Bayesian Analysis (2021) **16**, Number 2, pp. 545–573

Bayesian Effect Selection in Structured Additive Distributional Regression Models

Nadja Klein[*,‖], Manuel Carlan[†,**], Thomas Kneib[‡,††] Stefan Lang[§], and Helga Wagner[¶]

Abstract. We establish Bayesian effect selection for the broad class of structured additive distributional regression models using a spike and slab prior specification with scaled beta prime marginals for the importance parameters of blocks of regression coefficients. This enables us to model and select effects in all distributional parameters, such as location, scale, skewness or correlation parameters, for arbitrary distributions. The regression specifications encompass various effect types such as non-linear or spatial effects. Our spike and slab prior relies on a parameter expansion that separates blocks of regression coefficients into overall scalar importance parameters and vectors of standardised coefficients, and yields effective shrinkage and good sampling performance. Using constrained priors, it is possible to implement effect decompositions, where, for example, a non-linear effect can be decomposed into a linear component and the non-linear deviation from this linear effect; and to select both separately. We investigate some shrinkage properties, propose a way of eliciting prior hyperparameters and provide full posterior inference through Markov Chain Monte Carlo simulations. Using both simulated and real data sets, we show that our approach is applicable for data with various functional covariate effects, multilevel predictors and non-standard response distributions, such as bivariate Gaussian or zero-inflated Poisson.

Keywords: penalised splines, prior elicitation, parameter expansion, scaled beta prime distribution, shrinkage properties.

1 Introduction

The flexibility of modern regression methodology is both a blessing and a curse for applied researchers and statisticians alike since, on the one hand, added flexibility enables potentially more realistic models approximating the true data generating process but, on the other hand, poses additional challenges in the model building and model checking process. In this paper, we consider structured additive distributional regression models (Rigby and Stasinopoulos, 2005; Klein et al., 2015c) that combine additive predictors

[*]Correspondence should be directed to Nadja Klein, Nadja Klein is an Assistant Professor at the Humboldt Universität of Berlin, School of Business and Economics, Unter den Linden 6, 10099 Berlin, nadja.klein@hu-berlin.de

[†]Georg-August-Universität Göttingen, mcarlan@uni-goettingen.de

[‡]Georg-August-Universität Göttingen, tkneib@uni-goettingen.de

[§]Universität Innsbruck, stefan.lang@uibk.ac.at

[¶]Johannes-Kepler-Universität Linz, helga.wagner@jku.at

[‖]Nadja Klein acknowledges financial support by the German Research Foundation (DFG) through the Emmy Noether Grant KL 3037/1-1.

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

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

 https://doi.org/10.1214/20-BA1214

consisting of various types of regression effects, e.g. non-linear effects of continuous covariates, spatial effects or random effects (Kammann and Wand, 2003; Ruppert et al., 2003; Wood, 2017) with the possibility to model all parameters (e.g. location, scale or shape parameters) of arbitrary parametric response distributions in terms of covariates in a distributional regression approach. Examples of distributional models that fit well in this framework include the ones for univariate responses of any type (e.g. counts with zero-inflation/overdispersion, continuous responses with spikes, skewness, heavy tails or bounded support) as well as multivariate responses (such as multivariate normal, multivariate t, copula-based specifications with complex dependence structure, or Dirichlet regression for analyzing compositional data). As a consequence, an analyst is faced with the challenge not only of choosing an appropriate response distribution, (a task that we will not consider here, see for example Klein et al., 2015c, for practical solutions to this task) but also with determining the most appropriate subset of covariates along with their exact modelling alternative for multiple regression predictors.

For instance, in one of our empirical illustrations on childhood undernutrition with more than 20,000 observations, we analyse a bivariate response variable $\boldsymbol{y} = (y_1, y_2)'$ consisting of two scores for chronic and acute undernutrition. A previous study (Klein et al., 2015a) suggests a bivariate normal model in which not only the marginal expectations but also the marginal scale parameters and the correlation parameter depend on covariates. This leads to a distributional regression model with $K = 5$ distributional parameters $\vartheta_k \in \{\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\}$. In a semiparametric model with $i = 1, \ldots, n$ observations $\boldsymbol{y}_i$ each of these could be related to a predictor η_{ik} of the form

$$\eta_{ik} = \boldsymbol{x}_i'\boldsymbol{\beta_k} + f_{1,k}(cage) + f_{2,k}(mage) + f_{3,k}(mage) + f_{4,k}(mbmi) + f_{spat,k}(region), \tag{1}$$

where k refers to one of the five distributional parameters, $\boldsymbol{x}_i$ contains 13 binary/categorical covariates (and an intercept) with regression coefficients $\boldsymbol{\beta}_k$, $f_{j,k}(\cdot)$, $j = 1, 2, 3$, are non-linear smooth functions of child's age (*cage*), years of partner's education, mother's age (*mage*) and mother's body mass index (*mbmi*), and $f_{spat,k}$ are spatial effects based on regional information in the data. While effect selection (deciding which of the different effects should be included in the model) via a full search in the model space would already be challenging in a mean regression framework with only one single structured additive predictor, full effect selection in a distributional regression setting with multiple predictors is typically computationally prohibitive (the number of candidate models would be $2^{\text{number of effects} \times \text{number of predictors}} = 2^{17 \times 5}$). This is even more the case when one is interested in deciding whether the effect of a continuous covariate shall be included in a linear or non-linear form or whether it could be excluded completely from the model. In this paper, we address these challenges and develop the first contribution to general Bayesian effect selection for structured additive distributional regression models (Klein et al., 2015c).

While there has been extensive interest in spike and slab priors for Bayesian variable selection (i.e. the selection of effects in models with purely linear predictors) or function selection (selection of non-linear effects of continuous covariates) in previous years (see for example Clyde and George, 2004; O'Hara and Sillanpää, 2009, for reviews), most research has been restricted to mean regression with Gaussian errors, distributions from

the exponential family or survival models, and a focus on modelling the conditional expectation of the response. Furthermore, most approaches restrict the predictor specification to include either only linear effects or only non-linear effects of continuous covariates but do not enable the consideration of the decomposition of non-linear effects in linear and non-linear components or more complex effect types such as spatial effects. Classical Bayesian variable selection approaches for linear models based on spike and slab priors include for example Mitchell and Beauchamp (1988) or George and McCulloch (1997). Smith and Kohn (1996) utilise similar ideas for function selection in nonparametric Gaussian regression, while group variable selection has been considered in Zhang et al. (2014); Xu and Ghosh (2015).

Approaches that move beyond the framework of Gaussian models comprise Rossell and Rubio (2018) who propose Bayesian variable selection with non-local priors allowing for skewness and thicker tails compared to the Gaussian response distribution, and Wang et al. (2017) who consider variable selection after transforming the response. Chung and Dunson (2009) investigate variable selection for a distributional model constructed through a probit stick-breaking process, while Kundu and Dunson (2014) consider selection when the distribution of the errors is modelled non-parametrically.

Usually, the variable selection or shrinkage priors are directly imposed on scalar regression coefficients. In contrast, Ishwaran and Rao (2005) consider a hierarchical specification where the spike and slab structure is imposed on a higher level of the hierarchy, i.e. their prior variances. This is advantageous in situations where selection should take place on blocks of regression coefficients representing for example the coefficients of a basis expansion in nonparametric regression. This leads to function selection approaches for additive models, also considered in Yau et al. (2003); Cottet et al. (2008); Reich et al. (2009), who combine a spike with point mass at zero with a slab that has support only on the positive real numbers. In contrast, Zhu et al. (2010) specify both spike and slab as normal distributions (with very different variance components) and Panagiotelis and Smith (2008) assign a multivariate prior with spike at the origin and normal slab directly to the whole vector of basis coefficients. In either case, one typically observes poor mixing unless sampling from marginalized full conditionals. However, these are only available in closed form for Gaussian models or models that have a latent Gaussian representation such as the probit model (Zhu et al., 2010). One of the very few existing contributions to address selection on further distributional parameters is Cottet et al. (2008), who propose function selection in double exponential regression models, where both the mean and the dispersion parameter are linked to an additive predictor. The model space is restricted, since non-linear effects may enter the model only if the corresponding linear effect is included in the model, similar to Rossell and Rubio (2019), who consider cubic splines in additive survival models under censoring.

Our proposal is inspired by the approach of Scheipl et al. (2012) that introduces effect selection in generalized additive models for the exponential family with only one mean-related additive predictor. As Scheipl et al. (2012), we rely on a parameter expansion of the vector of the basis coefficients as originally proposed in Gelman et al. (2008), which allows us to expand the vector of basis coefficients in an importance parameter shared by all basis coefficients on the one hand and standardised basis coefficients on

the other hand. Effect selection is then performed by assigning a spike and slab prior to the squared importance parameter. More precisely, our paper makes the following important contributions:

- We propose a tractable solution to Bayesian effect selection based on spike and slab priors for structured additive distributional regression such that selection of general effect types is no longer restricted to additive mean regression models with exponential families as in Scheipl et al. (2012).
- We assign constrained multivariate normal priors to blocks of basis coefficients. The constraint allows us to keep sparse matrix structures, thus enabling efficient computations. In addition, we do not observe the strong dependence on the dimensionality of the basis coefficient vector identified in Scheipl et al. (2012), such that our method can also jointly select high dimensional coefficient vectors such as the ones induced by spatial effects. The constrained priors also give a natural decomposition of effects into the sum of simpler effects through projection, which broadens the scope of effect selection questions.
- Taking advantage of the modularity of Bayesian inference based on Markov chain Monte Carlo simulations, we extend the model to hierarchical multilevel specifications of the predictors following Lang et al. (2014).
- Formulating the spike and slab prior for the squared importance parameter yields scaled beta prime marginals, which have favourable shrinkage properties (Pérez et al., 2017). We examine some prior properties in detail and provide conditions for the propriety of the posterior.
- We develop rules for eliciting the hyperparameters of the spike and slab prior based on interpretable scaling criteria that are easily accessible to applied researchers. Based on the elicited parameters, we find that our new prior structure has similarly favourable shrinkage properties as the approach by Scheipl et al. (2012), while it avoids to arbitrarily fix the hyperparameters and is applicable in a much broader model class.

The rest of this paper is structured as follows: Section 2 summarises the specification of Bayesian effect selection priors for structured additive distributional regression models. Some properties of the effect selection prior are discussed in Section 3. Section 4 contains details on posterior estimation, software and implementation. Sections 5.1 and 5.2 evaluate the performance of our approach in simulations and three diverse applications. In Section 6 we conclude.

