{b}_j(x)^T \boldsymbol{\gamma}_{j2},$$

where $\boldsymbol{\gamma}_{j1} = (\gamma_{j11}, \dots, \gamma_{j1D_1})^T$ and $\boldsymbol{\gamma}_{j2} = (\gamma_{j21}, \dots, \gamma_{j2D_2})^T$ are partially reparameterized versions of the vectors of corresponding basis coefficients $\boldsymbol{\beta}_{j1}$ and $\boldsymbol{\beta}_{j2}$. The conditional transformation approach commonly involves response-covariate interactions (e.g., model types (2.6) and (2.8)), which is why we parametrize each partial transformation function generically as

$$
\begin{aligned}
h_j(y|x) &= \boldsymbol{c}_j(y,x)^T\boldsymbol{\gamma}_j = (\boldsymbol{a}_j(y)^T \otimes \boldsymbol{b}_j(x)^T)^T\boldsymbol{\gamma}_j \\
&= \sum_{d_1=1}^{D_1}\sum_{d_2=1}^{D_2} \gamma_{j,d_1d_2} B_{d_1}(y) B_{d_2}(x),
\end{aligned}
\tag{2.12}
$$

where the Kronecker product forms parametric interactions between the evaluated basis functions, and $\boldsymbol{\gamma}_j$ is a basis vector of dimension $D = D_1 D_2$. A collection of special cases can be found in Section 2.6.

We require all transformation functions to be strictly monotonically increasing solely in the direction of y but not in direction of the explanatory variables such that $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_j|\boldsymbol{x}) < F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_{j+1}|\boldsymbol{x})$ for all $y_j < y_{j+1}$. This property needs to be accomodated in the basis. For this, we adopt the approach of Pya and Wood (2015) for monotonically increasing smooth functions. The vector $\boldsymbol{\gamma}_j$ is reparameterized as $\boldsymbol{\gamma}_j = \boldsymbol{\Sigma}_j\tilde{\boldsymbol{\beta}}_j$, where $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_{D_1} \otimes \boldsymbol{I}_{D_2}$ and $\boldsymbol{\Sigma}_{D_1}$ is given by the lower triangular matrix of size D_1 such that $\Sigma_{D_1,kl} = 0$ if $k < l$ and $\Sigma_{D_1,kl} = 1$ if $k \geq l$. The vector $\tilde{\boldsymbol{\beta}}_j$ of dimension $D = D_1 D_2$ contains a mixture of unexponentiated and exponentiated β-coefficients given by

$$
\tilde{\boldsymbol{\beta}}_j = (\beta_{j,11}, \ldots, \beta_{j,1D_2}, \exp(\beta_{j,21}), \ldots \exp(\beta_{j,2D_2}), \ldots, \exp(\beta_{j,D_1D_2}))^T. \tag{2.13}
$$

and $\boldsymbol{I}_{D_2}$ is an identity matrix of size D_2. An unconditional transformation function $h_Y(y)$ is obtained by setting $D_2 = 1$ and a function of type $h(\boldsymbol{x})$ is obtained by setting $D_1 = 1$.

The vector of basis coefficients for the whole conditional transformation function $h(y|\boldsymbol{x})$ is given by $\boldsymbol{\gamma} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \ldots, \tilde{\boldsymbol{\beta}}_J^T)^T$ is based on $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_J^T)^T$. Matrix $\boldsymbol{\Sigma}$ is block diagonal with $\boldsymbol{\Sigma}_j$ as diagonal elements.

Of course, other basis specification could be employed to set up BDCTMs, as long as monotonicity along y is ensured. For example, the increasing splines considered in continuous ordinal regression (Manuguerra and Heller, 2010) would be a potential alternative. We rely on Bayesian P-splines and their tensor product interactions since these have been extensively studied in Bayesian structured additive regression and enable efficient and stable computations.

2.5 Priors

We adopt the principle of Bayesian P-splines (Lang and Brezger, 2004) and assume partially improper multivariate Gaussian priors for the unconstrained vectors $\boldsymbol{\beta}_{j1}$ and $\boldsymbol{\beta}_{j2}$ (the reparameterized vectors $\boldsymbol{\gamma}_{j1}$ and $\boldsymbol{\gamma}_{j2}$ are based on) such that

$$
\begin{aligned}
p(\boldsymbol{\beta}_{j1}|\tau_{j1}^2) &\propto \left(\frac{1}{\tau_{j1}^2}\right)^{\frac{\mathrm{rk}(\boldsymbol{K}_{j1})}{2\tau_{j1}^2}} \exp\left(-\frac{1}{2\tau_{j1}^2}\boldsymbol{\beta}_{j1}^T\boldsymbol{K}_{j1}\boldsymbol{\beta}_{j1}\right), \\
p(\boldsymbol{\beta}_{j2}|\tau_{j2}^2) &\propto \left(\frac{1}{\tau_{j2}^2}\right)^{\frac{\mathrm{rk}(\boldsymbol{K}_{j2})}{2\tau_{j2}^2}} \exp\left(-\frac{1}{2\tau_{j2}^2}\boldsymbol{\beta}_{j2}^T\boldsymbol{K}_{j2}\boldsymbol{\beta}_{j2}\right),
\end{aligned}
\tag{2.14}
$$

where τ_{j1}^2 and τ_{j2}^2 are marginal smoothing variances, rk$(\cdot)$ is the rank of a matrix, and $\boldsymbol{K}_{j1}$ and $\boldsymbol{K}_{j2}$ are potentially rank deficient prior precision matrices. The generic formulation of the precision matrix associated with $\boldsymbol{\gamma}_j$ is given by

$$\boldsymbol{K}_j = \frac{1}{\tau_{j1}^2}(\boldsymbol{K}_{j1} \otimes \boldsymbol{I}_{D_2}) + \frac{1}{\tau_{j2}^2}(\boldsymbol{I}_{D_1} \otimes \boldsymbol{K}_{j2}),$$

where precision matrices $\boldsymbol{K}_{j1}$ and $\boldsymbol{K}_{j2}$ control the penalty in the direction of y and x respectively. For unconditional transformation functions or pure covariate functions, $\boldsymbol{K}_{j1}$ and $\boldsymbol{K}_{j2}$ are respectively set to $\boldsymbol{0}$ such that only the prior precision of the corresponding effect is used. Specific choices are discussed in the next section. The model precision matrix $\boldsymbol{K}$ is given as the block diagonal matrix with matrices $\boldsymbol{K}_j$ as diagonal elements.

The smoothing variances τ_{j1}^2 and τ_{j2}^2 are associated with inverse gamma priors, $\tau_{j1}^2 \sim \text{IG}(a_{j1}, b_{j1})$ and $\tau_{j2}^2 \sim \text{IG}(a_{j2}, b_{j2})$. All model parameters are collected in $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J, \tau_{11}^2, \tau_{12}^2 \dots, \tau_{J1}^2, \tau_{J2}^2)^T$ with joint prior $p(\boldsymbol{\vartheta})$

2.6 Partial transformations

We start this section by introducing the two types of basis functions we use in $\boldsymbol{a}$, depending on whether Y is a count variable or discrete ordinal followed by a brief discussion of choices for $\boldsymbol{b}$ together with suitable precision matrices.

