Bayesian Analysis (2016) **11**, Number 4, pp. 1071–1106

Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression

Nadja Klein* and Thomas Kneib†

Abstract. The selection of appropriate hyperpriors for variance parameters is an important and sensible topic in all kinds of Bayesian regression models involving the specification of (conditionally) Gaussian prior structures where the variance parameters determine a data-driven, adaptive amount of prior variability or precision. We consider the special case of structured additive distributional regression where Gaussian priors are used to enforce specific properties such as smoothness or shrinkage on various effect types combined in predictors for multiple parameters related to the distribution of the response. Relying on a recently proposed class of penalised complexity priors motivated from a general set of construction principles, we derive a hyperprior structure where prior elicitation is facilitated by assumptions on the scaling of the different effect types. The posterior distribution is assessed with an adaptive Markov chain Monte Carlo scheme and conditions for its propriety are studied theoretically. We investigate the new type of scale-dependent priors in simulations and two challenging applications, in particular in comparison to the standard inverse gamma priors but also alternatives such as half-normal, half-Cauchy and proper uniform priors for standard deviations.

Keywords: Kullback–Leibler divergence, Markov chain Monte Carlo simulations, penalised complexity prior, penalised splines, propriety of the posterior.

1 Introduction

Structured additive regression (Fahrmeir et al., 2004; Kneib et al., 2009) provides an important framework for regression modelling in various areas of applications. It combines the flexibility of generalised additive models (Hastie and Tibshirani, 1990) with the additive inclusion of random effects, spatial components (Kammann and Wand, 2003) and further types of regression effects (Wood, 2006; Fahrmeir et al., 2013) on the conditional expectation. In particular, for some known response function h, the expectation $\mathbb{E}(y|\boldsymbol{x})$ of a response variable y given covariate information $\boldsymbol{x}$ is specified as

$$\mathbb{E}(y|\boldsymbol{x}) = h(\beta_0 + \sum_{j=1}^{J} f_j(\boldsymbol{x}))$$

where β_0 is an intercept and $f_j(\boldsymbol{x})$, $j = 1, \ldots, J$, are different types of functional effects depending on (a subset of) the covariate information $\boldsymbol{x}$. Utilising basis function expansions to represent the functional effects allows us to write the vector of function

*Chair of Statistics, Georg-August-University Göttingen, Germany, nklein@uni-goettingen.de
†Chair of Statistics, Georg-August-University Göttingen, Germany

 DOI: 10.1214/15-BA983

evaluations $\boldsymbol{f}_j = (f_j(\boldsymbol{x}_1), \ldots, f_j(\boldsymbol{x}_n))'$ for n individuals as the matrix vector product $\boldsymbol{f}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$ with suitable design matrices $\boldsymbol{Z}_j$ and vectors of regression coefficients $\boldsymbol{\beta}_j$ to be estimated. In a Bayesian model specification, a zero mean (conditionally) Gaussian prior with precision matrix $\frac{1}{\tau_j^2}\boldsymbol{K}_j$ is typically assumed for $\boldsymbol{\beta}_j$ where $\boldsymbol{K}_j$ is chosen to enforce, for example, smoothness or shrinkage of the coefficient vector $\boldsymbol{\beta}_j$ and the variance parameter τ_j^2 quantifies our prior uncertainty about the properties enforced by $\boldsymbol{K}_j$. Note that in many common model specifications such as Bayesian P-splines (Brezger and Lang, 2006) or Markov random fields (Rue and Held, 2005), the matrix $\boldsymbol{K}_j$ will be rank-deficient such that it cannot be inverted to a proper covariance matrix. Furthermore, note that our formulation uses a parameterisation in terms of the prior variance τ_j^2 while other formulations rely on the prior precision, i.e. the inverse variance, instead.

To complete the Bayesian model specification, suitable hyperpriors have to be augmented to the variance components τ_j^2. While the inverse gamma prior, $\tau_j^2|a,\theta \sim \text{IG}(a,\theta)$, is a natural, conjugate prior comprising flat priors for the variance and the standard deviation as degenerate special cases (see, e.g., Fahrmeir and Kneib, 2009), there has been considerable debate about the suitability of the inverse gamma distribution especially in the context of hierarchical random effects models (see, for example, Gelman, 2005, 2006; Hodges, 2013) or for overfitting models as demonstrated in Frühwirth-Schnatter and Wagner (2010, 2011). As a consequence, several alternatives such as half-normal, half-Cauchy or (proper) uniform priors for the standard deviation have been suggested as default priors in the literature (e.g. Gelman, 2006; Gelman et al., 2008; Polson and Scott, 2012). Another branch of the literature encouraged the choice of 'objective' priors (Bernardo, 1979; Berger, 2006; Ghosh, 2011) or Jeffreys priors (Berger et al., 2009). Unfortunately, prior elicitation of the hyperparameters of these priors, ensuring the propriety of the posterior and justification of the chosen distribution type with respect to axiomatic reasoning are often problematic in these cases.

Without relying on a specific modelling context, Simpson et al. (2014) develop a general approach for determining so-called penalised complexity priors reflecting that frequently hyperpriors are desired for parameters governing the deviation of a flexible model from a restrictive base model. Examples include (i) the normal distribution as a base model which is encompassed in the more flexible t-distribution model for the limiting case of increasing degrees of freedom or (ii) a first order autoregressive process for time series deviating from the base model of independence for increasing absolute value of the autoregressive parameter. In the case of structured additive regression models, an interesting base model is obtained when setting the smoothing variance to zero. For random effects, the base model would then correspond to a pure fixed effects model without random coefficients while for penalised splines with second order random walk prior, the base model would correspond to a parametric, linear effect.