2 Bayesian Effect Selection in Distributional Regression

2.1 Observation Model

Distributional Regression We develop our approach for the class of structured additive distributional regression (Klein et al., 2015c). Let $(\boldsymbol{y}_i, \boldsymbol{\nu}_i)$, $i = 1, \ldots, n$ denote n

independent observations of a response variable $\boldsymbol{Y} \in \mathcal{Y} \subseteq \mathbb{R}^p$, $p \geq 1$ and $\boldsymbol{\nu}$ the covariate vector comprising different types of covariate information such as discrete and continuous covariates or spatial information. We then assume that the conditional distribution of $\boldsymbol{y}_i$ given $\boldsymbol{\nu}_i$ is specified in terms of a K-parametric distribution with density

$$p(\boldsymbol{y}_i|\vartheta_{i1}, \ldots, \vartheta_{iK}), \tag{M1}$$

where $\boldsymbol{\vartheta}_i = (\vartheta_{i1}, \ldots, \vartheta_{iK})'$ is a collection of K scalar distributional parameters ϑ_{ik}, $k = 1, \ldots, K$. Various simpler models, such as generalized additive or survival models are included as special cases. However, in such mean regression models with $p(\cdot)$ from the exponential family, the focus is on modelling $\vartheta_i = \mathbb{E}(y_i)$, while all other $K - 1$ parameters are treated as fixed or nuisance parameters. In distributional regression in contrast, each of the distributional parameters ϑ_{ik} is related to regression effects and therefore depends on $\boldsymbol{\nu}_i$. More precisely, we assume that each ϑ_{ik} is related to a structured additive predictor η_{ik} via a one-to-one response function h_k, i.e. $h_k(\eta_{ik}) = \vartheta_{ik}$ and $\eta_{ik} = h_k^{-1}(\vartheta_{ik})$. The distributional regression framework allows for considerable flexibility where, for example, regression effects on the scale, the skewness, etc. can be studied in considerable detail, as we illustrate in our applications in Section 5.2.

Structured Additive Predictors Structured additive predictors are specified as

$$\eta_{ik} = \eta_{ik}^{\text{in}} + \eta_{ik}^{\text{sel}} = \sum_{l=1}^{L_k} f_{l,k}^{\text{in}}(\boldsymbol{\nu}_i) + \sum_{j=1}^{J_k} f_{j,k}^{\text{sel}}(\boldsymbol{\nu}_i), \tag{M2}$$

where the effects $f_{j,k}^{\text{sel}}(\boldsymbol{\nu}_i)$ represent various types of flexible functions depending on (different subsets of) the covariate vector $\boldsymbol{\nu}_i$ that are to be selected via spike and slab priors, while η_{ik}^{in} represents a second additive predictor consisting of all effects $f_{l,k}^{\text{in}}(\boldsymbol{\nu}_i)$ that are *not* under selection. The separation into two subsets of effects allows us to include specific covariate effects mandatorily in the model (e.g. based on prior knowledge or since these represent confounding effects that have to be included in the model). In the following, we will only discuss the specification of priors for $f_{j,k}^{\text{sel}}(\boldsymbol{\nu}_i) \equiv f_{j,k}$ and refer to Klein et al. (2015a) for handling η_{ik}^{in}.

Let therefore $f_{j,k}$ be effect j subject to selection in predictor k. It is then assumed that $f_{j,k}$ can be modelled as

$$f_{j,k}(\boldsymbol{\nu}_i) = \tau_{j,k} \sum_{d=1}^{D} \tilde{\beta}_{j,k,d} B_{j,k,d}(\boldsymbol{\nu}_i), \tag{M3}$$

where $B_{j,k,d}(\boldsymbol{\nu}_i)$, $d = 1, \ldots, D$ are appropriate basis functions, $\tilde{\boldsymbol{\beta}}_{j,k} = (\tilde{\beta}_{j,k,1}, \ldots, \tilde{\beta}_{j,k,D})'$ is the vector of (standardised) basis coefficients and $\tau_{j,k}$ is an importance parameter representing the overall relevance of $f_{j,k}(\boldsymbol{\nu})$. Due to the linear basis representation, the vector of function evaluations $\boldsymbol{f}_{j,k} = (f(\boldsymbol{\nu}_{j,k,1}), \ldots, f(\boldsymbol{\nu}_{j,k,n}))'$ can now be written as $\boldsymbol{f}_{j,k} = \tau_{j,k} \boldsymbol{B}_{j,k} \tilde{\boldsymbol{\beta}}_{j,k}$ where $\boldsymbol{B}_{j,k}$ is the $(n \times D)$ design matrix arising from the evaluation

of the basis functions $B_{j,k,d}(\boldsymbol{\nu}_i)$, $d = 1, \ldots, D$ at the observed $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$, see Section 2.3 below for some examples. The parameterisation in (M3) is equivalent to

$$f_{j,k}(\boldsymbol{\nu}_i) = \sum_{d=1}^{D} \beta_{j,k,d} B_{j,k,d}(\boldsymbol{\nu}_i) \tag{M3*}$$

but "redundant" in the sense that only the product $\boldsymbol{\beta}_{j,k} = \tau_{j,k}\tilde{\boldsymbol{\beta}}_{j,k}$ is identified. However, the importance parameter $\tau_{j,k}$ allows us to remove the effect from the predictor for $\tau_{j,k} = 0$, while the effect is considered to be of high importance if $\tau_{j,k}$ is large in absolute terms. We will place a spike and slab prior on the squared importance parameter $\tau_{j,k}$ to achieve effect selection.

2.2 The Normal Beta Prime Spike and Slab Prior

Constrained Prior for Regression Coefficients Since for many specific types of effects the vector of basis coefficients $\boldsymbol{\beta}_{j,k}$ is of relatively high dimension, it is often useful to enforce specific properties such as smoothness or shrinkage. In a Bayesian formulation, this can be facilitated by assuming constrained multivariate Gaussian priors

$$p(\boldsymbol{\beta}_{j,k}|\tau^2_{j,k}) \propto \exp\left(-\frac{1}{2\tau^2_{j,k}}\boldsymbol{\beta}'_{j,k}\boldsymbol{K}_{j,k}\boldsymbol{\beta}_{j,k}\right) \mathbb{1}\left[\boldsymbol{A}_{j,k}\boldsymbol{\beta}_{j,k} = \mathbf{0}\right], \tag{M4*}$$

where $\boldsymbol{K}_{j,k} \in \mathbb{R}^{D_{j,k} \times D_{j,k}}$ denotes the prior precision matrix implementing the desired smoothness properties, $\tau^2_{j,k}$ is a prior variance parameter controlling the degree of smoothness and the indicator function $\mathbb{1}[\boldsymbol{A}_{j,k}\boldsymbol{\beta}_{j,k} = \mathbf{0}]$ is included to enforce linear constraints on the regression coefficients via the constraint matrix $\boldsymbol{A}_{j,k}$. The latter is typically used to remove identifiability problems from the additive predictor (e.g. by centring the additive components of the predictor) but can also be used to remove the partial impropriety from the prior that comes from a potential rank deficiency of $\boldsymbol{K}_{j,k}$ with $\text{rk}(\boldsymbol{K}_{j,k}) = \kappa_{j,k} \leq D_{j,k}$.

We specify a prior of a very similar structure but on the vector of scaled basis coefficients $\tilde{\boldsymbol{\beta}}_{j,k}$,

$$p(\tilde{\boldsymbol{\beta}}_{j,k}) \propto \exp\left(-\frac{1}{2}\tilde{\boldsymbol{\beta}}'_{j,k}\boldsymbol{K}_{j,k}\tilde{\boldsymbol{\beta}}_{j,k}\right) \mathbb{1}\left[\boldsymbol{A}_{j,k}\tilde{\boldsymbol{\beta}}_{j,k} = \mathbf{0}\right] \tag{M4}$$

and assume that the constraint matrix $\boldsymbol{A}_{j,k}$ is chosen such that all rank deficiencies in $\boldsymbol{K}_{j,k}$ are effectively removed by setting

$$\boldsymbol{A}_{j,k} = \text{span}\left(\ker(\boldsymbol{K}_{j,k})\right), \tag{M5}$$

where $\ker(\boldsymbol{K}_{j,k})$ denotes the null space of $\boldsymbol{K}_{j,k}$ and $\text{span}\left(\ker(\boldsymbol{K}_{j,k})\right)$ is a representation of the corresponding basis. Removing all rank deficiencies does not only remove the non-propriety from the prior but also allows to make the relation between the original and the parameter expansion more explicit and to perform effect decomposition for the components of the additive predictor.

Effect Decomposition Assuming the constraint matrix from (M4) effectively restricts the parameter vector $\tilde{\boldsymbol{\beta}}_{j,k}$ to a lower dimensional space of dimension $\kappa_{j,k} = \text{rk}(\boldsymbol{K}_{j,k})$ and therefore removes certain special cases of effects from what can be represented in the basis expansion. For example, for Bayesian P-splines with second order random walk prior, the rank of the prior precision matrix is $\kappa_{j,k} = D_{j,k} - 2$ and the null space corresponds to constant and linear effects. Applying the constrained prior allows to select constant, linear effects and non-linear deviations separately. In general, an effect $f_{j,k}(\boldsymbol{\nu})$ can be decomposed into one component $f_{j,k,\text{unpen}}(\boldsymbol{\nu})$ that corresponds to the null space of the prior precision matrix and the remainder $f_{j,k,\text{pen}}(\boldsymbol{\nu})$ as

$$f_{j,k}(\boldsymbol{\nu}) = f_{j,k,\text{unpen}}(\boldsymbol{\nu}) + f_{j,k,\text{pen}}(\boldsymbol{\nu}).$$

To achieve separate effect selection for the two components of f, we assign distinct spike and slab priors. A related idea for cubic splines was used in Rossell and Rubio (2019).

Remark 1

- The specifications (M3), (M4) and (M3*), (M4*) seem to be equivalent to each other corresponding to a simple rescaling of the regression coefficients and the prior distribution as $\boldsymbol{\beta}_{j,k} = \tau_{j,k}\tilde{\boldsymbol{\beta}}_{j,k}$. However, this is only true if the prior distribution (M4) is indeed proper. To see this, assume that $\boldsymbol{K}_{j,k}$ is rank deficient and a constant effect is not penalised by the prior precision matrix. In this case, the traditional formulation of structured additive regression models (M3*) implies a constant effect if $\tau^2_{j,k}$ approaches zero while the rescaled version (M3) implies an effect equal to zero since the complete function is multiplied by $\tau_{j,k}$.
- Note, that both (M4*) and (M4) rely on the same precision matrix $\boldsymbol{K}_{j,k}$ and hence the constraint matrix $\boldsymbol{A}_{j,k}$ can be constructed independently of the parametrisation. The traditional way is an explicit mixed model decomposition (Fahrmeir et al., 2004; Wood, 2011) which is used by Scheipl et al. (2012) to perform effect selection for mean regression models. However, the explicit mixed model representation used by Scheipl et al. (2012) destroys the sparsity properties of the design matrices (such as band structures for B-splines) and causes full design matrices which in turn increases computation times. In order to keep the sparsity of the design matrices of functional effects (and hence to minimize computation time) we instead implicitly remove the improper part of $p(\boldsymbol{\beta}_{j,k}|\tau^2_{j,k})$ by sampling $\boldsymbol{\beta}_{j,k}$ directly from the constrained posterior using (M4*).