Smooth basis for count transformations: In case of a count response $Y \in \{0, \dots\}$, $\boldsymbol{a}$ consists of B-spline basis functions B_{d_1} i.e. $\boldsymbol{a}_j(y) = (B_1(y), \dots, B_{D_1}(y))^T$. It may be useful to parametrize the transformation function on the log-scale, i.e $\boldsymbol{a}_j(\log(y))$ or $\boldsymbol{a}_j(\log(y+1))$, where especially the latter can be beneficial numerically if there are many small and some large counts. Smooth monotonic effects of a count transformation subject to the reparameterization in (2.13) are supplemented with a penalty matrix $\boldsymbol{K}_{j1} = \boldsymbol{D}_1^T \boldsymbol{D}_1$ based on a $(D_1 - 2) \times D_1$ partial first-difference matrix $\boldsymbol{D}_1$ that is zero except that $D_{i,i+1} = -D_{i,i+2} = 1$ for $i = 1, \dots, D_1 - 2$ to achieve shrinkage towards a straight line (Pya and Wood, 2015).

Discrete basis for ordinal categorical data For ordered categorical responses, $Y \in \{1, \dots, c+1\}$ we assign one parameter to each category except for the reference category $c+1$ (Hothorn et al., 2018). As a basis, we use the unit vector $\boldsymbol{e}_c$ of length c, i.e. $\boldsymbol{a}_j(y_r) = \boldsymbol{e}_c(r)$, where

$$Y = r \iff \boldsymbol{e}_c(r) = (0, \dots, 1, \dots, 0)^T, \quad r = 1, \dots, c. \tag{2.15}$$

The corresponding precision matrix is $\boldsymbol{K}_{j1} = \boldsymbol{0}$.

Bases for covariates effects For covariate effects we allow linear bases $\boldsymbol{b}_j(\boldsymbol{z}) = (z_1, \dots, z_p)^T$ together with precision matrix $\boldsymbol{K}_{j2} = \boldsymbol{0}$ and B-spline bases for non-linear effects $\boldsymbol{b}_j(\nu) =$

$(B_1(\nu), \dots, B_{D_2}(\nu))^T$ with a second order random-walk precision matrix $\boldsymbol{K}_{j2}$. All bases involving B-spline basis functions can be centered around zero for identification purposes.

Transformation random effects $h_j(x) = \beta_g$ are based on the grouping indicator $g \in \{1, \dots, G\}$. The corresponding G-dimensional basis vector $\boldsymbol{b}_j(g)$ has entry one if x belongs to group g and zero otherwise. We set $\boldsymbol{K}_j = \boldsymbol{I}_G$ for i.i.d. random effects. Regular non-monotonic tensor splines as used in the forest health application in Section 5.2 can be retrieved by using the specification in (2.12) and setting $\boldsymbol{\gamma}_j = \boldsymbol{\beta}_j$.

2.7 Reference distribution

In the context of discrete conditional transformation models, the reference distribution function F_Z plays the role of the inverse link function controlling the interpretational scale of the impact of the explanatory variables. While it can be chosen arbitrarily in theory, we concentrate on distributions with log-concave densities for F_Z to guarantee uniqueness of the maximum likelihood estimate, which usually will also imply unimodality of the posterior. Furthermore, it is advised to consider characteristics such as right-skewness or the support of the count data distribution in the selection process. Prominent choices for F_Z are

- $F_{\text{SL}}(z) = (1 + \exp(-z))^{-1}$, that is, the standard logistic distribution,
- $\Phi(z)$, that is, the standard normal distribution, and
- $F_{\text{MEV}}(z) = 1 - \exp(-\exp(z))$, that is, the minimum extreme value distribution

This results in logit, probit or cloglog interpretations of the covariate effects. Setting $F_Z = F_{\text{SL}}$, for example, results in the discrete proportional odds model and $F_Z = F_{\text{MEV}}$ results in the proportional hazards model, with $h(\boldsymbol{x})$ becoming the log-odds ratio or the log-hazard ratio to $h(\boldsymbol{0})$, respectively (Hothorn et al., 2018).

To reflect specific properties of the data-generating process, other link functions that have been considered in the context of GLM, such as skew-logistic or t-distributed link functions to reflect strong asymmetry or heavy tails, may be considered. However, given the flexibility of the transformation function, we do not expect large gains from such specifications since both asymmetry and tail behaviour should be taken up by the transformation function, leaving only a small potential for improving the fit via the link function. We therefore suggest to stick to the defaults and to select the reference distribution according to preferences on model interpretation.

2.8 Transformation probability mass functions

In this section, we introduce the transformation probability mass functions (PMFs) resulting from the different sampling assumptions that come with count and ordinal categoric data as well as the resulting transformation likelihoods. To emphasize that $\boldsymbol{\gamma}$ are partially non-linear reparameterizations of $\boldsymbol{\beta}$, we write $\boldsymbol{\gamma}(\boldsymbol{\beta})$. Following Hothorn et al. (2018), the log-transformation PMF of a conditionally independent (count) response Y with unbounded support $Y \in \{0, 1, \dots, \}$ is given by

$$\log(f_Z(y|\boldsymbol{\beta})) = \begin{cases} \log[F_Z(\boldsymbol{c}(y_k, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta}))] & k = 1 \\ \log[F_Z(\boldsymbol{c}(y_k, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta})) - F_Z(\boldsymbol{c}(y_{k-1}, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta})] & k > 1. \end{cases} \tag{2.16}$$

In case of an ordinal categorical response with bounded support $Y \in \{y_1, \dots, y_{c+1}\}$, the corresponding conditional distribution function needs to take the additional constraint for the reference category $c + 1$, $P(Y \leq y_{c+1} | \boldsymbol{X} = \boldsymbol{x}) = F_Z(h(y_{c+1} | \boldsymbol{X} = \boldsymbol{x})) = 1$ into account. The transformation PMF is then given by

$$f_Z(y|\boldsymbol{\beta}) = \begin{cases} [F_Z(\boldsymbol{c}(y_k, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta})))] & k = 1 \\ [F_Z(\boldsymbol{c}(y_k, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta})) - F_Z(\boldsymbol{c}(y_{k-1}, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta}))] & k = 2, \dots, c \\ [1 - F_Z(\boldsymbol{c}(y_c, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta}))] & k = c + 1. \end{cases} \tag{2.17}$$

With the convention $F_Z(h(y_0)) = F_Z(h(-\infty)) = 0$ and $F_Z(h(y_{c+1})) = F_Z(h(\infty)) = 1$, the conditional PMF simplifies to

$$f_Z(y_k|\boldsymbol{\beta}) = F_Z(\boldsymbol{c}(y_k, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta}) - F_Z(\boldsymbol{c}(y_{k-1}, \boldsymbol{x})^T \boldsymbol{\gamma}(\boldsymbol{\beta})) \tag{2.18}$$

encompassing count and ordered categoric models in a unified framework (Hothorn et al., 2018). Based on (2.18), the transformation log-likelihood for independent observations $(y_i, \boldsymbol{x}_i)$, $i = 1, \dots, n$ is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log(F_Z(\boldsymbol{c}(y_i, \boldsymbol{x}_i)^T \boldsymbol{\gamma}(\boldsymbol{\beta})) - F_Z(\boldsymbol{c}(y_i - 1, \boldsymbol{x}_i)^T \boldsymbol{\gamma}(\boldsymbol{\beta}))).$$

The likelihood is chosen according to the discrete response structure only, while the transformation function determines whether excess zeros are accounted for or if the category-specific effects are included, for example. With all building blocks in mind, a BDCTM can be fully specified by the set $\{\boldsymbol{\vartheta} | F_Z, \boldsymbol{c}, \pi_{\vartheta}(\cdot)\}$ of unknown model parameters $\boldsymbol{\vartheta}$, given a choice for the basis $\boldsymbol{c}$, the reference distribution F_Z and the joint prior π_{ϑ} (Carlan et al., 2020).