The penalised complexity prior of Simpson et al. (2014) is developed as follows: The deviation between the base model and the more flexible alternative is measured in terms of the Kullback–Leibler divergence. Since usually the base model would be favoured unless there is evidence in the data for the necessity of the more flexible alternative, an exponential prior is assigned to the Kullback–Leibler distance such that the mode of the prior corresponds to the base model. The speed of the exponential decay is determined by a hyperparameter that, in structured additive regression, can

be elicited based on prior assumptions about the scaling of the model components. We therefore refer to the priors as scale-dependent hyperpriors. Based on the prior for the Kullback–Leibler distance, one can then derive the induced prior for the parameter of interest (i.e. the smoothing variance in case of structured additive regression). The main advantages of the approach by Simpson et al. (2014) are (i) the derivation based on a simple set of axiomatic assumptions (preference of the base model, exponential decay in the distance between base model and alternative) and (ii) the assistance they offer for prior elicitation by requesting statements on the expected scaling from the data analyst.

The idea of divergence-based prior constructions can be traced back at least to Jeffreys (1961) where it led to the derivation of the famous class of Jeffreys priors. García-Donato and Sun (2007) and Bayarri and García-Donato (2008) considered divergence-based priors in the context of Bayesian hypothesis testing. However, they utilise the Kullback–Leibler divergence for the complete models to be tested while the penalised complexity prior only takes the divergence between prior structures into account. This allows for much more versatile derivations since the priors can be used as building blocks in a variety of models but also comes at the price of more restrictive assumptions on prior independence. We will return to this issue later in Section 3.1.

In this paper, we utilise the approach of Simpson et al. (2014) to develop scale-dependent priors for the variance parameters in structured additive distributional regression and to compare it to other types of priors with parameters chosen according to the same scaling criterion. More specifically, the main contributions of our paper are as follows:

- While Simpson et al. (2014) already consider some special cases of effects comprised in structured additive regression (random walk priors, random effects), they are restricted to indicator basis functions that induce a design matrix $\boldsymbol{Z}_j$ of the zero / one incidence type. We construct scale-dependent hyperpriors for the general case of arbitrary basis functions which makes the determination of the scaling factor more demanding.

- We develop Markov chain Monte Carlo simulation inference with scale-dependent priors instead of the integrated nested Laplace approximation framework (INLA, Rue et al., 2009; Lindgren et al., 2011) of Simpson et al. (2014). This involves the derivation of a suitable proposal density for the scale-dependent prior based on a quadratic approximation to the log-full conditional of the log-variance parameters. This is particularly relevant since we include scale-dependent hyperpriors not only in mean regression models for Gaussian responses or responses from the exponential family, but rather consider the general framework of distributional regression (Klein, Kneib and Lang, 2015; Klein, Kneib, Lang and Sohn, 2015) where further moments or general shape parameters of the conditional response distribution can be related to a predictor (similar as in generalised additive models for location, scale and shape, Rigby and Stasinopoulos, 2005) which is currently (to the best of our knowledge) not possible with INLA.

- We establish theoretical results by providing sufficient (and sometimes necessary) conditions to arrive at a proper posterior distribution. The question of propriety

arises naturally due to the partially improper priors for the vectors of regression coefficients for several effect types in structured additive regression.

- We study the potential of scale-dependent hyperpriors to enhance numerical stability in situations with a flat likelihood and give empirical evidence in favour of the proposed prior as compared to other prior structures in further simulation studies and two applications.

- In addition to the common inverse gamma priors that lead to Gibbs sampling updates, we consider half-normal, half-Cauchy and proper uniform priors for standard deviations that have been advocated as hyperpriors for variance parameters in Bayesian mixed effects models (Gelman, 2005, 2006; Hodges, 2013). To derive suitable proposal densities, we again consider quadratic approximations of the log-full-conditional for the log-variances. The parameters of the hyperpriors are chosen according to a similar scaling criterion for the effects as for the scale-dependent priors.

Accordingly, the rest of the paper is structured as follows. In Section 2, we first introduce the generic model formulation in structured additive distributional regression in order to then construct the scale-dependent hyperpriors for the variance components. Furthermore, we provide a practicable strategy to optimise the scale parameter of the priors, illustrate it along two exemplary effect types and make further alternative prior structures applicable (see Table 1 for a summary of the supported prior specifications). The main theoretical results on the propriety of the posterior are treated in Section 3 while required further theoretical results are in Supplement A (Klein and Kneib, 2015). In the subsequent Section 4, we describe how inference is performed with an adaptive Metropolis–Hastings algorithm for all unknown parameters and describe suitable adaptations for alternative hyperprior structures. Basic results from simulations capturing various scenarios are briefly summarised in Section 5 while all simulations are documented in more detail in Supplement B and G. The good performance of scale-dependent priors is illustrated along applications on patent citations and geoadditive regressions in the analysis of childhood undernutrition in Zambia in Section 6 before we end with a discussion in Section 7. Further theoretical details and derivations can be found in Supplement C to F.