Last, we highlight again that in combination with the distributional flexibility, we are the first to enable Bayesian effect selection in structured additive distributional regression. We show later the good performance of our approach in these but also simpler models.

Beta Prime Spike and Slab Prior on Squared Importance Parameter To achieve function selection in our model, we place a spike and slab prior specification on the squared importance parameter $\tau^2_{j,k}$. This hierarchical prior relies on a mixture of one prior concentrated around zero such that it can effectively be thought of as representing

zero (the spike component) and a more dispersed, mostly noninformative prior (the slab) and is specified via the hierarchy

$$\begin{aligned}
\tau^2_{j,k}|\delta_{j,k},\psi^2_{j,k} &\sim \text{Ga}\left(\frac{1}{2}, \frac{1}{2r_{j,k}(\delta_{j,k})\psi^2_{j,k}}\right),\\
\delta_{j,k}|\omega_{j,k} &\sim \text{Bi}(1,\omega_{j,k}),\\
\psi^2_{j,k} &\sim \text{IG}(a_{j,k},b_{j,k}),\\
\omega_{j,k} &\sim \text{Beta}(a_{0,j,k},b_{0,j,k}),\\
r_{j,k} \equiv r(\delta_{j,k}) &= \begin{cases} r_{j,k} > 0 \text{ small} & \delta_{j,k} = 0,\\ 1 & \delta_{j,k} = 1. \end{cases}
\end{aligned} \tag{M6}$$

The scale parameter $\psi^2_{j,k}$ determines the prior expectation of $\tau^2_{j,k}$, which is $\psi^2_{j,k}$ for $\delta_{j,k} = 1$ and $r_{j,k}\psi^2_{j,k}$ for $\delta = 0$ with $r_{j,k} \ll 1$ being a fixed small value and hence the indicator $\delta_{j,k}$ determines whether a specific effect $\boldsymbol{\beta}_{j,k} = \tau_{j,k}\tilde{\boldsymbol{\beta}}_{j,k}$ is included in the model ($\delta_{j,k} = 1$) or excluded from the model ($\delta_{j,k} = 0$). The parameter $\omega_{j,k}$ is the prior probability for an effect being included in the model and the remaining parameters $a_{j,k}$, $b_{j,k}$, $a_{0,j,k}$, $b_{0,j,k}$ and $r_{j,k}$ are hyperparameters of the spike and slab prior. We will discuss prior elicitation for these parameters in detail in Section 3.2.

Marginalising over $\psi^2_{j,k}$, both the spike and the slab component $p(\tau^2_{j,k}|\delta_{j,k})$ are scaled beta prime distributions with shape parameters 1/2 and $a_{j,k}$ and scale parameter $2r(\delta_{j,k})b_{j,k}$ (Pérez et al., 2017). Therefore we call the hierarchical prior on $\boldsymbol{\beta}_{j,k} = \tau_{j,k}\tilde{\boldsymbol{\beta}}_{j,k}$ specified by (M4)–(M6) the Normal Beta Prime Spike and Slab (NBPSS) prior, see Section 3 for a detailed discussion of the properties of the NBPSS prior. Equations (M1) to (M6) define our complete model specification for effect selection in structured additive distributional regression.

2.3 Special Cases

We briefly discuss some of the components $f_{j,k}$ used later in our empirical evaluations:

- For linear effects of continuous covariates, the columns of the design matrix $\boldsymbol{B}_{j,k}$ are equal to the different covariates. For binary/categorical covariates, the basis functions represent the chosen coding, e.g. dummy or effect coding and the design matrix then consists of the resulting dummy or effect coding columns. While for linear effects not under selection, flat improper priors (with $\boldsymbol{K}_{j,k} = \boldsymbol{0}$) are common standard, our effect selection prior corresponds to conditionally i.i.d. Gaussian priors. Informative Gaussian priors can also be used for effects not under selection to achieve a Bayesian ridge regression prior that enforces shrinkage of the effects towards zero (with $\boldsymbol{K}_{j,k} = \boldsymbol{I}$).
- For a non-linear effect of a continuous covariate x we employ Bayesian P-splines (Lang and Brezger, 2004). The i-th row of the design matrix $\boldsymbol{B}_{j,k}$ then contains the B-spline basis functions $B_{j,k,1}(x_i), \ldots, B_{j,k,D}(x_i)$ evaluated at the observed covariate value x_i. If not stated otherwise, we will use cubic B-splines with 20 inner

knots (resulting in effects of dimension $D = 22$) and second order random walk prior in all our empirical applications. The constrained prior removes constant and linear effects from the spline, as mentioned in Section 2.2.

- Spatial effects for a discrete set of geographical regions are modelled via Gaussian Markov random fields (GMRFs) with precision matrix given by an adjacency matrix encoding the neighbourhood relation between the regions (Rue and Held, 2005) and a design matrix with entries (i, s) equal to one if observation i is located in region s and zero otherwise. We consider the simplest form of GMRFs and define two regions as neighbours if they share common borders.
- Multilevel structured additive regression models as proposed by Lang et al. (2014) allow for hierarchical prior specifications for regression effects where each parameter vector may again be assigned an additive predictor $\tilde{\boldsymbol{\eta}}_{j,k}$, i.e. the vector $\boldsymbol{\beta}_{j,k}$ is decomposed as $\boldsymbol{\beta}_{j,k} = \tilde{\boldsymbol{\eta}}_{j,k} + \boldsymbol{\varepsilon}_{j,k}$.

3 Properties of the NBPSS prior

In the following, we discuss some relevant properties of the NBPSS prior hierarchy, including elicitation of hyperparameters, shrinkage properties and propriety of the posterior. For prior elicitation and shrinkage properties, the marginal distribution of $\boldsymbol{\beta}_{j,k} = \tau_{j,k}\tilde{\boldsymbol{\beta}}_{j,k}$ plays a crucial role. We will therefore start with deriving this marginal distribution. For notational convenience we drop the index for the distribution parameter k and the index for the effect j in Sections 3.1 and 3.2.

3.1 Marginal Distribution

The marginal prior for the squared importance parameter τ^2 is given by the mixture

$$p(\tau^2) = p(\tau^2|\delta = 1)\mathbb{P}(\delta = 1|a_0, b_0) + p(\tau^2|\delta = 0)\mathbb{P}(\delta = 0|a_0, b_0) \tag{2}$$

of two scaled beta prime distributions $\text{BP}(1/2, a, 2b)$ and $\text{BP}(1/2, a, 2rb)$ with mixture weight of the slab given by $\mathbb{P}(\delta = 1|a_0, b_0) = a_0/(a_0 + b_0)$. The NBPSS prior can alternatively be derived by assuming a mixture of two scaled t distributions for the importance parameter $\tau = \pm\sqrt{\tau^2}$. Specifying this prior hierarchically, the first equation in (M6) is replaced by $\tau|\delta, \psi^2 \sim \text{N}\left(0, r(\delta)\psi^2\right)$ and as a consequence posterior sampling for $\boldsymbol{\beta}$ would no longer be possible with Gibbs steps as the corresponding conditional posterior would depend on the likelihood function. Marginalising over ψ^2, δ and ω, the prior $p(\tau)$ is a mixture of two scaled t distributions with $2a$ degrees of freedom, location parameter 0, scale parameters b/a and rb/a and mixture weights $a_0/(a_0 + b_0)$ and $b_0/(a_0+b_0)$, respectively. Thus, the prior on the importance parameter τ is basically the normal-mixture-of-inverse gamma (NMIG) prior of Ishwaran and Rao (2005), who considered scalar regression coefficients $\boldsymbol{\beta}$ that are conditionally normal given the inverse gamma distributed variance parameter τ^2 (but with one level of hierarchy less) on the one hand, and, on the other hand related to the parameter-expanded NMIG (peNMIG) specification of Scheipl et al. (2012).

The implied marginal density for $\boldsymbol{\beta} = \tau\tilde{\boldsymbol{\beta}}$ can be derived as

$$p(\boldsymbol{\beta}) = \int_{-\infty}^{\infty} p(\tau) p_{\tilde{\boldsymbol{\beta}}}(\boldsymbol{\beta}/\tau)|\tau|^{-D} d\tau, \tag{3}$$

where $p_{\tilde{\boldsymbol{\beta}}}$ is given in equation (M4), and we approximate (3) numerically.

3.2 Prior Elicitation

In the following, we discuss prior elicitation for the NBPSS prior hyperparameters a, b, a_0, b_0 and r. More precisely, we argue that suitable default values can be suggested for a, a_0, and b_0 based on theoretical arguments while providing intuitive and user-friendly criteria for the elicitation of b and r. In the literature, default values have often been suggested from simulation-based evidence (e.g. in Scheipl et al., 2012) but we prefer to determine b and r in a more transparent and adaptive way.

Theoretical properties of the scaled beta prime distribution have been discussed in Pérez et al. (2017). From this, it follows that for both spike and slab moments of order less than a exist and the variance decreases with a. Furthermore, for small values of a, the spike and the slab component will overlap such that moves from $\delta = 0$ to $\delta = 1$ are possible. However to guarantee the existence of moments, a should not be too small either. We therefore recommend to set $a = 5$ as a default; but different values can be supplied by the user in our implementation.

For the prior inclusion parameter ω, a sensible default is to use $a_0 = b_0 = 1$ which corresponds to a flat prior on the unit interval. Of course, one can also choose fixed values for ω in case strong prior knowledge on the prior inclusion probability of the size of the expected model is available. As the marginal prior inclusion probability is given by $\mathbb{P}(\delta = 1|a_0, b_0) = a_0/(a_0 + b_0)$, a_0 and b_0 can be chosen to reflect prior assumptions on the inclusion probability of effects.