3 Posterior inference

For Bayesian inference, we rely on MCMC simulation techniques. We sketch the most relevant parts of the algorithm in this section.

Update of the basis coefficients: The log-full conditional of the basis coefficients (up to an additive constant) is given by

$$\log(p(\boldsymbol{\beta}|\cdot)) \propto l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{K} \boldsymbol{\beta},$$

where the second term arises from the multivariate Gaussian prior. The gradient of the unnormalized log-posterior is needed for inference and is given by

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{f_Z(\boldsymbol{c}(y_i, \boldsymbol{x}_i)^T \boldsymbol{\Sigma} \tilde{\boldsymbol{\beta}}) \boldsymbol{c}(y_i, \boldsymbol{x}_i)^T \boldsymbol{\Sigma} \boldsymbol{C} - f_Z(\boldsymbol{c}(y_i - 1, \boldsymbol{x}_i)^T \boldsymbol{\Sigma} \tilde{\boldsymbol{\beta}}) \boldsymbol{c}(y_i - 1, \boldsymbol{x}_i)^T \boldsymbol{\Sigma} \boldsymbol{C}}{F_Z(\boldsymbol{c}(y_i, \boldsymbol{x}_i)^T \boldsymbol{\Sigma} \tilde{\boldsymbol{\beta}}) - F_Z(\boldsymbol{c}(y_i - 1, \boldsymbol{x}_i)^T \boldsymbol{\Sigma} \tilde{\boldsymbol{\beta}})} - \boldsymbol{K} \boldsymbol{\beta},$$

where $\boldsymbol{C}$ is a diagonal matrix of size D with entries

$$C_{dd} = \begin{cases} 1 & \text{if } \tilde{\beta}_d = \beta_d \\ \exp(\beta_d), & \text{otherwise.} \end{cases}$$

Strong dependencies among the variables (which are partly due to the monotonicity restriction) complicate the sampling from the posterior distribution. This is further impeded by the mixed linear-non-linear dependence of the transformation function on $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$, respectively. Therefore, we use the No-u-turn sampler (NUTS, Hoffman and Gelman, 2014) with dual averaging (Nesterov, 2009) for efficient exploration of the target distribution. The adaptive and dynamic nature of NUTS enables a streamlined estimation process that abolishes the need for costly preliminary tuning runs (needed for setting the number of leapfrog steps and the step size parameter) at the expense of some computation time per iteration. In the following, we distinguish between the burn-in period, which determines the number of samples that gets thrown out at the beginning of a Markov chain, and the warm-up period, which controls the length of the adaptive phase of the algorithm. Due to the high dependence between parameter blocks, all basis coefficients are updated in one step, followed by successive updates of the smoothing variances.

Update of the smoothing variances: In the univariate case, updating the smoothing variance is straightforward by using the full-conditional:

$$\tau_j^2 | \cdot \sim \mathrm{IG}\left(a_j + \frac{\mathrm{rk}(\boldsymbol{K}_j)}{2}, b_j + \frac{1}{2} \boldsymbol{\beta}_j^T \boldsymbol{K}_j \boldsymbol{\beta}_j\right),$$

where $\boldsymbol{K}_j$ is specified as shown in Section 2.6. However, in case of tensor splines based on a multivariate Gaussian prior with precision matrix,

$$\frac{1}{\tau_{j1}^2}(\boldsymbol{K}_{j1} \otimes \boldsymbol{I}_{D_2}) + \frac{1}{\tau_{j2}^2}(\boldsymbol{I}_{D_1} \otimes \boldsymbol{K}_{j2}), \tag{3.1}$$

we need to consider the generalized determinant of (3.1) when updating the smoothing variances. This aggravates sampling, which is why we introduce an anisotropy parameter $\omega_j \in (0, 1)$, resulting in an alternative representation of the precision given by

$$\frac{1}{\tau_j^2} \boldsymbol{K}_j = \frac{1}{\tau_j^2} \left[\omega_j (\boldsymbol{K}_{j1} \otimes \boldsymbol{I}_{D_2}) + (1 - \omega_j)(\boldsymbol{I}_{D_1} \otimes \boldsymbol{K}_{j2})\right],$$

where ω_j controls how much prior information is assigned to each of the two covariates of the tensor spline. For the BDCTM, we consider a discrete prior for ω_j, which allows to pre-compute a finite set of generalized determinants that can be used within the MCMC simulations (see Kneib et al., 2019) for a detailed explanation of this approach).

In the following, the hyperparameters of the inverse gamma prior are set to $a_{j1} = a_{j2} = 1$, $b_{j2} = b_{j2} = 0.001$, resulting in good and stable performance in all investigated cases.

Numerical stability: Klein et al. (2015a) observed numerical problems if zero-inflation was wrongfully assumed when in fact, for example, a simple Poisson model was due. One reason is that the estimated predictor for the probability of an extra zero tends towards minus infinity in log-space. This is usually not an issue in models of type (2.5) as the coefficients that are related to the zero component are not exp-transformed. In cumulative models with category-specific effects, however, flat sections can lead to divergent transitions in which case weakly identified coefficients have to be dropped from the model (Pya and Wood, 2015). This issue can be remedied by adding $\epsilon = 10e^{-6}$ to the diagonal of the precision matrix in this case. Moreover, the target acceptance rate can be increased to up to .99 to keep transitions in check.

Software: All computations were carried out in R version 4.1.0 (R Core Team, 2020). To improve computation time, likelihoods and score functions were implemented via the package `Rcpp` (Eddelbuettel et al., 2011). The mass matrix adaption scheme was adopted from `adnuts` (Monnahan and Kristensen, 2018).

4 Simulation study

In this section, we present a simulation experiment that highlights the possible advantages of the count transformation approach in general and that compares our Bayesian approach with the likelihood-based linear count transformation model by Siegfried and Hothorn (2020).

Count transformation models can mimic most well-known models for count data. Therefore, a meaningful simulation study in this setting needs to consider the sensitivity of the flexible transformation function with respect to the true data-generating process. In other words, it needs to investigate to what extent the flexible transformation function is able to accommodate eventual overdispersion and other characteristics of possibly complex data generating processes.

Simulation design: We use a similar simulation design to Siegfried and Hothorn (2020) with the following properties:

- One covariate is generated via $z \sim \mathbb{U}[0, 1]$.
- Conditional on z, we consider five different count data generating processes (DGPs)
  - *Poisson* with mean and variance $\mathbb{E}(Y|z) = \mathbb{V}(Y|z) = \exp(1.2 + 0.8z)$,
  - *Negative Binomial* with $\mathbb{E}(Y|z) = \exp(1.2 + 0.8z)$ and variance $\mathbb{V}(Y|z) = \mathbb{E}(Y|z) + \mathbb{E}(Y|z)^2/3$, and
  - Three different count data-generating processes according to $F_Z(\boldsymbol{a}_{(8)}(\log(y + 1))^T\boldsymbol{\gamma} - z\beta)$, $\beta = 0.8$ with the reference functions $F_Z = F_{\text{SL}}$ (*logit*), $F_Z = \Phi$ (*probit*), $F_Z = F_{\text{MEV}}$ (*cloglog*).