2 Scale-Dependent Hyperpriors

2.1 Distributional Regression

Observation Model We consider the construction of scale-dependent hyperpriors in the general class of Bayesian structured additive distributional regression, recently developed in Klein, Kneib, Lang and Sohn (2015) for univariate responses and extended in Klein, Kneib, Klasen and Lang (2015) to multivariate responses. In these models, it is assumed that the (not necessarily scalar) response variable y given covariates $\boldsymbol{x}$ has a parametric distribution with density

$$p(y|\vartheta_1, \ldots, \vartheta_K) \tag{M1}$$

where $\{\vartheta_k, k = 1, \ldots, K\}$ is a collection of distributional parameters each taking values in a subset of $\mathbb{R}$. Compared to mean regression where $p(\cdot)$ usually belongs to the exponential family and where $K-1$ parameters are treated as fixed or nuisance parameters, in distributional regression each of the distribution parameters is linked to a structured additive predictor η_k via a suitable one-to-one transformation h_k, i.e. $h_k(\eta_k) = \vartheta_k$. The predictors are then composed additively as

$$\eta_k = \beta_{0,k} + \sum_{j=1}^{J_k} f_{j,k}(\boldsymbol{x}) \tag{M2}$$

where, in turn, each function $f_{j,k}(\boldsymbol{x})$ depending on (different subsets of) $\boldsymbol{x}$ is represented by a linear combination of basis functions. Hence, after dropping the dependence on the distributional parameter (index k) and the order of the functions (index j), a typical function $f(\boldsymbol{x})$ is then specified as

$$f(\boldsymbol{x}) = \sum_{d=1}^{D} \beta_d B_d(\boldsymbol{x}) \tag{1}$$

where $B_d(\boldsymbol{x})$, $d = 1, \ldots, D$, is a set of appropriate basis functions while $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)'$ is the vector of corresponding basis coefficients. To ensure identifiability of the model, specific constraints $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ with appropriate matrices $\boldsymbol{A}$ representing, for example, centring of the functional effects are added such that β_0 corresponds to the overall level of the predictor. Specific examples include penalised splines, where the basis functions correspond to B-splines whereas the amplitudes of these basis functions are estimated with the vector of coefficients or Markov random fields where the coefficients represent spatial effects for a prespecified set of regions and the basis functions are indicator functions mapping the individual observations to these regions (see Fahrmeir et al., 2013, for details and further examples).

Since in many cases the vector of basis coefficients $\boldsymbol{\beta}$ will be of rather high dimension, it is important to enforce specific properties of the estimates such as smoothness or shrinkage to regularise estimation and therefore to reduce variability of estimates. In a likelihood-based framework, this can be achieved by adding quadratic penalties $\lambda\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta}$ to the likelihood, where the smoothing parameter λ determines the impact of the penalty on the estimation result. Data-driven estimates for all smoothing parameters contained in a structured additive regression model can then, for example, be determined via (approximate) restricted maximum likelihood estimation (Fahrmeir et al., 2004) or based on generalised cross-validation (Wood, 2006).

Prior Assumptions for Regression Coefficients In a Bayesian treatment of structured additive regression models, the stochastic analogue to quadratic penalties are (partially improper) multivariate Gaussian priors

$$p(\boldsymbol{\beta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{\frac{\mathrm{rk}(\boldsymbol{K})}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta}\right) \tag{2}$$

with fixed positive (semi-)definite precision matrix $\boldsymbol{K}$, variance parameter τ^2 and $\mathrm{rk}(\cdot)$ denoting the rank of a matrix. For the overall constants β_0, we typically assume flat

priors

$$p(\beta_0) \propto const \tag{3}$$

similar as for linear effects with design matrix containing covariates and precision matrix $\boldsymbol{K} = \boldsymbol{0}$.

Mixed Model Representation Due to the potential rank deficiency, the null space of the precision matrix $\boldsymbol{K}$ in (2) will in some cases be non-trivial. For a proper construction of the hyperprior in the following, we introduce a mixed model type representation (Wand, 2000; Ruppert et al., 2003; Fahrmeir et al., 2004) for the vector of function evaluations $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))' = \boldsymbol{Z}\boldsymbol{\beta}$ of n individuals as

$$\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{Z}(\tilde{\boldsymbol{U}}\boldsymbol{\beta}_{unpen} + \tilde{\boldsymbol{V}}\boldsymbol{\beta}_{pen}) = \boldsymbol{U}\boldsymbol{\beta}_{unpen} + \boldsymbol{V}\boldsymbol{\beta}_{pen} \tag{4}$$

such that $\tilde{\boldsymbol{V}}'\boldsymbol{K}\tilde{\boldsymbol{V}} = \boldsymbol{I}$ and $\tilde{\boldsymbol{U}}'\boldsymbol{K}\tilde{\boldsymbol{U}} = \boldsymbol{0}$ hold. The columns of $\tilde{\boldsymbol{U}}$ are then a basis of the nullspace of $\boldsymbol{K}$ and $\tilde{\boldsymbol{V}}$ can be obtained from the spectral decomposition of the prior precision matrix $\boldsymbol{K} = \boldsymbol{\Gamma}\boldsymbol{\Omega}_+\boldsymbol{\Gamma}'$ (with $\boldsymbol{\Omega}_+$ the diagonal matrix of positive eigenvalues and $\boldsymbol{\Gamma}$ the corresponding orthonormal matrix of eigenvectors) as $\tilde{\boldsymbol{V}} = \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1}$ where $\boldsymbol{L} = \boldsymbol{\Gamma}\boldsymbol{\Omega}_+^{1/2}$. Furthermore, the spectral decomposition of $\boldsymbol{K}$ delivers the generalised inverse $\boldsymbol{K}^- = \boldsymbol{\Gamma}\boldsymbol{\Omega}_+^-\boldsymbol{\Gamma}'$ with diagonal matrix

$$\boldsymbol{\Omega}_+^-[d,d] = \begin{cases} 1/\boldsymbol{\Omega}_+[d,d] & \text{if } \boldsymbol{\Omega}_+[d,d] > 0, \\ 0 & \text{otherwise} \end{cases}$$

and the generalised determinant of $\boldsymbol{K}^-$, $|\boldsymbol{K}^-| = \prod_{\boldsymbol{\Omega}_+[d,d]>0} \boldsymbol{\Omega}_+^-[d,d]$.