For the elicitation of b and r, we propose an approach inspired by the principled approaches of Simpson et al. (2017) and Klein and Kneib (2016). More precisely, we consider marginal probability statements on the supremum norm $\sup_{\boldsymbol{\nu}\in\mathcal{D}} |f(\boldsymbol{\nu})|$ over a certain set of covariate values $\mathcal{D}$ conditional on the status of the inclusion/exclusion parameter δ. Given $\delta = 1$ (inclusion of the effect), the marginal distribution of $f(\boldsymbol{\nu})$ does no longer depend on r, such that the parameter b can be determined from

$$\mathbb{P}\left(\sup_{\boldsymbol{\nu}\in\mathcal{D}} |f(\boldsymbol{\nu})| \leq c \,\middle|\, \delta = 1 \right) = \alpha. \tag{4}$$

This is the probability that the supremum norm of an effect is smaller than a pre-specified level c for all design points $\boldsymbol{\nu} \in \mathcal{D}$, such that α and c should be small. Basically we formulate the prior such that it is unlikely that the supremum norm of $|f(\boldsymbol{\nu})|$ stays below a pre-specified level if it is indeed an informative effect that should be included. Both the level c and the prior probability α have to be specified by the analyst according to her/his prior beliefs. To derive r, we proceed similarly but consider the probability

$$\mathbb{P}\left(\sup_{\boldsymbol{\nu}\in\mathcal{D}} |f(\boldsymbol{\nu})| \leq c \,\middle|\, \delta = 0 \right) = 1 - \alpha \tag{5}$$

now conditioning on non-inclusion. Since in this case we would rather be interested in making the probability of not exceeding the threshold c large, the probability is reversed to $1 - \alpha$. Note that the absolute value of the effects can be taken without loss of generality due to the centring constraint of each function to ensure identifiability.

The basic idea of these two equations is that such prior statements can be much more easily elicited in applications, in particular in distributional regression where the application of response functions such as the exponential function or the logit transform induce default ranges of plausible effect sizes. Of course, the levels c as well the probability levels α can be chosen to be distinct for the inclusion/exclusion criteria in (4) and (5) but we suppress this possibility notationally both for simplicity and since in most cases it seems plausible to choose the same parameter settings anyway.

To access the probabilities in (4) and (5), we derive the marginal distribution of $\sup|f(\boldsymbol{\nu})|$. For a single covariate value $\boldsymbol{\nu}$, the function evaluation is given by $f(\boldsymbol{\nu}) = \tau(B_1(\boldsymbol{\nu}), \ldots, B_D(\boldsymbol{\nu}))\tilde{\boldsymbol{\beta}} = \tau \boldsymbol{b}_\nu' \tilde{\boldsymbol{\beta}} = \boldsymbol{b}_\nu' \boldsymbol{\beta}$ and the marginal density is

$$p(\boldsymbol{b}_\nu' \boldsymbol{\beta} \mid \delta) = \int_0^\infty p(\boldsymbol{b}_\nu' \boldsymbol{\beta} | \tau^2) p(\tau^2 \mid \delta) \mathrm{d}\tau^2,$$

where $\boldsymbol{b}_\nu' \boldsymbol{\beta} | \tau^2 \sim \mathrm{N}(0, \tau^2 \boldsymbol{b}_\nu' \boldsymbol{K}^- \boldsymbol{b}_\nu)$ (with $\boldsymbol{K}^-$ denoting the generalized inverse of $\boldsymbol{K}$) and $p(\tau^2)$ is given in (2). Note that using the generalized inverse effectively removes the portion of $f(\boldsymbol{\nu})$ that corresponds to the null space of $\boldsymbol{K}$ such that we take the constraint in (M4) into account. The integrals above are scalar integrals for each covariate $\boldsymbol{\nu}$ which can be solved numerically. However, obtaining the supremum over a large set $\mathcal{D}$, numerical integration easily becomes computationally intractable. We hence determine the distribution of the supremum based on simulations from the hierarchical NBPSS prior. In the Online Appendix B (Klein et al., 2020), we show how to determine r and b independently of each other. For given design matrix $\boldsymbol{B} = (\boldsymbol{b}_{\nu_1}', \ldots, \boldsymbol{b}_{\nu_n}')'$, precision matrix $\boldsymbol{K}$, probability level α and threshold c, these can be computed for general functional effects using the R package `sdPrior` (Klein, 2018).

Of course, the elicitation rules described in this section can similarly be used to elicit hyperparameters for other types of hierarchical prior structures as discussed for example in Klein and Kneib (2016) for the specific case of scale-dependent hyperpriors. In particular, one may consider informative regularization priors also for the fixed effects or the unpenalized parts of nonparametric effects not under selection. Especially in situations with sparse data, this may have beneficial effects on mixing and convergence of the MCMC sampling scheme. Last, we recommend standardizing the continuous covariates included with linear, parametric effects since this enhances numerical stability of the optimization routine.

3.3 Shrinkage Properties

Regularisation and shrinkage properties of certain prior settings in regression specifications can be studied by considering the marginal distribution of the regression coefficients and/or functional effects.

Constraint Regions We compare the prior specified in (M4)–(M6) with a standard NMIG prior applied directly to the coefficients in $\boldsymbol{\beta}$ and the parameter expanded prior (peNMIG) of Scheipl et al. (2012).

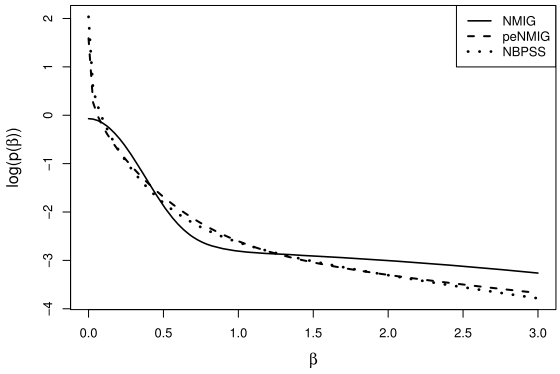

Figure 1: Univariate marginal log-densities for a standard NMIG prior (solid line), the peNMIG prior of Scheipl et al. (2012, dashed line) and the NBPSS prior (dotted line). Hyperparameters are set to $a_0 = b_0 = 1$, $a = 5$, $b = 50$, $r = 0.005$.

Figure 1 shows the univariate marginal log-densities where the most distinct difference is between the standard NMIG prior compared to the peNMIG and NBPSS priors. While the standard NMIG prior resembles the shape of a normal distribution with a finite asymptote at zero, both parameter expanded priors feature a spike in zero. As we will show and discuss in the next section, this spike is indeed infinite. Figure 2 supplements the univariate considerations by bivariate marginal log-densities. We differentiate between two situations: First, we consider two parameters that depend on the same value τ^2, i.e. parameters belonging to the same function $f(\boldsymbol{\nu})$, while in the second case we consider parameters depending on different importance parameters. This distinction is important since the standard NMIG prior always assumes independent components with separate hyperparameters. As a consequence, the peNMIG and NBPSS priors deviate from the standard situation in two ways: First by the parameter expansion itself and second by making the parameters depend on the same hyperparameter. To disentangle the effect of these two deviations, we rely on the separate presentations. We make the following important observations:

- The NBPSS and peNMIG priors share the same qualitative behaviour while deviating considerably from the standard NMIG prior regardless of whether the case of shared or distinct τ^2 is considered.

- The univariate marginal densities qualitatively resemble the ones of the original spike and slab prior of Mitchell and Beauchamp (1988) with tails that are heavy enough to induce a re-descending score function which ensures robustness of the Bayesian estimators (see also the next subsection).

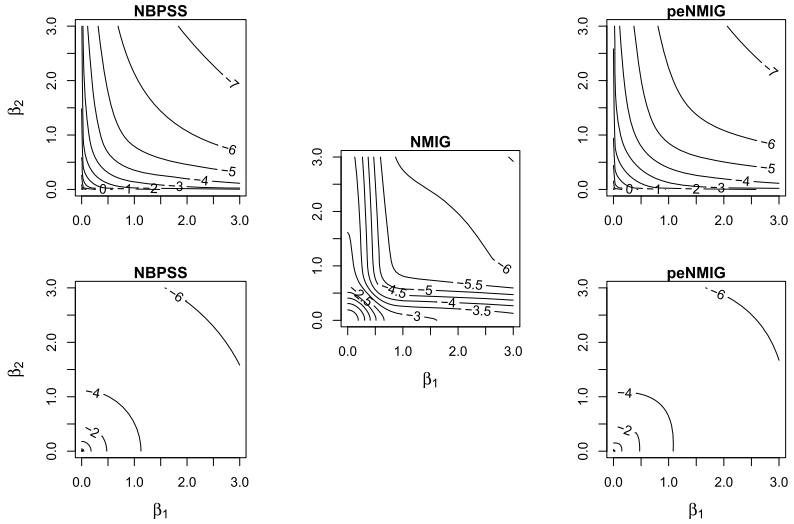

Figure 2: Contour lines of bivariate marginal log-densities for a standard NMIG prior (middle panel), the peNMIG (right column) and the NBPSS prior (left column). The first row panels show results for parameters with distinct hyperparameters and the second row panels show results for parameters sharing the same τ. For the standard NMIG, the hyperparameters are by construction assumed to be distinct and no changes in the row are possible.

- For distinct parameters, we observe contours similar to the convex shape of L_q priors with $q < 1$ for the peNMIG and NBPSS priors which implies weak shrinkage of large effects while small coefficients are strongly shrunken to zero.

- For the case of shared τ^2, the shapes of the contours imply simultaneous shrinkage of both parameters instead of the strong shrinkage towards the coordinate axes observed for distinct importance parameters. This is exactly the desired type of shrinkage for parameters belonging to one effect $f(\boldsymbol{\nu})$ to completely remove the effect from the model specification.

- As already noted in Section 2.2, the specification of the prior in Scheipl et al. (2012) differs from ours insofar as they consider the mixed model decomposition of effects. Additionally, Scheipl et al. (2012) use a bimodal prior for the standardized regression effects with modes at $+1$ and -1. This effectively bounds the coefficients away from zero and thus encourages sampling from one mode of the posterior, while we instead explore the full posterior. Consequently, the conditional posterior of $\tilde{\boldsymbol{\beta}}$ of NBPSS is a standard normal distribution $p_{\text{NBPSS}}(x) = \text{N}(x; 0, 1)$, while the one of peNMIG is a mixture of two normals with modes, $p_{\text{peNMIG}}(x) = 0.5\,\text{N}(x; 1, 1) + 0.5\,\text{N}(x; -1, 1)$. Taking the ratio explains the slightly heavier tails of peNMIG in Figures 1 and 2.

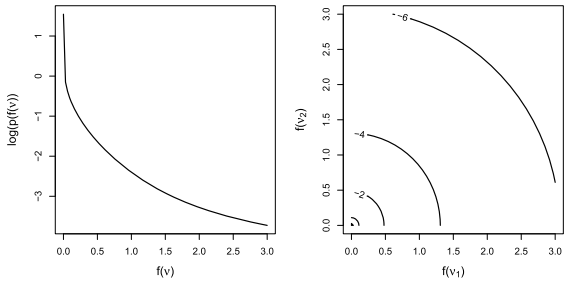

Figure 3: Univariate (left) and bivariate (right) marginal log-densities of $f(\boldsymbol{\nu})$. The hyperparameters have been fixed at $a = 5$, $b = 50$, $r = 0.005$ and $a_0 = b_0 = 1$.