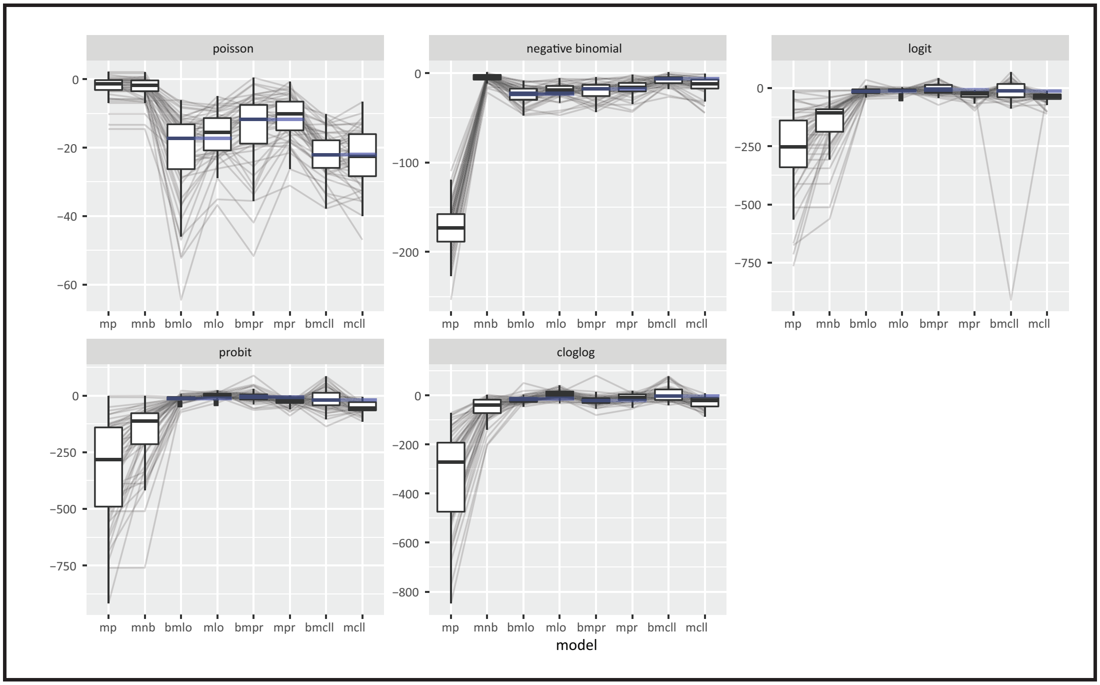

Figure 1 Comparison of count data-generating processes on basis of centered out-of-sample log-likelihoods obtained from the respective model. Larger out-of-sample log-likelihoods indicate a better performance of the corresponding model.

- Each dataset was estimated by their corresponding true (oracle) models, that is, a Poisson GLM (mp), a negative binomial GLM (mnb), BDCTMs ($bmlo$ denotes the logistic model, $bmpr$ denotes the probit model and $bmcll$ denotes the cloglog model) and a frequentist count transformation model (Siegfried and Hothorn, 2020) implemented in the **R**-package `cotram` (Siegfried and Hothorn, 2021), where mlo stands for the logistic model, mpr stands for the probit model and $mcll$ for the cloglog model. Each model type was estimated for each DGP, resulting in $5 \times 8 = 40$ models in total.
- Training and validation sample sizes are set to 250 and 750, respectively.
- The simulation experiment was repeated in 100 replications with a total iteration number of 2,000 and a burn-in and warm-up phase of length 1,000, such that 1,000 iterations are being used for computing the estimates.

Each model fit is quantified by means of the centered out-of-sample log-likelihood resulting from the difference between the out-of-sample log-likelihoods of the models and the out-of-sample log-likelihoods of the true data-generating processes evaluated on a hold-out sample, taking a predictive perspective that implicitly controls for differenes in complexity between the models. The results presented in Figure 1 confirm most of the findings of Siegfried and Hothorn (2020) regarding the merits of the count transformation approach.

Based on these results, we can make the following statements:

- The Poisson model, being the most rigid model, shows the worst performance with respect to the out-of-sample log-likelihood, if misspecified.
- As expected, the negative binomial model performs well for the Poisson and the overdispersed case, but shows inferior performance in the remaining scenarios.
- The fit of both the BDCTM and the `cotram` model is robust for all considered DGPs, effectively redeeming the promise of providing a flexible model framework for count data that is applicable in many situations.
- The BDCTM seems to perform better than `cotram` in the more complicated scenarios and worse especially in settings where a simple Poisson model would be due; this may be less surprising considering BDCTM's spline-based nature in comparison to `cotram`'s use of Bernstein polynomials.

The simulation study confirms the robustness of the BDCTM in the presence of different data-generating processes. Its fit is satisfactory in all investigated cases and highly competitive in the more complicated scenarios. While the Poisson distribution only works well in simple scenarios, the negative binomial distribution also works quite well for most scenarios (except the Poisson case). Still, BDCTMs outperform negative binomial regression uniformly over all but the Poisson and the negative binomial scenario.

5 Applications

We illustrate possible applications of the BDCTM in this section. For better readability, we add the number of basis functions to the basis, for $\boldsymbol{a}_{(q)}$. Code required for reproducing the following applications is openly accessible.[*]

5.1 Patent citations with excess zeros

Similar to an author of a scientific publication, an inventor who applies for a patent has to cite all existing patents their work is based on. We analyze the citation number ($ncit : y$) of patents granted by the European Patent Office (EPO). The considered dataset includes five dummys and three continuous variables. The available continuous covariates are the grant year ($year$), the number of the designated states ($ncountry$) and the number of patent claims ($nclaims$). (For a full description of the explanatory variables in the data set of $n = 4{,}805$ observations, see Jerak and Wagner 2006). A high rate of zeros ($\approx 46\%$) and a big spread ncit $\in \{0, \ldots, 40\}$ hint on the presence of zero-inflation and overdispersion. A rigorous investigation of this presumption has to consider whether this is holds conditional on the covariates. We let the sampler run for 2,000 iterations with a burn-in and warm-up phase of length 1,000 such that 1,000 iterations are obtained for inference.

[1] Source code available at https://github.com/manucarl/bdctm_showcase.