As a consequence, the dimension of vector $\boldsymbol{\beta}_{pen}$ equals the rank of the precision matrix $\boldsymbol{K}$, i.e. $\dim(\boldsymbol{\beta}_{pen}) = \text{rk}(\boldsymbol{K}) = \kappa$ and $\boldsymbol{\beta}_{pen}$ follows a proper i.i.d. normal prior, $\boldsymbol{\beta}_{pen}|\tau^2 \sim \text{N}(0, \tau^2\boldsymbol{I})$, while $\boldsymbol{\beta}_{unpen}$ is the unpenalised part of dimension $D - \kappa$ with zero precision matrix and flat prior. In this way, $1/\tau^2$ can be interpreted as the precision of the deviation from the null space which facilitates the construction of the scale-dependent hyperprior based on the base model defined by this null space. For a spline component with second order random walk prior, for example, the null space corresponds to a linear effect in the covariate and τ^2 can therefore be interpreted as a measure for the deviation from this simpler base model, as explained in more detail in the following.

2.2 Derivation of the New Hyperprior Structure

We will now discuss the construction of a new type of prior structure for the variance parameters τ^2 by applying the general definition of penalised complexity priors developed in Simpson et al. (2014). The basic reasoning of this derivation relies on the following principles discussed in Simpson et al. (2014):

Principle 1: *Occam's razor.* The hyperprior should invoke the principle of parsimony in the sense that a suitable, simple base model for each effect is preferred unless the data provide convincing evidence for a more complex modelling alternative.

Principle 2: *Measure of complexity.* The increased complexity between two models is measured by the unidirectional measure

$$d(p||p_b) = \sqrt{2\,\mathrm{KLD}(p||p_b)}$$

where $\mathrm{KLD}(p||p_b)$ denotes the Kullback–Leibler divergence between the base model represented by density p_b and the alternative represented by density p, i.e.

$$\mathrm{KLD}(p||p_b) = \int p(u) \log\left(\frac{p(u)}{p_b(u)}\right) \mathrm{d}u. \tag{5}$$

Principle 3: *Constant rate penalisation.* This assumption implies an exponential prior $p_d(d) = \lambda \exp(-\lambda d)$ on the distance scale d such that there is a constant rate of decay in the distance prior from the base model to stronger deviations from this base model as quantified by the KLD.

Theorem 1. *Let* $\frac{1}{\tau^2}\boldsymbol{K}$ *be the precision matrix of the flexible model for a vector of regression coefficients* $\boldsymbol{\beta}$ *and* $\frac{1}{\tau_b^2}\boldsymbol{K}$ *the precision matrix of the base model where* $\tau_b^2 \to 0$. *Furthermore, let* $p(\tau^2)$ *be the prior for* τ^2 *depending on a hyperparameter* θ. *If the prior* $p(\cdot)$ *is constructed according to Principles 1 to 3 discussed above, we obtain a Weibull prior with shape parameter* $a = 1/2$ *and scale parameter* θ *(determined by the rate of decay* λ *specified in Principle 3), i.e.*

$$p(\tau^2) = \frac{1}{2\theta}\left(\frac{\tau^2}{\theta}\right)^{-1/2} \exp\left(-\left(\frac{\tau^2}{\theta}\right)^{1/2}\right). \tag{6}$$

This result is a consequence of the proof in Appendix A2 of Simpson et al. (2014) applying the change of variable theorem. In Supplement A.1, we provide a detailed proof that derives the prior based on Principles 1 to 3 to allow for a better understanding of how the principles lead to the implied prior structure. This proof is an extended and adapted version of the one in Appendix A2 of Simpson et al. (2014).

Remark 1.

(i) As indicated by Simpson et al. (2014), the invariance property of the prior is obtained 'for free'. For example, a type-2 Gumbel distribution is obtained as the prior for $\xi = 1/\tau^2$.

(ii) As already noted by Simpson et al. (2014), the exponential decay assumption can of course be replaced by alternative priors for the distance if desired. We stick to the simple exponential prior since it is a convenient default in cases where no additional prior knowledge on the distance is available.

(iii) Principle 4.*(User-defined scaling)* of Simpson et al. (2014) controls the decay-rate $r = \exp(-\lambda)$ by imposing the condition

$$\mathbb{P}(q(\tau^2) \leq c) = 1 - \alpha \tag{7}$$

for an interpretable transformation $q(\cdot)$ of τ^2 and some user-defined values $c > 0$ and $\alpha \in (0, 1)$. The probability in (7) depends on the intensity λ via the density

of $q(\tau^2)$ such that solving the expression with respect to λ yields the exact prior specification for τ^2. We discuss the choice of $q(\cdot)$, c and α in the subsequent section.

2.3 Choosing the Scale Parameter – User-Defined Scaling

Compared to the models with direct linkage between a parameter of interest and its marginal precisions considered in Simpson et al. (2014) allowing for simple forms of the transformation $q(\cdot)$ in (7) such as $q(\tau) = \tau$, we are interested in relating the scale parameter θ to the functions f rather than directly to the variances τ^2. This means that the user has some knowledge about the scale of f providing the specification of a certain interval the function f falls into with a high marginal probability, i.e.