We also study the implied constraint regions for the marginal prior of function evaluations $f(\boldsymbol{\nu}) = \boldsymbol{b}_\nu'\boldsymbol{\beta}$, which can be derived in complete analogy by utilising that $\boldsymbol{b}_\nu'\tilde{\boldsymbol{\beta}} \sim \text{N}(0, \boldsymbol{b}_\nu'\boldsymbol{K}^-\boldsymbol{b}_\nu)$ with a generalised inverse $\boldsymbol{K}^-$. In contrast, the marginal prior for the peNMIG prior is not numerically accessible since it involves a complex mixture of 2^D components (where D is the dimension of $\boldsymbol{\beta}$) due to the bimodal prior for the elements of $\tilde{\boldsymbol{\beta}}$. Figure 3 depicts marginal densities for the effect $f(\boldsymbol{\nu})$ evaluated at one (left panel) and two (right hand panel) randomly chosen covariate values of a sequence of $n = 100$ equidistant values in $[-\pi, \pi]$. The resulting design matrix $\boldsymbol{B}$ is based on cubic Bayesian P-splines with $D = \dim(\boldsymbol{\beta}) = 22$. Hence, the bivariate plot corresponds to the situation of one shared importance parameter since we are interested in shrinkage of the effect evaluations for the same effect at different covariate values. Qualitatively, the behaviour from the marginal densities of the regression coefficient is translated to the function evaluations, i.e. we observe a peak in zero and simultaneous shrinkage.

Tail Behaviour and Behaviour in the Origin Visually, the marginal prior for $\boldsymbol{\beta}$ features a distinct peak at zero as shown in the previous section. We now investigate more closely, whether this spike is finite or infinite by considering the behaviour of $p_{\boldsymbol{\beta}}(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}}$. Using (3) we obtain

$$\begin{aligned} p(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}} &= 2p_{\tilde{\boldsymbol{\beta}}}(\mathbf{0}) \left(\int_0^1 \underbrace{p_\tau(\tau)}_{\geq p_\tau(1)} \tau^{-D} d\tau + \underbrace{\int_1^\infty p_\tau(\tau)\tau^{-D} d\tau}_{\geq 0} \right) \\ &\geq 2p_\tau(1)p_{\tilde{\boldsymbol{\beta}}}(\mathbf{0}) \int_0^1 \tau^{-D} d\tau = \infty, \end{aligned}$$

since $\int_0^1 \tau^{-D} d\tau$ diverges for $D \geq 1$ and therefore the marginal prior for $\boldsymbol{\beta}$ indeed has an infinite spike in zero. Note that we have shown that the multivariate parameter expanded prior has a spike in zero, while Scheipl et al. (2012) have only shown the result for the univariate marginal prior. This together with heavy tails can be considered to induce particularly beneficial shrinkage properties, similar to the horseshoe prior and

also the normal-Jeffreys' prior, for which both robustness for large effects and very efficient estimation of sparse coefficient vectors have been shown (Polson and Scott, 2010). Van der Pas et al. (2016) specified conditions on the prior distribution in the sparse multivariate normal means model to obtain posterior contraction at the minimax rate. These conditions require tails at least as heavy as Laplace but not too heavy to recover both nonzero and zero means with the optimal rate, see also Ročková (2018) for a detailed investigation of the spike and slab lasso. However, Condition (3) of Van der Pas et al. (2016) is not fulfilled for the NBPSS prior with the current choice of hyperparameters. For the Bayesian linear regression model with Gaussian errors, (near) minimax posterior contraction rates were obtained recently in Castillo et al. (2015) for a spike and slab prior with discrete spike and Laplace slab, in Ročková and George (2018) for the spike and slab lasso, and in Bai and Ghosh (2019) for a (one-component) Normal beta prime prior. Yet no results are available for distributional regression models.

The tail behaviour of the marginal prior for $\boldsymbol{\beta}$ can be studied by looking at the score function of $p(\boldsymbol{\beta})$ which consists of the elements

$$\frac{\partial}{\partial \beta_d} p_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = -\int p_\tau(\tau) p_{\tilde{\boldsymbol{\beta}}}(\boldsymbol{\beta}/\tau) \frac{\beta_d}{\tau^2} |\tau|^{-D} d\tau.$$

Figure A.1 of the Online Appendix visualizes the resulting score function and compares it to the score function of the NMIG and peNMIG priors. From the graphical representation we find that all three prior structures have heavy tails such that the score functions are re-descending (i.e. they approach zero as their argument tends to infinity) which induces Bayesian robustness of the resulting estimates. The score functions of the peNMIG and NBPSS priors resemble the shape of L_q priors with q close to zero, while the shape of the score function for the NMIG prior shows a more complex non-monotonously shape around zero.

3.4 Properties of the Posterior

Propriety Building on the theoretical results derived in Klein et al. (2015b) and Klein and Kneib (2016), we obtain sufficient (and in the special case of Gaussian mean regression also necessary) conditions for the propriety of the posterior under our Bayesian effect selection specification. While in Section 2 we avoid to explicitly reparameterize the design matrices to remove the null space of the precision matrices $\boldsymbol{K}_{j,k}$ (both for effects with NBPSS prior and for effects not under selection), we employ a mixed model representation of the predictors $\boldsymbol{\eta}_k$ in (M2) here as this greatly simplifies the derivation of sufficient conditions for the propriety of the posterior. To keep the presentation concise, we refer the reader to the Online Appendix A for the general strategy, the main differences to the theoretical results derived in Klein et al. (2015b) and Klein and Kneib (2016) and detailed proofs.

Consistency In general, posterior consistency is another important property which has been studied extensively in the literature for a variety of variable selection priors. However, most papers deal with the Gaussian linear model under various scenarios (too many to give an extensive literature overview here). For instance, for classical

spike and slab priors such as Gaussian mixtures (George and Mc Culloch, 1993) and Laplace mixtures (Ročková, 2018; Ročková and George, 2018) the asymptotic theory of Castillo and van der Vaart A. (2012) and Ročková (2018) suggests that using $a_0 = 1$ and $b_0 = p$ yields optimal posterior concentration when the number of unknown parameters p is $o(n)$. In contrast, in our model the number of all regression parameters p (which is far larger than the number of covariates) in the K distributional parameters is typically considerably smaller than n. Hence, the complexity of our model is mostly due to specifying a very flexible, semiparametric regression model for each distributional parameter rather than the large number of covariates in the predictors.

For instance, ignoring smoothing in $\boldsymbol{\beta}$ through the multivariate normal priors, p would be computed as the sum of all coefficients in each of the functional effects, i.e. $p = \sum_{k=1}^{K} \sum_{j=1}^{J_k+L_k} D_{j,k}$, where $D_{j,k}$ is the number of basis functions of the j-th effect $f_{j,k}$ in predictor η_k (which is associated to distribution parameter ϑ_k). However, due to the shrinkage in $\boldsymbol{\beta}$ and due to our main goal of general effect selection rather than the selection of single coefficients, the more interesting case is to count the total number of effects $\tilde{p} = \sum_{k=1}^{K}(J_k + L_k)$.

Keeping these tasks in mind, unfortunately, the available literature is rare. A first start could be the work of Rossell and Rubio (2019), who consider non-local priors with discrete spike for survival models but allow for non-linear regression splines without smoothing. Interestingly, these authors show that a similar effect decomposition as we employ can help gaining efficiency. Still, the discussion above implies a number of differences to our setting, such that we leave studying the posterior consistency in depth in our model class for future work.

4 Posterior Estimation

In this section, we describe the basic steps of our Markov chain Monte Carlo (MCMC) sampler. In each MCMC sweep $m = 1, \ldots, M$, this sampler involves two main loops; one over the distribution parameters $k = 1, \ldots, K$ and a second inner loop over the all $l = 1, \ldots, L_k$ and $j = 1, \ldots, J_k$ effects in predictor k, see Algorithm 1 of the Online Appendix C.2.

Update of the Basis Coefficients Due to the modular structure of MCMC simulation algorithms, no changes in the MCMC scheme developed by Klein et al. (2015c) are required for updating the basis coefficients $\boldsymbol{\beta}_{j,k}$ when supplementing them with a NBPSS prior on τ^2 instead of the standard inverse gamma prior. We therefore apply iteratively weighted least squares based approximations to the log full conditional and generate proposals from the multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}_{j,k}, \boldsymbol{P}_{j,k}^{-1})$ with expectation and precision matrix given by

$$\boldsymbol{\mu}_{j,k} = \boldsymbol{P}_{j,k}^{-1}\boldsymbol{B}_{j,k}'\boldsymbol{W}_{j,k}(\tilde{\boldsymbol{y}}_{j,k} - \boldsymbol{\eta}_{-j,k}), \qquad \boldsymbol{P}_{j,k} = \boldsymbol{B}_{j,k}'\boldsymbol{W}_{j,k}\boldsymbol{B}_{j,k} + \frac{1}{\tau_{j,k}^2}\boldsymbol{K}_{j,k}, \tag{6}$$

where $\boldsymbol{\eta}_{-j,k} = \boldsymbol{\eta}_{j,k} - \boldsymbol{B}_{j,k}\boldsymbol{\beta}_{j,k}$ is the predictor without the effect currently updated and the working observations $\tilde{\boldsymbol{y}}_{j,k}$ and weights $\boldsymbol{W}_{j,k}$ are determined based on first and second derivatives of the log-likelihood with respect to the predictor.

Update of the Smoothing Variance for Effects not Subject to Selection For effects not subject to selection, we consider an inverse gamma prior $\tau^2_{j,k} \sim \text{IG}(a_{j,k}, b_{j,k})$ for the smoothing variances such that the update of $\tau^2_{j,k}$ can be done via a simple Gibbs sampling step drawing from $\tau^2_{j,k}|\cdot \sim \text{IG}(a'_{j,k}, b'_{j,k})$, with updated parameters $a'_{j,k} = \frac{\text{rk}(\boldsymbol{K}_{j,k})}{2} + a_{j,k}$, $b'_{j,k} = \frac{1}{2}\boldsymbol{\beta}'_{j,k}\boldsymbol{K}_{j,k}\boldsymbol{\beta}_{j,k} + b_{j,k}$.

Update of the Squared Importance Parameter for Effects Subject to Selection The full conditional $p(\tau^2_{j,k}|\boldsymbol{\beta}_{j,k}, \delta_{j,k}, \psi^2_{j,k})$ is a generalised inverse Gaussian distribution $\text{GIG}(p, q, c)$, with $p = -0.5\,\text{rk}(\boldsymbol{K}_{j,k}) + 0.5$, $q = 1/(r(\delta_{j,k})\psi^2_{j,k})$, $c = \boldsymbol{\beta}'_{j,k}\boldsymbol{K}_{j,k}\boldsymbol{\beta}_{j,k}$. This has the advantage that τ^2 can be generated independently of the likelihood in an efficient Gibbs step which is no longer possible when the prior is formulated for the importance parameter τ as in (Scheipl et al., 2012) where a Metropolis-Hastings update is required, see the Online Appendix C.1.