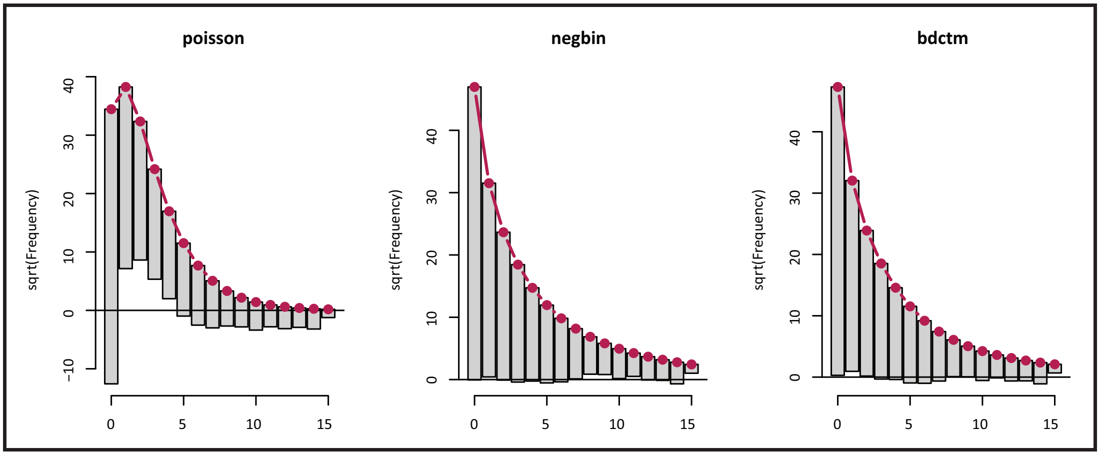

Figure 2 Patent citations. Rootograms of the linear Poisson, the linear negative binomial and the simple linear BDCTM model.

We start our investigation with the simple linear transformation model (BDCTM$_{\text{lin}}$):

$$F_{\text{SL}}(\boldsymbol{a}_8(\log(\lfloor y+1 \rfloor))^T\boldsymbol{\gamma} - \boldsymbol{z}^T\boldsymbol{\beta}), \tag{5.1}$$

where the linear predictor $\boldsymbol{z}^T\boldsymbol{\beta}$ contains all available covariates. As a first in-sample assessment of the practical capabilities of our transformation approach, we want to inspect to what extend the observed frequencies $\text{obs}_r = \sum_{i=1}^n \mathbb{1}(y_i = r)$ in the data set match the expected frequencies $\text{exp}_r = \sum_{i=1}^n (r; \hat{\boldsymbol{\gamma}}_i)$ derived from the model. Figure 2 displays the rootograms as introduced by Kleiber and Zeileis (2016) obtained from the model in Equation (5.1), from a Poisson and from a negative binomial GLM with all covariates included in the predictors. Rootograms make use of a horizontal reference line (at zero) to highlight the discrepancies between observed and expected frequencies. The Poisson model clearly underfits the zeros and exhibits an undulating pattern, overpredicting counts between 1 and 4, and underpredicting the rest, which is a sign of substantial overdispersion. The flexible transformation function of BDCTM is able to emulate the overdispersion-robust negative binomial model, which is reflected in the bars being closely aligned with the x-axis.

In summary, this first visual inspection of the goodness-of-fit confirms that BDCTM is able to ameliorate the impact of overdispersion on the model fit.

We also want to pursue the assumption of excess zeros. For this, we consider a two-component model (BDCTM$_{\text{hurdle-lin}}$) in the vein of (2.5) with $h(\boldsymbol{x}) = h_0(\boldsymbol{x}) = \boldsymbol{z}^T\boldsymbol{\beta}$:

$$F_{\text{SL}}(\boldsymbol{a}_8(\log(\lfloor y+1 \rfloor))^T\boldsymbol{\gamma} - \boldsymbol{z}^T\boldsymbol{\beta} + \mathbb{1}(y=0)(\beta_0 - \boldsymbol{z}^T\boldsymbol{\beta})),$$

where, again, $\boldsymbol{z}$ contains all explanatory variables in the data set. As GLM analogues, we consider the zero-inflated versions of the Poisson and of the negative binomial models. Previous analyses of the data set revealed that assuming non-linear relationships for the continuous covariates can improve the estimation results (Klein et al., 2015a). This does not automatically hold for BDCTM, because the explanatory variables impact the response on a different scale (the scale of

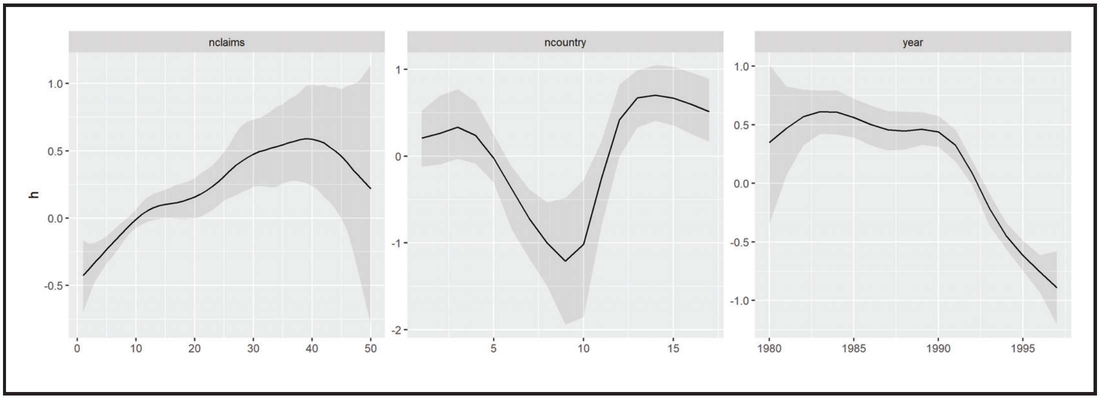

Figure 3 Patent citations. Posterior mean estimates of the effects of nclaims, ncountry and year on the log-odds ratio, together with 95% credible intervals. Remaining covariates are held constant at their mean or are set to zero in case of dummy variables. Estimates belong to BDCTM$_{\text{nl}}$.

the transformation). Therefore, we estimated models of type (2.4) and (2.5), while replacing the covariate functions with additive functions of type (2.3), that is, $h(\boldsymbol{x}) = h_0(\boldsymbol{x}) = \boldsymbol{z}^T\boldsymbol{\beta} + f(ncountry) + f(year) + f(nclaims)$, where $\boldsymbol{z}$ now only contains the discrete covariables. In what follows, we refer to these partially non-linear models as BDCTM$_{\text{nl}}$ and BDCTM$_{\text{hurdle-nl}}$, respectively. Figure 3 shows the estimated non-linear effects of *ncountry*, *year* and *nclaims* on the log-odds ratio from model BDCTM$_{\text{nl}}$.

In the next step, we compared all models in terms of *randomized quantile residuals* as proposed by Rigby et al. (2008). For every observation y_i, we computed residuals $\hat{r}_i = \Phi^{-1}(u_i)$ where Φ^{-1} is the quantile function of the standard normal distribution and u_i is randomly drawn from $\mathbb{U}(F(y_i - 1)|\hat{\boldsymbol{\gamma}}), F(y_i|\hat{\boldsymbol{\gamma}}))$ with plugged in estimates $\hat{\boldsymbol{\gamma}}$. $F(\cdot|\hat{\boldsymbol{\gamma}})$ is the estimated conditional distribution function. Residuals obtained from the true model follow a standard normal distribution, which is why deviations can be checked by quantile-quantile plots. Figure 4 shows the Q-Q plots of the considered models. Again, the Poisson model reveals a lack of fit represented by the strong deviations from the normal line, which also holds true for its zero-inflated counterpart to a somewhat lesser extend. The negative binomial models provide a considerably better fit but seem to be surpassed by the BDCTMs, which indicate the best aptitude for infering the distribution of patent citations while at the same time providing a flexible ‘sans souci’ approach, abolishing the need to search for the ‘right’ count distribution in general.