$$\mathbb{P}(|f(\boldsymbol{x})| \leq c \, \forall \boldsymbol{x} \in \mathcal{D}) \geq 1 - \alpha \tag{8}$$

where $\alpha \in (0, 1)$ and $c > 0$ are chosen in advance and $\mathcal{D}$ denotes the domain of $\boldsymbol{x}$. The absolute value can be taken without loss of generality due to the centring constraint in the additive predictor (M2) for each function to ensure identifiability. For simplification purposes, we reduce the simultaneous statement above to a finite-dimensional problem by using a subset of points $\mathcal{X}_P = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P\}$ from $\mathcal{D}$ together with the Bonferroni inequality to arrive at

$$\mathbb{P}(|f(\boldsymbol{x}_p)| \leq c \, \forall \boldsymbol{x} \in \mathcal{X}_P) \geq 1 - \sum_{p=1}^{P} \mathbb{P}\left(|f(\boldsymbol{x}_p)| \geq c\right). \tag{9}$$

The marginal density of $f(\boldsymbol{x}_p) = (B_1(\boldsymbol{x}_p), \ldots, B_D(\boldsymbol{x}_p))\boldsymbol{\beta} = \boldsymbol{z}_p'\boldsymbol{\beta}$ can be obtained by integrating τ^2 out, i.e. computing the integral

$$p(\boldsymbol{z}_p'\boldsymbol{\beta}) = \int_0^\infty p(\boldsymbol{z}_p'\boldsymbol{\beta}, \tau^2) \mathrm{d}\tau^2 = \int_0^\infty p(\boldsymbol{z}_p'\boldsymbol{\beta}|\tau^2) p(\tau^2) \mathrm{d}\tau^2$$

where $\boldsymbol{z}_p'\boldsymbol{\beta}|\tau^2 \sim \mathrm{N}(0, \tau^2 \boldsymbol{z}_p'\boldsymbol{K}^- \boldsymbol{z}_p)$. Hence, θ can be chosen such that

$$\sum_{p=1}^{P} \left(1 - \int_{-c}^{c} \int_0^\infty p_{\boldsymbol{z}_p'\boldsymbol{\beta}}(u|\tau^2) p(\tau^2) \mathrm{d}\tau^2 \mathrm{d}u\right) = \alpha \tag{10}$$

is fulfilled and (10) can be solved numerically. An implementation for given precision matrix $\boldsymbol{K}$, design matrix $\boldsymbol{Z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_P')$, probability level α and the threshold c is provided in the R-package `sdPrior` (Klein, 2015), available from CRAN (`https://cran.r-project.org/web/packages/sdPrior/`). We will illustrate the application of user-defined scaling in practice along two important predictor components in the following.

Bayesian P-Splines Bayesian P-splines (Lang and Brezger, 2004) are a Bayesian version of penalised regression with B-splines (P-splines, Eilers and Marx, 1996), replacing the difference penalties with Gaussian random walk priors. In this case, the prior precision matrix is given by $\boldsymbol{K} = \boldsymbol{D}'\boldsymbol{D}$ where $\boldsymbol{D}$ denotes a difference matrix of appropriate

order. While for fixed order of the penalty $\boldsymbol{K}$ always has the same structure, the design matrix $\boldsymbol{Z}$ varies depending on the specification of the knots of the spline basis and the distribution of the observed covariate values. To empirically evaluate how much the optimal value θ obtained from (10) depends on these factors and to investigate the loss of efficiency through the inequality when switching from a simultaneous statement to a pointwise approximation, we conducted several simulation experiments whose results are summarised below (a more detailed documentation is provided in Supplement B). Let therefore $\mathcal{X}_P = \{x_1, \ldots, x_P\}$ be the set of realisations of x at which the probability statement is evaluated, $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.5, 0.7\}$ the probability statements under consideration and $c = 3$ the threshold of interest. The latter is motivated from the fact that with the most common link functions such as the log link, the probit link or the logistic link, there is no more variability in the desired parameter if the predictor exceeds the range from -3 to 3. Of course, in practice the threshold c as well as the other settings can (and should) be chosen according to the prior knowledge of the analyst to induce a user-defined amount of scaling. As mentioned above, this can easily be done using the functionality provided in the R-package `sdPrior`. Finally, let $r = 1, \ldots, R = 1000$ be the number of simulation replications. Then, the results of our empirical evidence can be summarised as follows:

1. *Bonferroni inequality.* Let here $\mathcal{X}_P$ be deterministic equidistant grids within the interval $[-3, 3]$ of sizes $P \in \{1, 2, 5, 10, 15, 30, 40, 50, 100\}$. As expected, the desired level is maintained pointwise for all P while for the simultaneous statement, independently from α, larger P induce more conservative statements, compare Figure B1.

   Note that in general the usage of the Bonferroni inequality only makes the statement in (9) more conservative which is not of very high relevance for our purposes. Furthermore, in simulations the results for θ (and the resulting functional estimates for f) proved to be very robust across different values for α so that we expect that the estimation results for the functional effects f will not change (or only marginally) in practice. As a consequence, we did not invest additional efforts in reducing the loss of efficiency induced by the Bonferroni inequality.

2. *Optimal scale parameter based on observed covariate values.* The distribution of x varies between a uniform, normal and gamma distribution with parameters chosen such that they have equal quantile ranges $dx \in \{20, 10, 6, 4, 2, 1\}$. The level α is fixed to 0.01. As expected, the larger $P \in \{1, 2, 3, 4, 5, 10, 15, 20\}$, the smaller θ. However, all absolute values for $P > 1$ lie in the interval $[0.002, 0.005]$ such that the differences are relatively small in absolute terms, see Figure B6. For $P = 1$, we obtain $\theta \approx 0.008$. The interval range dx does not seem to influence θ even though some variation within the interval range is visible.

3. *Optimal scale parameter based on knots of the design matrix.* Now, $x_1 \leq \cdots \leq x_P$ is the sequence of $P = 22$ knots corresponding to the design matrices $\boldsymbol{Z}$ of cubic B-splines and second order random walk prior obtained for univariate, normal or gamma distributed covariates x in samples of sizes $n = 50, 10, 1000$, dx as above and $\alpha = 0.01$ as before. As expected, for all three distributions, ranges dx and sample sizes n, the optimal θ is approximately the same (Figure B7).