Updates for the Hyperparameters of the NBPSS prior For the hyperparameters of the NBPSS prior, we obtain Gibbs sampling steps via the following full conditionals:

- Inclusion indicator $\delta_{j,k}$:

$$p(\delta_{j,k} = 1|\cdot) = \frac{1}{1 + \frac{\varphi(\tau_{j,k};0;r_{j,k}\psi^2_{j,k})(1-\omega_{j,k})}{\varphi(\tau_{j,k};0;\psi^2_{j,k})\omega_{j,k}}},$$

  with $\varphi(\cdot;\mu,\sigma^2)$ the density of a normal distribution with mean μ and variance σ^2.

- Hyper-variance $\psi^2_{j,k}$: $\psi^2_{j,k}|\cdot \sim \text{IG}\big(a_{j,k} + 0.5, b_{j,k} + \frac{\tau^2_{j,k}}{2r_{j,k}(\delta_{j,k})}\big)$.

- Inclusion probability $\omega_{j,k}$: $\omega_{j,k}|\cdot \sim \text{Beta}(a_{0,j,k} + \delta_{j,k}, b_{0,j,k} + 1 - \delta_{j,k})$.

Implementation Implementation was done in a developer version of BayesX (Belitz et al., 2015), which is available from the authors on request. The software makes use of methods for efficient storing of large data sets and sparse matrix algorithms for sampling from multivariate Gaussian distributions and also allows us to access existing procedures for example for computing simultaneous credible bands for nonparametric effects (Krivobokova et al., 2010). Hyperparameter elicitation is integrated in the R-package `sdPrior` (Klein, 2018).

5 Empirical Evaluations

5.1 Simulations

To evaluate the performance of the NBPSS prior for effect selection in distributional regression, we conducted extensive simulations under various settings. We distinguish different scenarios for the predictor complexity, models including and excluding spatial effects, four selected response distributions, varying sample sizes, correlated and uncorrelated covariates and a set of user-defined parameters for hyperprior elicitation.

Simulation Design

- We consider Gaussian responses with effects only on the expectation, a Gaussian location-scale model, Poisson regression and zero-inflated Poisson models.
- We specify four test functions $f_1(x) = x$, $f_2(x) = x + \frac{(2x-2)^2}{5.5}$, $f_3(x) = -x + \pi\sin(\pi x)$ and $f_4(x) = 0.5x + 15\phi(2(x-0.2)) - \phi(x+0.4)$.
- We distinguish two scenarios in terms of the predictor complexity:
  - **low sparsity** in which out of 16 included covariates 12 have non-zero influence. The true linear predictor is $\eta = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + 1.5\,(f_1(x_5) + f_2(x_6) + f_3(x_7) + f_4(x_8)) + 2(f_1(x_9) + f_2(x_{10}) + f_3(x_{11}) + f_4(x_{12}))$ and we simulate the two cases with and without additional spatial effect $f_{spat}(s)$, labeled as "spatial/non-spatial". These settings are used for η_μ in the homoscedastic Gaussian and the Gaussian location-scale model, as well as for η_λ in the Poisson and the zero-inflated Poisson model.
  - **high sparsity** in which out of eight included covariates four have non-zero influence. The true linear predictor is $\eta = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$ and we again simulate the two cases with and without additional spatial effect $f_{spat}(s)$. These settings are used for η_{σ^2} in the Gaussian location-scale model and for η_π in the zero-inflated Poisson model.
- We generate the covariates x either as i.i.d. realizations from $U[-2, 2]$ or from an $AR(1)$ process with correlation $\rho = 0.7$ and standardize x in order to facilitate prior elicitation.
- We simulate 150 replications for each combination of the settings.
- We use six combinations of α and c for the elicitation of the prior hyperparameters b and r arising from the pairwise combination of $\alpha = 0.05, 0.1, 0.2$, and $c = 0.1, 0.2$.
- We consider the sample sizes $n = 200; 1{,}000$ for Gaussian, $n = 500; 2{,}000$ for Poisson, $n = 1{,}000; 2{,}000$ for Gaussian location-scale and zero-inflated Poisson responses. The sample sizes have been chosen to reflect a challenging (small sample size) and a relatively informative (large sample size) setting, taking the different complexity of the model structures into account.

As a competitor for the single parameter distributions Gaussian and Poisson, we consider the peNMIG prior of Scheipl et al. (2012) implemented in the R-package `spikeSlabGAM` (Scheipl, 2016). We refrain from comparison with further variable selection priors mentioned in the introduction as these usually lack applicability beyond the framework of generalized linear models. Hyperparameter elicitation for the NBPSS prior was performed with the package `sdPrior` (Klein, 2018) and estimation was done with the current developer version of BayesX (Belitz et al., 2015).

Results In the following, we restrict ourselves to the main conclusions, a detailed description about simulation settings and evaluation including complete graphical evidence is provided in the Online Appendix D.

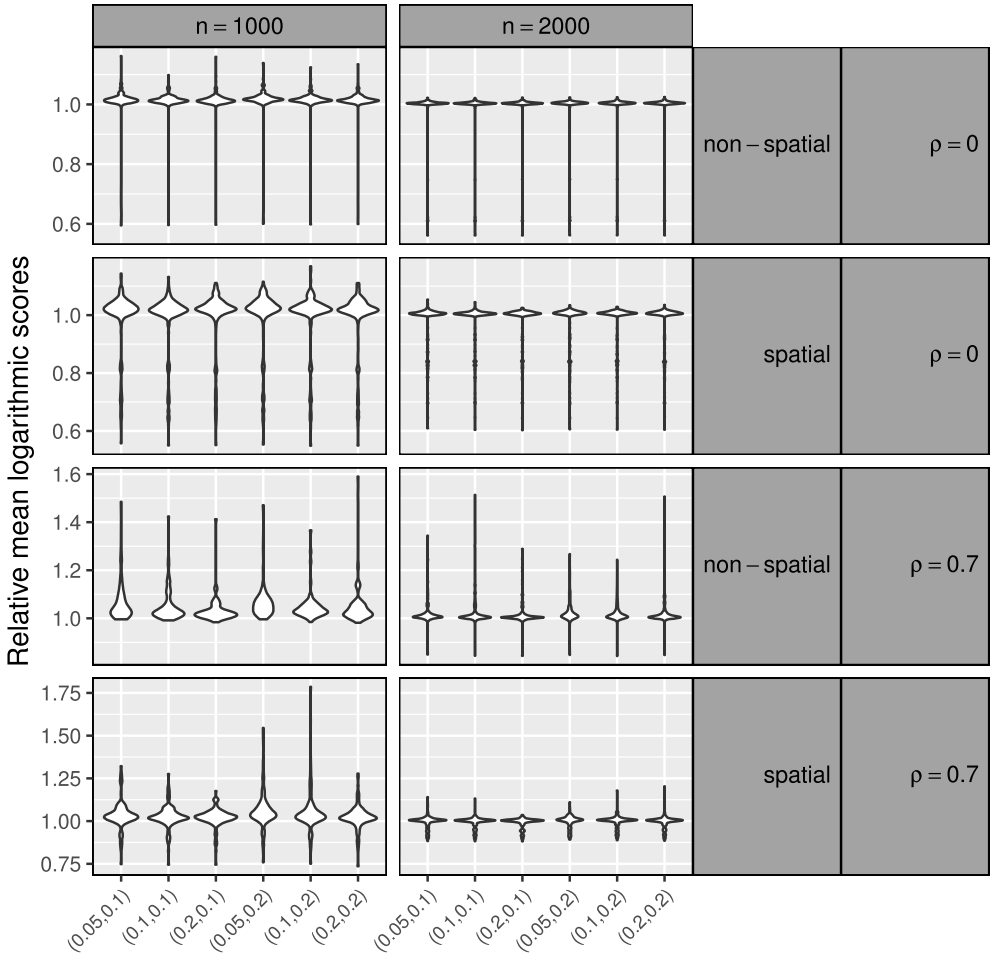

Figure 4: Violin plots of relative mean log-scores (MLS), i.e. MLS with NBPSS prior divided by MLS of oracle model, in the zero-inflated Poisson model. The log-scores are averaged over 5,000 new test data observations for each replicate. Columns are sample sizes $n = 1{,}000; 2{,}000$, rows 1 and 3 belong to the non-spatial scenarios, rows 2 and 4 to the spatial ones. Covariates are uncorrelated in rows 1 and 2 and correlated in rows 3 and 4. Boxplots within a column/row correspond to different combinations of (α, c).

- As a general outcome, the NBPSS prior results in very good performance for the selection of relevant effects even in challenging distributional regression settings with effect selection on multiple distributional parameters, where no competing Bayesian variable selection approach is available so far. Evidence for that is given in Figure 4 and the Figures in the Online Appendix Part D, showing posterior inclusion probabilities and the ratio between predictive NBPSS log-scores and oracle log-scores (i.e. log-scores arising from a model with given, true predictor

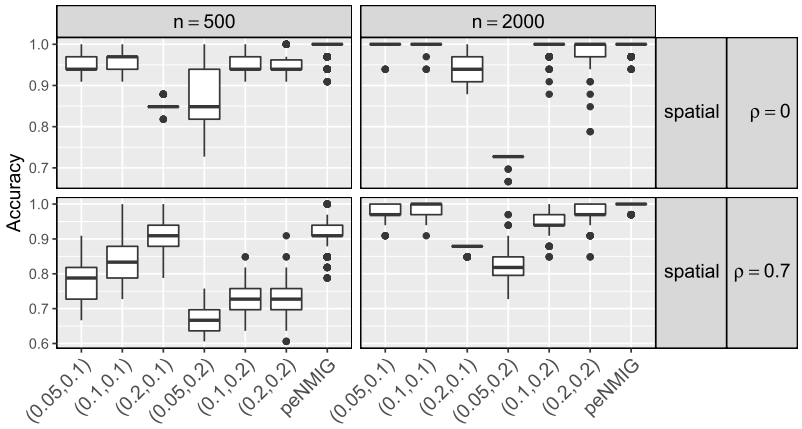

Figure 5: Overall accuracy (measured by the sum of true positives and true negatives divided by the total number of effects) for the Poisson model. Columns are sample sizes $n = 500; 2{,}000$. Covariates are uncorrelated in row 1 and correlated in row 2 and both include a spatial effect (see the Online Appendix D for the non-spatial scenarios). The boxplot on the right of each subplot shows the peNMIG prior, the remaining ones correspond to different choices for (α, c) and the NBPSS prior.

specification), respectively, in the zero-inflated Poisson model. The log-scores have been computed from test data sets with 5,000 observations.