For a more rigorous assessment of the out-of-sample performance, we conclude our analysis with an evaluation based on *proper scoring rules.* Originally proposed by Gneiting and Raftery (2007), they serve as summary measures for the predictive power of a model. Based on data $y_1, \ldots, y_R$ in a validation sample and estimated probabilities $\hat{\boldsymbol{p}}_r = (\hat{p}_{r0}, \hat{p}_{r1}, \ldots)$ obtained from the predictive distribution $\hat{p}_{rk} = f(y_r = k|\hat{\boldsymbol{\gamma}})$, scores are computed by taking the sum of the individual score contribution $S = \sum_{r=1}^{R} S(\hat{\boldsymbol{p}}_r, y_r)$. We consider the three most prominent scores:

- Brier score: $S(\hat{\boldsymbol{p}}_r, y_r) = -\sum_k (\mathbb{1}(y_r = k) - \hat{p}_{rk})^2$,
- Logarithmic score: $S(\hat{\boldsymbol{p}}_r, y_r) = \log(\hat{p}_{r y_r})$ (out-of-sample likelihood), and
- Spherical score: $S(\hat{\boldsymbol{p}}_r, y_r) = \frac{\hat{p}_{r y_r}}{\sqrt{\sum_k \hat{p}_{rk}^2}}$.

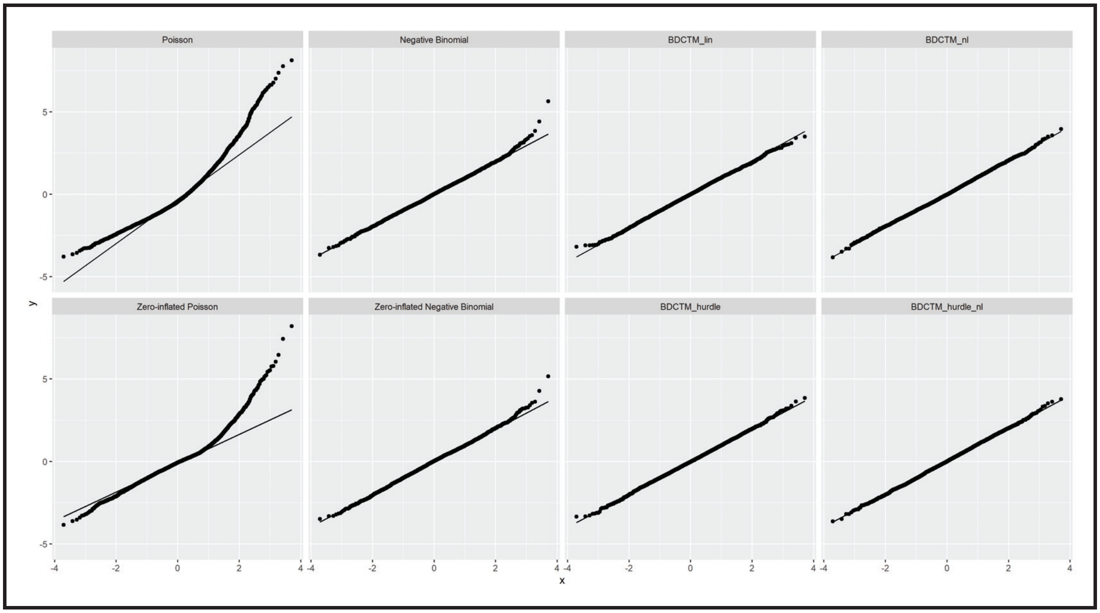

Figure 4 Patent citations. Comparison of quantile residuals obtained by BDCTM models with and without addititional zero component with various generalized linear and zero-inflated models.

The probabilistic forecasts collected in $\hat{\boldsymbol{p}}_r$ for the responses y_r are assessed by 10-fold cross-validation. Table 1 shows the score sums obtained from the four BDCTM models introduced in this section, together with the Watanabe Information Criterion for Bayesian models (WAIC, Watanabe (2010)). The `cotram` model is specified equivalently to $\text{BDCTM}_{\text{lin}}$, which is why their similar performance in terms of quadratic and spherical score is not surprising. Note that the logarithmic score considers only one probability of the predictive distribution and is therefore vulnerable to outliers and extreme observations, which could explain the better performance of $\text{BDCTM}_{\text{lin}}$ in that regard. Both, considering excess zeros and non-linear effects, come with improved predictive power, culminating in the $\text{BDCTM}_{\text{hurdle-nl}}$'s dominating performance across all measures besides the WAIC where the zero component did not lead to improvements. The scores could be further improved by a model selection procedure as shown in Klein et al. (2015a).

Table 1 Patent citations. Score sums of all models obtained via 10-fold cross-validation. Calculation of the WAICs on basis of the whole data set. Best results are depicted in bold font

Model	Logarithmic	Quadratic	Spherical	WAIC
$\text{BDCTM}_{\text{lin}}$	−8119.67	−3444.53	2530.84	6257.85
$\text{BDCTM}_{\text{hurdle-lin}}$	−8091.47	−3438.87	2534.39	6224.634
BDCTM_{nl}	−8110.94	−3440.52	2533.98	**6040.573**
$\text{BDCTM}_{\text{hurdle-nl}}$	**−8044.69**	**−3427.77**	**2543.44**	6184.174
cotram	−8174.92	−3443.07	2531.23	-

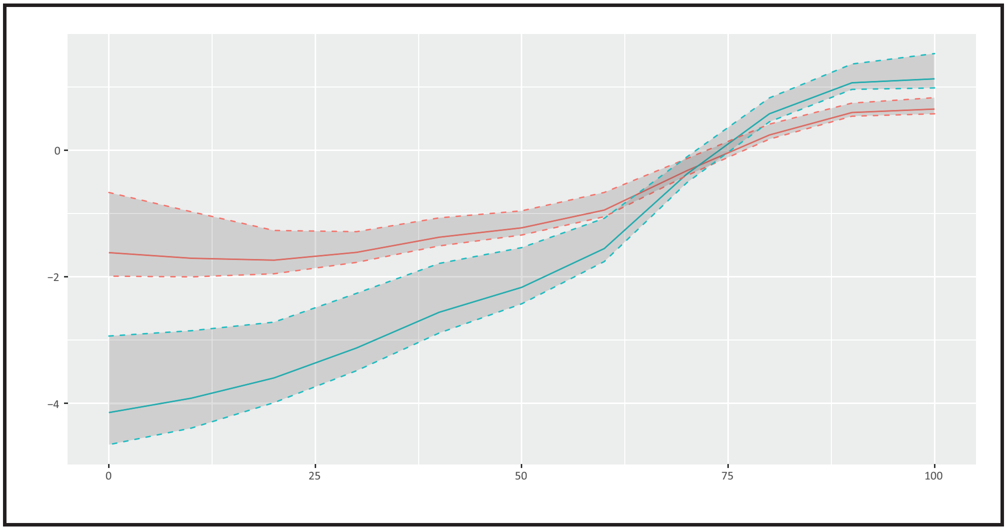

Figure 5 Forest health: estimated non-linear category-specific effect of *canopy*, “no defoliation” in red, “severe defoliation” in blue, together with 95%-credible intervals.