Based on the results so far, we decided for a pointwise selection of θ ($P = 1$) depending on the chosen values of α and c. Exemplary values for $c = 3$ are listed in Table B1 of the supplement.

Markov Random Fields Suppose that spatial information in terms of an index $s \in \{1, \ldots, S\}$ representing the location in one of S geographical regions is given and that a spatial effect should be included in a regression specification. In a Markov random field specification, each region would be associated with its own regression coefficient such that the $n \times S$ design matrix $\boldsymbol{Z}$ is of zero/one incidence type, linking regions and individual observations. To obtain spatial smoothness, the $S \times S$ precision matrix $\boldsymbol{K}$ is constructed based on the set of indices δ_s consisting of the neighbouring regions of s obtained from the symmetric neighbourhood relation $s \sim t \Leftrightarrow t \in \delta_s$. Then, the off-diagonal entries of $\boldsymbol{K}$ are given as $\boldsymbol{K}[s,t] = \boldsymbol{K}[t,s] = -\mathbb{1}\{s \sim t\},\ s \neq t$ while the diagonal elements are given by $\boldsymbol{K}[s,s] = |\delta_s|$ such that $\text{rk}(\boldsymbol{K}) = S - 1$. Hence, unlike for penalised splines where the precision matrix has a fixed structure only depending on the difference order, the structure of the precision matrix varies with the neighbourhood structure and in particular the connectivity of the geographical units. In such a case, our `sdPrior` package can be used to compute θ for a given map and neighbourhood structure (and in fact any design matrix $\boldsymbol{Z}$ and precision matrix $\boldsymbol{K}$).

Note that in general, the same hyperprior for the smoothing variance can induce very different amounts of smoothing for different effect types as has been shown by Sørbye and Rue (2014) for Gaussian Markov random field priors. This problem is, however, avoided in our case by explicitly including the structure of $\boldsymbol{K}$ in the determination of θ (compare (8) to (10)).

Since conditionally β_s given all other regions collected in $\boldsymbol{\beta}_{-s}$ follows a normal distribution with expectation equal to the mean of the neighbouring regions and variance $\tau^2 |\delta_s|^{-1}$, the base model in case of Markov random fields corresponds to the deterministic situation of a zero spatial effect.

2.4 Scaling Alternative Prior Structures

While Principles 1 to 3 for penalised complexity priors give rise to the Weibull prior for the smoothing variance, Principle 4 can also be applied to choose hyperparameters of other prior structures, such as inverse gamma priors or half-normal, half-Cauchy and proper uniform priors for standard deviations (Gelman, 2005, 2006; Hodges, 2013).

For example, in case of an inverse gamma prior with $a = 1$ and θ small (which defines a 'flat' prior for the inverse smoothing variance, i.e. the precision), $\boldsymbol{z}_p'\boldsymbol{\beta}$ marginally follows a t-distribution, $\boldsymbol{z}_p'\boldsymbol{\beta} \sim \text{t}(2a, 0, \theta \boldsymbol{z}_p' \boldsymbol{K}^- \boldsymbol{z}_p / a)$. This considerably simplifies optimising (10) with respect to θ since it avoids the necessity of numerical integration to obtain the marginal probability statements.

Unfortunately, however, Principle 4 is not generally applicable. For example, in case of the inverse gamma prior $\text{IG}(\theta, \theta)$ with $\theta = \epsilon$ and ϵ small, there are certain ranges of the threshold c and the probability level α for which there is no solution of (10) with respect to θ. The inherent reason is that ϵ impacts the shape of the inverse gamma

density in a complex way such that reducing ϵ does not allow to concentrate more probability mass of the marginal distribution close to zero.

In contrast, both the half-normal and the half-Cauchy prior can be transferred to the scaling principle. Since both priors are usually assumed for the standard deviation τ, we first derive the implied priors for τ^2 and then study the resulting marginal prior for $\boldsymbol{z}_p'\boldsymbol{\beta}$. For the half-normal prior $\tau \sim \text{HN}(0, \theta^2)$, i.e. the normal distribution with expectation zero and variance θ^2 truncated to the positive half-axis, we obtain a gamma distribution with shape parameter 1/2 and scale parameter $(2\theta^2)^{-1}$ for τ^2 (see Supplement D.1 for details) such that the marginal density of $\boldsymbol{z}_p'\boldsymbol{\beta}$ is given by

$$p(\boldsymbol{z}_p'\boldsymbol{\beta}) = \frac{1}{\pi\theta\sqrt{\boldsymbol{z}_p'\boldsymbol{K}^-\boldsymbol{z}_p}} K_0(\boldsymbol{z}_p'\boldsymbol{\beta}/\sqrt{\theta^2\boldsymbol{z}_p'\boldsymbol{K}^-\boldsymbol{z}_p})$$

with $K_0(x)$ denoting the modified Bessel function of second kind and order 0. The half-Cauchy prior $\tau \sim \text{HC}(0, \theta^2)$ with location parameter 0, scale parameter θ and density proportional to $(1+(\tau/\theta)^2)^{-1}$ implies a generalised beta prime distribution with density

$$p(\tau^2) = \frac{1}{\pi\theta^2}\left(1 + \frac{\tau^2}{\theta^2}\right)^{-1}\left(\frac{\tau^2}{\theta^2}\right)^{-1/2}$$

for τ^2 where the three shape parameters are given by 1/2, 1/2, and 1 while the scale parameter corresponds to θ^2 (see Supplement D.2 for details). In this case, the marginal distribution of $\boldsymbol{z}_p'\boldsymbol{\beta}$ can only be approximated numerically. Due to the heavy tails of the Cauchy prior, the resulting numerical optimisation of (10) can be unstable especially for a large number of design points P or small values of the probability level α. We discuss these numerical challenges in more detail in Supplement B where we conduct the Monte Carlo experiment on the marginal probabilities and the impact of the Bonferroni inequality, with settings described previously in the paragraph on Bayesian P-splines.