- In the simple exponential family framework with only one single regression predictor, the NBPSS prior turns out to be a strong competitor to the peNMIG prior (see Figure 5 for overall accuracy results of the Poisson model). Selection of large coefficient blocks such as spatial effects works well for all types of response distributions, while these are particularly problematic with peNMIG due to severe mixing problems. On the other hand, the explicit reparameterisation of non-linear effects used with the peNMIG prior (as compared to the constrained sampling approach that NBPSS is based on) seems to have some advantages in separating the linear and non-linear part of non-linear effects in cases where the true effect is close to linear and at the same time covariates are strongly correlated.

- To further investigate effect separation in different distribution parameters, we conducted an experiment in Subsection D.5 of the Online Appendix with two scenarios: One where the same covariate enters both the mean and the variance predictor and one, where the covariate is just added as noise. Based on a threshold of 0.5 for the posterior mean inclusion probability (the median probability model, Barbieri and Berger, 2004), most replications would select the correct model. There are slightly more false positives than false negatives. Crucially though, false positives and false negatives in one predictor do not seem to (spuriously) affect the respective posterior mean inclusion probability in the other predictor; and misclassification occurs only for a small proportion of replicates.

- Coinciding with previous evidence on Bayesian effect selection, we find a strong impact of hyperprior parameter choice on the resulting effect selection performance. Our interpretable yet flexible way of eliciting hyperprior parameters equips data analysts with an intuitive approach for choosing these hyperparameters. More precisely, changing the probability α and the threshold c can help to balance between the true positive and false negative rates of effect selection. Choosing α and c smaller, results in more conservative, i.e. sparser models. Based on our simulations, we suggest $\alpha = c = 0.1$ as default values in our applications.

5.2 Applications

In this section, we demonstrate the efficacy of the NBPSS prior and its applicability for non-Gaussian, discrete or multivariate data. Core information about the different data sets *Patents*, *Nigeria* and *House prices* can be found in Table 1. Note here that all our examples have a large number of unknown coefficients, since each effect in a predictor typically induces a complete vector of coefficients and each predictor (of which we have K as distributional parameters) is an additive decomposition of such effects.

Data set	n	$\sum_{k=1}^{K} J_k$	distribution	time
Nigeria	23,042	108	bivariate normal	5.92 min
Patents	4,805	22	zero-inflated Poisson	0.25 min
House prices	98,354	26	Gaussian location-scale	3.75 min

Table 1: Summaries for the data sets *Patents*, *Nigeria*, and *House prices*. Columns 2 to 4 show the number of observations, number of effects in all predictors in the full model and the distribution for the response. The last column reports the computing time required for estimating 1,000 subsequent MCMC sweeps with the NBPSS prior.

Bivariate Analysis of Undernutrition The *Nigeria* data have been extracted from Demographic and Health Surveys (DHS, https://dhsprogram.com/) containing nationally representative information about the population's health and nutrition status in numerous developing and transition countries. Here we use data from Nigeria collected in 2013. Overall there are 23,042 observations after removing outliers and inconsistent observations from the data, see Table E.1 of the Online Appendix for a full description of variables. We use *stunting* and *wasting* as the bivariate response vector, where *stunting* refers to stunted growth measured as insufficient height of the child with respect to its age (chronic undernutrition), while *wasting* refers to insufficient weight for height (acute undernutrition). We assume the two indicators are jointly normally distributed with marginal means, scales and correlation parameter depending on covariates. Specifically, the model equations for all $K = 5$ predictors of the distributional parameters are specified according to (1). The four non-linear effects f_1 to f_4 of *cage* (age of the child in months), *edupartner* (years of partner's education), *mage* (age of the mother in years), *mbmi* (body mass index of the mother) are decomposed into their linear and non-linear part as described in Section 2.2. For the scale parameters, we used an exponential response function and for ρ the response function $h(x) = x/\sqrt{(1+x^2)}$. The deviance information criterion (DIC)/Watanabe-Akaike information criterion (WAIC)

Figure 6: Nigeria: Posterior means and 95% credible intervals for the linear effects of all model parameters (left column for stunting, right column for wasting, top row for ρ, middle row for σ, bottom row for μ) together with posterior inclusion probabilities (those with $\mathbb{P}(\delta|\cdot) > 0.5$ are highlighted in bold). Since ρ acts on both responses, the effects are only shown in the first column. Red corresponds to results for the NBPSS prior and blue to the full model with inverse gamma prior for the hypervariances.

of the full model and model with NBPSS prior are 159,101/159,190 and 159,101/159,173, respectively and hence slightly better for the NBPSS prior model.

Figures 6 and 7 show the posterior means together with their 95% posterior credible intervals of linear and non-linear effects for the full model (blue) and the model with NBPSS prior (red) as well as posterior inclusion probabilities $\mathbb{P}(\delta|\cdot) = 1$. For the function estimates $f_{j,k} = f_{j,k,lin} + f_{j,k,nonlin}$, Figure 7 shows the corresponding non-linear part

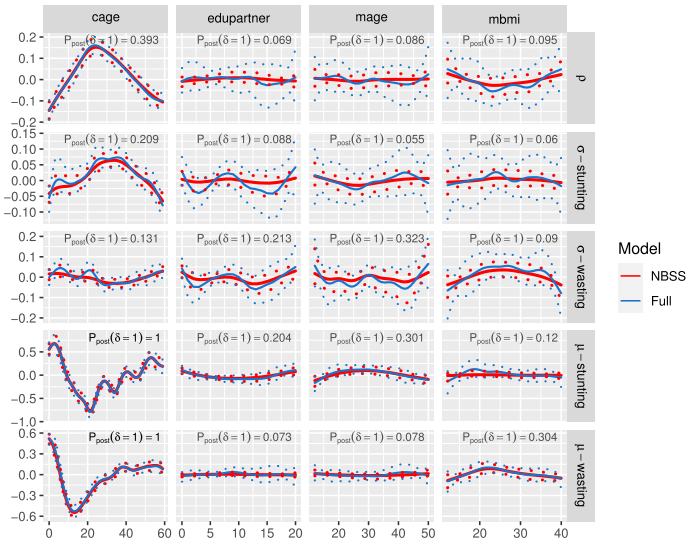

Figure 7: Posterior mean estimates and 95% credible intervals for non-linear effects of *cage*, *edupartner*, *mage*, *mbmi* (column-wise) in $\rho, \sigma_{\text{stunting}}, \sigma_{\text{wasting}}, \mu_{\text{stunting}}, \mu_{\text{wasting}}$ (row-wise) together with posterior inclusion probabilities. Red corresponds to the NBPSS prior and blue to the full model.

$f_{j,k,nonlin}$ separately from the linear part $f_{j,k,lin}$ in Figure 6, while the sum of the two components can be found in the Online Appendix E. We see that both models yield very similar point estimates, however the NBPSS prior results in smoother estimates (against overfitting) and more narrow credible intervals. In Section D.6 of the Online Appendix we empirically find that posterior coverage is appropriate when employing the NBPSS prior: Coverage rates for estimation without NBPSS prior are far too conservative, while the ones from the NBPSS prior are closer to the nominal level but with a slight tendency to be anti-conservative. Spatial effects of the five distribution parameters with the NBPSS prior are visualized in the Online Appendix E. While we omit the ones of the full model, tendencies are similar as for the remaining effects.

Based on marginal inclusion probabilities in the Online Appendix E, we find that the regional effect is relevant in all distribution parameters, i.e. not only the marginal means but also the scales and the correlation between *stunting* and *wasting*. Interestingly, chronic undernutrition measured by *stunting* seems to be mostly driven by variables describing the life situation of the children. In contrast, besides the region of residence, the mother's nutritional status measured by *mbmi* has a relevant effect only for acute undernutrition (*wasting*).

Number of Patent Citations The *Patents* data set contains the number of citations of patents granted by the European Patent Office (EPO). An inventor who applies for a patent has to cite all related, already existing patents his patent is based on. Klein et al.

(2015b) use this data set to illustrate their developed methodology on Bayesian zero-inflated and overdispersed count data and conducted variable selection in a stepwise forward approach based on the deviance information criterion (DIC). We focus on zero-inflated Poisson (ZIP) models with a detailed analysis in the Online Appendix Part F. From that we can conclude that the ZIP model with effect selection through the NBPSS is clearly favoured over the stepwise forward selection of Klein et al. (2015b) in terms of various predictive criteria.

Hedonic House Prices Understanding determinants not only of expected house prices but also their variability is important for the risk management of financial institutions relying on real estate as a part of their portfolio. Distributional regression relying, for example, on a normal response model for price per square metre is therefore a very promising approach since it allows to identify not only covariates related to expected house prices but also to the variability of house prices around their expectation. An additional complication often arising in the regression-based, hedonic approach to house price assessment is spatially structured but at least partially unobserved heterogeneity that requires the inclusion of spatial effects in a hierarchical multilevel regression model. We employ a Gaussian hierarchical location-scale model, see the Online Appendix Part G for details. In summary, we find that the NBPSS prior demonstrates its effect selection and shrinkage and regularization abilities also in hierarchical settings.

6 Summary and Discussion

In this paper, we have introduced Bayesian effect selection based on spike and slab priors to the class of structured additive distributional regression models. We considered a constrained prior construction that enables effect decomposition and efficient computations via sparse matrix structures, provided simple rules for prior elicitation and derived shrinkage properties of the NBPSS prior highlighting its favourable properties. In simulations, we have demonstrated that the NBPSS prior is applicable even to the selection of high-dimensional coefficient blocks in more than one distribution parameter. The method promises wide applicability, which we illustrate along three different examples including zero-inflated count data, a bivariate Gaussian model and a hierarchical location-scale specification for hedonic housing prices. Instead of arbitrarily fixing hyperparameters of the inverse gamma priors, we have suggested an intuitive and interpretable way for hyperprior elicitation, which is easily accessible by applied users. This is an important feature since results react sensitively with respect to the actual choices of hyperparameters.

Yet, the NBPSS prior controls the flexibility of each effect separately since priors are assumed to be independent and does not allow to control the overall complexity of the predictor. However, the NBPSS prior could be extended to achieve also global shrinkage properties, e.g. by specifying the scale parameter in the prior on τ^2 as a product of a global and a local parameter (similar as in Polson and Scott, 2010). Another direction for future research is the consideration of hierarchical spike and slab priors based on our parameter expansion and the NBPSS prior for the importance parameter. In such

a setup, one would not only (de-)select complete blocks of regression coefficients but could perform separate selection decisions for sub-groups or single elements of this block. In a recent paper, Bai and Ghosh (2019) consider a high-dimensional Bayesian linear regression model with $p \gg n$ and it would be interesting to work on such extensions for general effect selection.