5.2 A partial proportional odds model for forest health assessment

This short analysis involving non-linear category-specific effects is based on data from the forest of Rothenbuch (Spessart) over the years 1982–2004. Every year, the health status is evaluated and categorized by the response variable $defol$ measuring defoliation grades. Since data is sparse in some of the original nine categories (0%, 12.5%, . . . , 100%), we aggregated them into the three defoliation grades: 1 = no (0%), 2 = weak (12.5% − 37.5%) and 3 = severe ($\geq$ 50%). Among others, the dataset comes with the covariates $canopy$ (canopy density in percentage), x, y (x- and y-coordinates of location) and id (tree location identification number.). (Check Fahrmeir et al. (2013) for a full description of the dataset). The goal of this analysis is to determine the effect of the covariates on the degree of defoliation. Since the forest data is notorious for confounding and high autocorrelation, we let the sampler run for 10,000 iterations with a burn-in and warm-up phase of length 1000.

For this, we set up the partial proportional odds model

$$\begin{aligned} F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_r) = F_{\mathrm{SL}}(&\boldsymbol{e}(defol)^T\boldsymbol{\gamma}_1 + (\boldsymbol{e}(defol)^T \otimes \boldsymbol{b}_{(10)}(canopy)^T)^T\boldsymbol{\gamma}_2 \\ &- \boldsymbol{b}(id)^T\boldsymbol{\beta}_3 \\ &- (\boldsymbol{b}_{(10)}(x)^T \otimes \boldsymbol{b}_{(10)}(y)^T)^T\boldsymbol{\beta}_4), \end{aligned}$$

where we assume non-linear category-specific shifts of $canopy$, a transformation random effect for the tree location groups and a spatial non-linear effect on the basis of a tensor spline for the coordinates x and y. Figure 5 shows the estimated non-linear category-specific effect for $canopy$. The section for $0 \leq canopy \leq 25$ displays almost parallel curves, which then vary more and more individually until they even cross. The variance of the estimated random effect for id is 2.42, and the standard deviation is 1.55. Figure 6 shows the estimated random intercepts. In a preliminary run, we

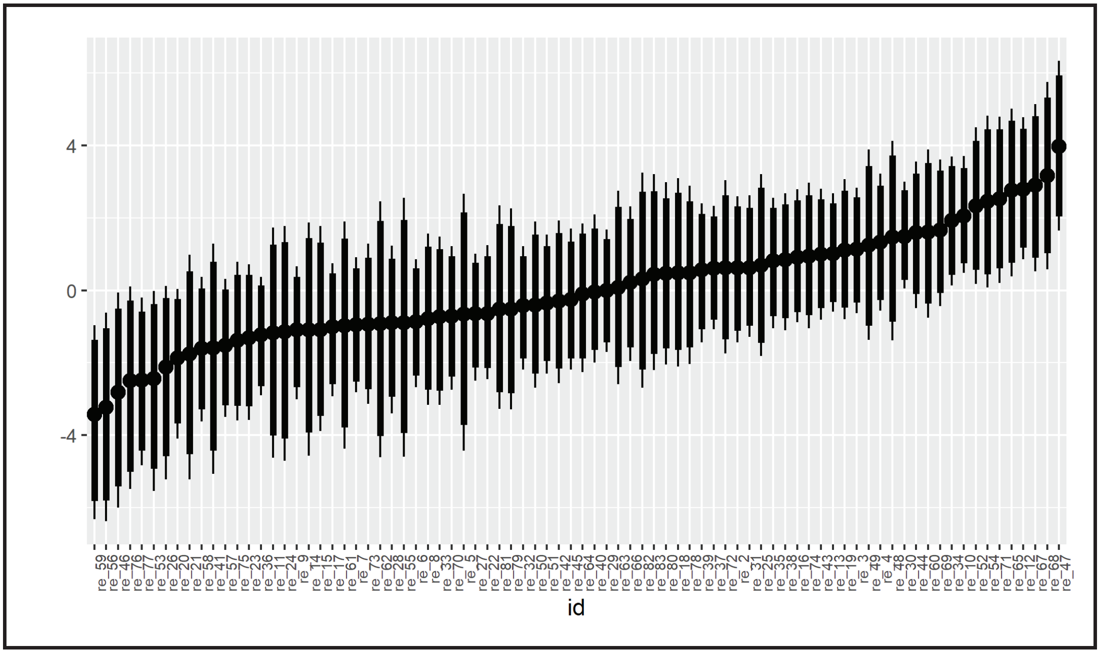

Figure 6 Forest health: median-sorted estimated random intercepts for tree location groups.

observed the same problems with confounding in location-specific effects as Fahrmeir et al. (2013), which could be improved to some extend by adding the spatial effect. It is displayed in Figure 7.

6 Discussion

With the BDCTM, we present a novel Bayesian model framework for discrete data that combines cumulative link models with models for count data through directly modeling the conditional distribution function. Approaching these discrete data structures from the transformation perspective allows us to unify models that are usually treated seperately under the same umbrella. The BDCTM is flexible in the sense that it permits the user to control interpretability by means of choosing a reference distribution in conjunction with an additive transformation function. Estimating the conditional distribution function directly makes deriving distributional aspects such as the conditional quantiles straightforward by numerical inversion of $F_Z(h(y|\boldsymbol{x})$ (Siegfried and Hothorn, 2020). Furthermore, our Bayesian inferential procedure lets us obtain credible intervals and other quantities of interest without having to rely on large sample approximations. All high-dimensional effects are joined with suitable prior specifications, resulting in smooth effects across the board.

We demonstrate BDCTM's ability to handle under- or overdispersion in an adaptive fashion without restrictive distributional assumptions in Sections 4 and 5. A short investigation of a nonlinear non-proportional odds model highlights the versatility of our approach. In a model selection context, the unifying scope of the transformation function turns out to be a valuable simplification

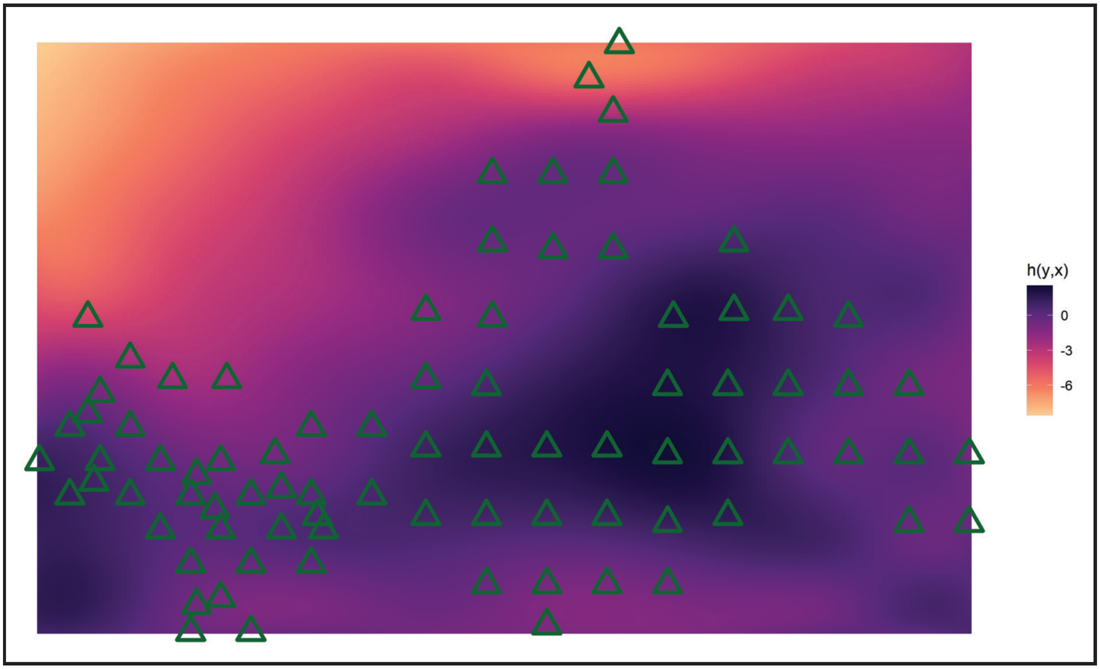

Figure 7 Forest health: estimated two-dimensional spatial effect with triangles indicating observed tree locations based on 2nd order penalties.