As a final alternative prior, we consider proper uniform priors $\tau \sim \text{U}(0, \theta)$ where θ should be chosen according to the scaling criterion. Unfortunately, this prior turns out to be difficult to deal with due to the convexity of the log-density. In particular, the general principle for constructing adaptive proposal densities discussed in Section 4.1.2 relies on a quadratic approximation of the log-full conditional which does not fit with the convex form of the log-density. We therefore constructed an approximation based on the sigmoid function

$$p_{\text{approx}}(\tau) = \frac{\theta^{-1}}{1 + \log(1 + \exp(-\tilde{c}))/\tilde{c}}\left(1 - \frac{\exp(\tau\tilde{c}/\theta - \tilde{c})}{1 + \exp(\tau\tilde{c}/\theta - \tilde{c})}\right)$$

where $\tilde{c} := \theta/s$ and s controls the precision of the approximation, see Supplement C for details on the approximation and the choice of $\tilde{c}$. Note that this approximation also avoids possible difficulties due to the inherent non-differentiability of the uniform prior with respect to τ in θ. Density and cumulative distribution function for the distribution of τ^2 under the approximate uniform distribution for τ are available in closed form and the corresponding implementations (including random number generation) have been added to the package `sdPrior`.

For all four priors (inverse gamma with $a = 1$ for τ^2, half-normal, half-Cauchy and approximate uniform τ), the scale-dependent determination of θ is implemented in the R-package `sdPrior`. In addition, it is possible to derive the implied prior densities on the distance scale to facilitate the comparison with the construction principle for scale-dependent priors (see Supplement E). Figure 1 shows the prior densities for the variance (see Table 1 for details on the parameterisations) as well as the distance scale for five different prior specifications with parameters chosen according to the scaling criterion where $\alpha = 0.01$, $c = 3$ and $P = 1$ for a spline component with B-spline basis of degree 3, 20 inner knots and second order difference matrix. There are some rather striking features of these prior densities:

- Most importantly, as already noted by Simpson et al. (2014), the inverse gamma prior puts zero density on the base model as can be seen in the right panel of Figure 1. In fact, even a large amount of models close to the base model are assigned a prior probability that is approximately zero. As a consequence, the inverse gamma prior inherently favours the more complex, nonlinear model even in case of complete absence of evidence for nonlinearity in the data. This is in fact also true for the inverse gamma distribution with both parameters equal and small (not included in Figure 1 due to its rather different shape), although the prior starts to deviate from zero earlier.

- All other priors follow the principle of penalised complexity and prior preference of the base model, such that the prior density of the base model on the distance scale is different from zero. What differs is the speed of the decay of the prior on the distance scale.

- All priors except the inverse gamma are convex on the variance scale. In contrast, only the scale-dependent and the proper uniform prior are convex on the distance scale. This corresponds to stronger discrimination between base model and alternatives close to the base model for the scale-dependent and the proper uniform prior while half-normal and half-Cauchy put almost the same prior density on the base model and small deviations (see the inlay in the right panel of Figure 1).

- On the distance scale, the tail behaviour of the half-normal and the uniform prior are pretty similar with a fast decay to zero. In contrast, both half-Cauchy and scale-dependent prior have heavier tails.

Marginal densities of $\boldsymbol{z}_p'\boldsymbol{\beta}$ in Figure B8 (Supplement B) indicate that all but the inverse gamma priors induce a similar structure (but different absolute values of the densities) with a distinct peak at zero that increases for smaller values of α. Resulting from the marginal t-distribution, inverse gamma priors imply that the maximum of the densities is equal for all α. Its computation is available via the R-package `sdPrior`.

For the four alternative prior specifications (inverse gamma, half-Cauchy, half-normal and proper uniform), we repeated the Monte Carlo experiment (1) with optimised scale parameter for each of the priors. As indicated before, some problems can occur with

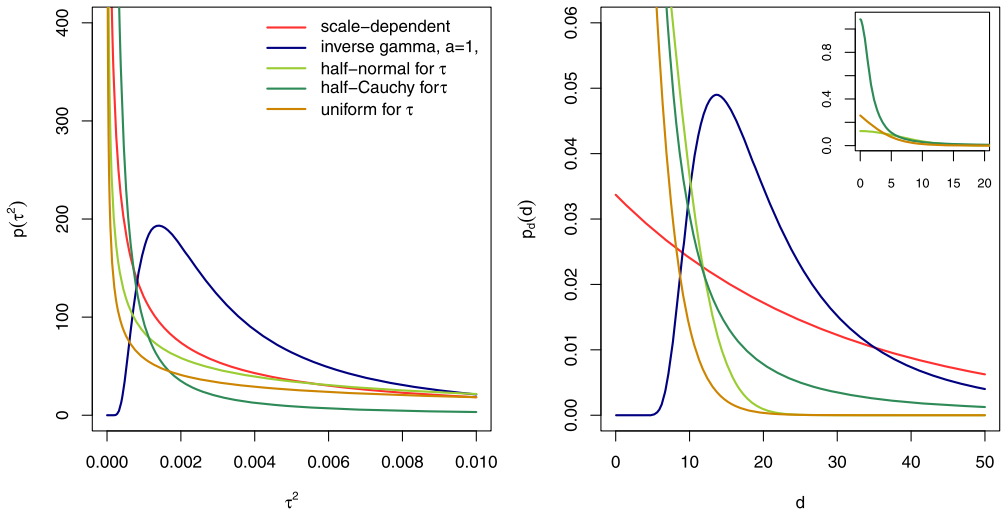

Figure 1: Illustration of densities for different hyperpriors for τ^2. The left panel shows the priors $p(\tau^2)$, the right panel the resulting priors on the distance scale. The scale parameter θ of the priors is the value resulting from a pointwise optimisation criterion with $\alpha = 0.01$, $c = 3$, $P = 1$ for a spline component with B-spline basis of degree 3, 20 inner knots and second order difference matrix.