As in distributional regression the propriety of the posterior is not trivial, however, care has to be taken with respect to the specific prior choices (Ghosh et al., 2018). Alternatively, if interest is rather in smoothing and shrinkage than in explicit effect selection shrinkage priors like the double gamma prior (Bitto and Frühwirth-Schnatter, 2019) or penalised complexity priors (Simpson et al., 2017) might be used. Also, it is conceptually straightforward to include Bayesian quantile or expectile regression models into the NBPSS prior framework and we aim to do so in a future work.

Supplementary Material

Supplementary Material to "Bayesian Effect Selection in Structured Additive Distributional Regression Models" (DOI: 10.1214/20-BA1214SUPP; .pdf). This contains extensive additional material, including tables and figures referred to in the text, organized into Parts A–G.

References

Bai, R. and Ghosh, M. (2019). "On the beta prime prior for Scale Parameters in High-Dimensional Bayesian Regression models." *Statistica Sinica, in press*, doi:10.575/ss.202019.0037. 559, 569

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2015). "BayesX – Software for Bayesian inference in structured additive regression models. Version 3.0.2." Available from http://www.bayesx.org. 561, 562

Barbieri, Maria Maddalena and Berger, James O. (2004). "Optimal predictive model selection." *Annals of Statistics*, 32: 870–897. doi: https://doi.org/10.1214/009053604000000238. 564

Bitto, A. and Frühwirth-Schnatter, S. (2019). "Achieving Shrinkage in a Time-Varying Parameter Model Framework." *Journal of Econometrics*, 210: 75–97. MR3944764. doi: https://doi.org/10.1016/j.jeconom.2018.11.006. 569

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 559

Castillo, I. and van der Vaart A. (2012). "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40: 2069–2101. MR3059077. doi: https://doi.org/10.1214/12-AOS1029. 560

Chung, Y. and Dunson, D. B. (2009). "Nonparametric Bayes Conditional Distribution Modeling With Variable Selection." *Journal of the American Statistical Asso-*

ciation, 104(488): 1646–1660. MR2750582. doi: https://doi.org/10.1198/jasa.2009.tm08302. 547

Clyde, M. and George, E. I. (2004). "Model uncertainty." *Statistical Science*, 19(1): 81–94. MR2082148. doi: https://doi.org/10.1214/088342304000000035. 546

Cottet, R., Kohn, R. J., and Nott, D. J. (2008). "Variable Selection and Model Averaging in Semiparametric Overdispersed Generalized Linear Models." *Journal of the American Statistical Association*, 103: 661–671. MR2524000. doi: https://doi.org/10.1198/016214508000000346. 547

Fahrmeir, L., Kneib, T., and Lang, S. (2004). "Penalized structured additive regression for space-time data: A Bayesian perspective." *Statistica Sinica*, 14: 731–761. MR2087971. 551

Gelman, A., Van Dyk, D., Huang, Z., and Boscardin, W. J. (2008). "Using Redundant Parameterizations to Fit Hierarchical Models." *Journal of Computational and Graphical Statistics*, 17: 95–122. MR2424797. doi: https://doi.org/10.1198/106186008X287337. 547

George, E. and Mc Culloch, R. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88: 881–889. 560

George, E. and McCulloch, R. (1997). "Approaches to Bayesian Variable selection." *Statistica Sinica*, 7: 339–374. 547

Ghosh, J., Li, Y., and Mitra, R. (2018). "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression." *Bayesian Analysis*, 13(3): 359–383. MR3780427. doi: https://doi.org/10.1214/17-BA1051. 569

Ishwaran, H. and Rao, S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *The Annals of Statistics*, 33: 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 547, 553

Kammann, E. E. and Wand, M. P. (2003). "Geoadditive models." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 52: 1–18. MR1963210. doi: https://doi.org/10.1111/1467-9876.00385. 546

Klein, N. (2018). *sdPrior: Scale-Dependent Hyperpriors in Structured Additive Distributional Regression*. R package version 0.6. 555, 561, 562

Klein, N., Carlan, M., Kneib, T., Lang, S., and Wagner, H. (2020). "Supplementary Material of "Bayesian Effect Selection in Structured Additive Distributional Regression Models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1214SUPP. 555

Klein, N. and Kneib, T. (2016). "Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression." *Bayesian Analysis*, 11: 1107–1106. MR3545474. doi: https://doi.org/10.1214/15-BA983. 554, 555, 559

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015a). "Bayesian Structured Additive Distributional Regression for Multivariate Responses." *Journal of the Royal Statistical*

Society. Series C (Applied Statistics), 64: 569–591. MR3367789. doi: https://doi.org/10.1111/rssc.12090. 546, 549

Klein, N., Kneib, T., and Lang, S. (2015b). "Bayesian Generalized Additive Models for Location, Scale and Shape for Zero-Inflated and Overdispersed Count Data." *Journal of the American Statistical Association*, 110: 405–419. MR3338512. doi: https://doi.org/10.1080/01621459.2014.912955. 559, 567, 568

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015c). "Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany." *The Annals of Applied Statistics*, 9: 1024–1052. MR3371346. doi: https://doi.org/10.1214/15-AOAS823. 545, 546, 548, 560

Krivobokova, T., Kneib, T., and Claeskens, G. (2010). "Simultaneous Confidence Bands for Penalized Spline Estimators." *Journal of the American Statistical Association*, 105: 852–863. MR2724866. doi: https://doi.org/10.1198/jasa.2010.tm09165. 561

Kundu, S. and Dunson, D. B. (2014). "Bayes Variable Selection in Semiparametric Linear Models." *Journal of the American Statistical Association*, 109(505): 437–447. MR3180575. doi: https://doi.org/10.1080/01621459.2014.881153. 547

Lang, S. and Brezger, A. (2004). "Bayesian P-splines." *Journal of Computational and Graphical Statistics*, 13: 183–212. MR2044877. doi: https://doi.org/10.1198/1061860043010. 552

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K., and Kneib, T. (2014). "Multilevel Structured Additive Regression." *Statistics and Computing*, 24: 223–238. MR3165550. doi: https://doi.org/10.1007/s11222-012-9366-0. 548, 553

Mitchell, T. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association*, 83: 1023–1032. MR0997578. 547, 556

O'Hara, R. and Sillanpää, M. (2009). "A Review of Bayesian Variable Selection Methods: What, How, and Which." *Bayesian Analysis*, 4: 85–118. MR2486240. doi: https://doi.org/10.1214/09-BA403. 546

Panagiotelis, A. and Smith, M. S. (2008). "Bayesian Identification, Selection and Estimation of Functions in High-Dimensional Additive Models." *Journal of Econometrics*, 143: 291–316. MR2389611. doi: https://doi.org/10.1016/j.jeconom.2007.10.003. 547

Pérez, M.-E., Pericchi, L. R., and Raméz, I. C. (2017). "The scaled beta2 distribution as a robust prior for scales." *Bayesian Analysis*, 12(3): 615–637. MR3655869. doi: https://doi.org/10.1214/16-BA1015. 548, 552, 554

Polson, N. G. and Scott, J. G. (2010). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics*, 9. Oxford Univ. Press. MR3204017. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. 559, 568

Reich, B. J., Storlie, C. B., and Bondell, H. (2009). "Variable Selection in Bayesian Smoothing Spline ANOVA Models: Application to Deterministic Computer Codes." *Technometrics*, 51: 110–120. MR2668168. doi: https://doi.org/10.1198/TECH.2009.0013. 547

Rigby, R. A. and Stasinopoulos, D. M. (2005). "Generalized additive models for location, scale and shape (with discussion)." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54: 507–554. MR2137253. doi: https://doi.org/10.1111/j.1467-9876.2005.00510.x. 545

Rossell, D. and Rubio, F. J. (2018). "Tractable Bayesian variable selection: beyond normality." *Journal of the American Statistical Association*, 113: 1742–1758. MR3902243. doi: https://doi.org/10.1080/01621459.2017.1371025. 547

Rossell, D. and Rubio, F. J. (2019). "Additive Bayesian variable selection under censoring and misspecification." ArXiv:1907.13563. 547, 551, 560

Ročková, V. (2018). "Bayesian estimation of sparse signals with a continuous spike and slab prior." *The Annals of Statistics*, 46: 401–437. MR3766957. doi: https://doi.org/10.1214/17-AOS1554. 559, 560

Ročková, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 113(521): 431–444. MR3803476. doi: https://doi.org/10.1080/01621459.2016.1260469. 559, 560

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. New York/Boca Raton: Chapman & Hall/CRC. MR2130347. doi: https://doi.org/10.1201/9780203492024. 553

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press. MR1998720. doi: https://doi.org/10.1017/CBO9780511755453. 546

Scheipl, F. (2016). *spikeSlabGAM: Bayesian Variable Selection and Model Choice for Generalized Additive Mixed Models*. R package version 1.1.11. 562

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). "Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models." *Journal of the American Statistical Association*, 107: 1518–1532. MR3036413. doi: https://doi.org/10.1080/01621459.2012.737742. 547, 548, 551, 553, 554, 556, 557, 558, 561, 562

Simpson, D., Rue, T. G., H. Martins, Riebler, A., and Sørbye, S. H. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 554, 569

Smith, M. S. and Kohn, R. (1996). "Nonparametric regression using Bayesian variable selection." *Journal of Econometrics*, 75: 317–343. 547

Van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10: 976–1000. MR3486423. doi: https://doi.org/10.1214/16-EJS1130. 559

Wang, L., Yuanyuan Tang, Y., Debajyoti, S., Pati, D., and Stuart Lipsitz, S. (2017). "Bayesian variable selection for skewed heteroscedastic response." ArXiv:1602.09100v2. 547

Wood, S. N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73: 3–36. MR2797734. doi: https://doi.org/10.1111/j.1467-9868.2010.00749.x. 551

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. New York/Boca Raton: Chapman & Hall/CRC, 2nd edition. 546

Xu, X. and Ghosh, M. (2015). "Bayesian Variable Selection and Estimation for Group Lasso." *Bayesian Analysis*, 10(4): 909–936. MR3432244. doi: https://doi.org/10.1214/14-BA929. 547

Yau, P., Kohn, R., and Wood, S. (2003). "Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression." *Journal of Computational and Graphical Statistics*, 12: 23–54. MR1965210. doi: https://doi.org/10.1198/1061860031301. 547

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K.-A. (2014). "Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(4): 595–620. MR3258055. doi: https://doi.org/10.1111/rssc.12053. 547

Zhu, H., Vannunci, M., and Cox, D. D. (2010). "A Bayesian Hierarchical Model for Classification with Selection of Functional Predictors." *Biometrics*, 66: 463–473. MR2758826. doi: https://doi.org/10.1111/j.1541-0420.2009.01283.x. 547

Acknowledgments

We thank a referee, the Associate Editor, the Editor-in-Chief and the Editorial Board for their careful reading of our paper and helpful comments.