because there is just one "predictor" that has to be constructed. Though not shown in this article, it is possible to establish a relationship between overdispersion and the covariate effects by including full non-linear interactions between the count response and the respective explanatory variable. Constructing the conditional transformation function can be difficult as informed decisions about which effects to include and to interact with the response are required. Therefore, it would be desirable to develop an effect selection strategy via spike and slab priors in the spirit of Klein et al. (2021) for the BDCTM that could effectively tell the user what kind of effect is impacting the regular count process, the zero component or overdispersion.

As demonstrated in Section 5.2, our cumulative link transformation approach can be supplemented with category-specific linear or non-linear effects by modeling them as response-covariate interactions. This way, popular models such as (non-)proportional odds or hazards models can be retrieved simply by specifying the reference distribution. Both the count and the ordinal model could be supplemented with a more flexible link function as proposed by Aranda-Ordaz (1983), that is,

$$F(h) = 1 - (\lambda \exp(h) + 1)^{-\lambda^{-1}},$$

which depends on an auxiliary parameter $\lambda \in]0, \infty[$, mitigating between the log-log link for $\lambda \to 0$ and the logistic link when $\lambda \to 1$. Horowitz (2001) avoided specifying the link function entirely. A

Bayesian version would entail prior distributions on the space of nonparametric continuous reference distribution.

To conclude, we believe that in this article, the BDCTM is established as a flexible, modular modeling framework in the world of discrete data that is competitive in many modern scenarios.

Supplementary material

Supplementary material is available online.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

Thomas Kneib received financial support from the DFG within the research project KN 922/9-1. The work of Manuel Carlan was supported by DFG via the research training group 1644.

References

Aranda-Ordaz F (1983) An extension of the proportional-hazards model for grouped data. *Biometrics*, **39**, 109–17.

Brezger A and Lang S (2006) Generalized structured additive regression based on bayesian p-splines. *Computational Statistics & Data Analysis*, **50**, 967–91.

Cameron A and Trivedi P (1998) *Regression Analysis of Count Data.* London: Cambridge University Press.

Carlan M, Kneib T and Klein N (2020) *Bayesian Conditional Transformation Models*. arXiv e-prints, page arXiv:2012.11016.

Dey DK, Ghosh SK and Mallick BK (2000) *Generalized Linear Models: A Bayesian Perspective*. Boca Raton: CRC Press.

Doksum KA and Gasko M (1990) On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review*, **58**, 243–52.

Dunson DB (2005) Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, **100**, 618–27.

Eddelbuettel D, François R, Allaire J, Ushey K, Kou Q, Russel N, Chambers J and Bates D (2011) Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, **40**, 1–18.

Fahrmeir L, Kneib T, Lang S and Marx B (2013) *Regression: Models, Methods and Applications*. New York: Springer.

Frühwirth-Schnatter S and Wagner H (2006) Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika*, **93**, 827–41.

Frühwirth-Schnatter S, Frühwirth R, Held L and Rue H (2009) Improved auxiliary mixture sampling for hierarchical models of non-gaussian data. *Statistics and Computing*, **19**, 479–92.

Ghosh SK, Mukhopadhyay P and Lu J-CJ (2006) Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, **136**, 1360–75.

Gneiting T and Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, **102**, 359–78.

Hastie TJ and Tibshirani RJ (1990) *Generalized Additive Models*, volume 43. Boca Raton: CRC press.

Hilbe JM (2011) *Negative Binomial Regression*. London: Cambridge University Press.

Hoffman MD and Gelman A (2014) The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, **15**, 1593–1623.

Horowitz JL (2001) Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, **69**, 499–513.

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

Hothorn T, Möst L and Bühlmann P (2018) Most likely transformations. *Scandinavian Journal of Statistics*, **45**, 110–34.

Jerak A and Wagner S (2006) Modeling probabilities of patent oppositions in a bayesian semiparametric regression framework. *Empirical Economics*, **31**, 513–33.

Kleiber C and Zeileis A (2016) Visualizing count data regressions using rootograms. *The American Statistician*, **70**, 296–303.

Klein N, Kneib T and Lang S (2015a) Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, **110**, 405–19.

Klein N, Kneib T, Lang S, Sohn A (2015b) Bayesian structured additive distributional regression with an application to regional income inequality in germany. *The Annals of Applied Statistics*, **9**, 1024–52.

Klein N, Carlan M, Kneib T, Lang S and Wagner H (2021) Bayesian effect selection in structured additive distributional regression models. *Bayesian Analysis*, **16**, 545–73.

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**, 1–39.

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

Lavine M and Mockus A (1995) A nonparametric bayes method for isotonic regression. *Journal of Statistical Planning and Inference*, **46**, 235–48.

Manuguerra M and Heller GZ (2010) Ordinal regression models for continuous scales. *The International Journal of Biostatistics*, **6**.

McCullagh P (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **42**, 109–27.

Monnahan CC 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**, e0197954.

Nelder JA and Wedderburn RW (1972) Generalized linear models. *Journal of the Royal Statistical Society: Series A*, **135**, 370–84.

Nesterov Y (2009) Primal-dual subgradient methods for convex problems. *Mathematical Programming*, **120**, 221–59.

Peterson B and Harrell FE (1990) Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **39**, 205–17.

Pya N and Wood SN (2015) Shape constrained additive models. *Statistics and Computing*, **25**, 543–59.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Rigby B, Stasinopoulos M and Akantziliotou C (2008) Instructions on how to use the gamlss package in r. *Computational Statistics and Data Analysis*, **2**, 194–95.

Rodrigues J (2003) Bayesian analysis of zero-inflated distributions. *Communications in Statistics-Theory and Methods*, **32**, 281–89.

Siegfried S and Hothorn T (2020) Count transformation models. *Methods in Ecology and Evolution*, **11**, 818–27.

Siegfried S and Hothorn T (2021) *Count Transformation Models: The cotram Package*. URL https://CRAN.R-project.org/package=cotram. R package version 0.2.1.

Sokal RR and Rohlf FJ (1981) *Biometry: The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman Ltd.

Tutz G (2011) *Regression for Categorical Data*, volume 34. London: Cambridge University Press.

Watanabe S (2010) Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–94.

Winkelmann R (2008) *Econometric Analysis of Count Data*. New York: Springer.