half-Cauchy priors while no inconsistencies for the other priors can be identified, compare Figures B1 to B5 (Supplement B). Note that based on our pointwise selection, i.e. $P = 1$, the desired coverage is maintained for all probability levels α (and thus no biased estimation in that respect is expected), as can be seen in Figure B4.

2.5 Priors for the Error Variance in Gaussian Mean Regression

Distributional regression comprises several simpler models such as structured additive mean regression as special cases. In particular, the Gaussian mean regression model

$$y = \eta + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \tau_\varepsilon^2) \tag{11}$$

is an important simplification which, however, requires the additional specification of a prior for the error variance τ_ε^2. While it may be tempting to use the scale-dependent hyperprior also in this case, it is important to note that the base model would correspond to a model with zero variance, i.e. a deterministic model where all residuals are equal to zero. This is in contrast to the application of scale-dependent priors for regression effects in structured additive regression where the base model is also deterministic but this simply implies that complex effects reduce to simpler ones. In case of the error variance, an interpolating model with zero variances would certainly not be accepted as a useful base model (and could usually not be estimated anyway if the number of parameters is smaller than the sample size). As a consequence, we will stick to the usual $\mathrm{IG}(a_\varepsilon, \theta_\varepsilon)$ prior as default choice in Gaussian mean regression in the following.

Name	Density	Information
SD(α)	$p(\tau^2) \propto \left(\tau^2/\theta\right)^{-1/2} \exp(-(\tau^2/\theta)^{1/2})$	scale-dependent prior for τ^2
HN(α)	$p(\tau^2) \propto (\tau^2)^{1/2-1} \exp\left(-\tau^2/(2\theta^2)\right)$	gamma prior for τ^2 / half-normal prior for τ
HC(α)	$p(\tau^2) \propto \left(1+\tau^2/\theta^2\right)^{-1} \left(\tau^2/\theta^2\right)^{-1/2}$	generalised beta prime prior for τ^2/half-Cauchy prior for τ
U(α)	$p(\tau^2) \propto (\tau^2)^{-1/2} \left(1 - \frac{\exp\left((\tau^2)^{1/2}\tilde{c}/\theta - \tilde{c}\right)}{1+\exp\left((\tau^2)^{1/2}\tilde{c}/\theta-\tilde{c}\right)}\right)$	approximate uniform prior for τ^2/proper uniform prior for τ
IG(α)	$p(\tau^2) \propto (\tau^2)^{-2} \exp\left(-\theta/\tau^2\right)$	flat prior for $1/\tau^2$ for $\theta \to 0$
IG(ϵ, ϵ)	$p(\tau^2) \propto (\tau^2)^{-\epsilon-1} \exp\left(-\epsilon/\tau^2\right)$	'Jeffreys prior'/flat prior on log-scale for $\epsilon \to 0$
IG($-1, 0$)	$p(\tau^2) \propto const$	flat prior for τ^2
IG($-1/2, 0$)	$p(\tau^2) \propto 1/\sqrt{\tau^2}$	flat prior for τ

Table 1: Overview of available hyperpriors for τ^2. For further details on the densities, see Supplement C to E.

Still, when investigating propriety of the joint posterior distribution in Section 3, we will see that models with a scale-dependent prior on τ^2_ε also arise naturally when applying a certain reparameterisation. However, these models are then a technical device to prove propriety rather then being a sensible model specification to use in practice.

3 Propriety of the Posterior Distribution

Since partially improper priors (2) for the vectors of regression coefficients are employed in structured additive distributional regression, a natural and important question that arises is whether the joint posterior distribution is proper. Recently, Klein, Kneib and Lang (2015) found sufficient conditions for the propriety in the general framework of structured additive distributional regression when the usual inverse gamma hyperpriors for the smoothing variances τ^2 are used. The results are based on the work of Sun et al. (2001) who derived several upper and lower bounds for the required integrals. Unfortunately, these bounds are very specific to the inverse gamma case and thus cannot be directly transferred to our model class with scale-dependent priors. We will therefore establish adapted bounds in Lemma 2 that will allow us to generalise the results of Klein, Kneib and Lang (2015) to distributional regression with scale-dependent hyperpriors. We will treat the Gaussian mean regression case first and will then consider the general framework of distributional regression. Note that unlike in the case of inverse gamma priors, the distributional regression case cannot be immediately traced back to the Gaussian case by applying a mixed model representation. The reason is that the latter results in a working Gaussian model with scale-dependent prior also for the error variance. As a consequence, additional considerations are required. Further theoretical details and proofs will be given in Supplement A while we restrict ourselves to the main results in the rest of this section.

3.1 Conditional Independence Assumptions and Posterior Distribution

Before the posterior distribution can be derived using Bayes' theorem with the observation model from (M1), (M2) and the prior specifications in (2), (3) and (6), we complete the Bayesian model formulation by the following conditional independence assumptions.

a. Conditional Independence Assumptions

(a.1) Given η_k, $k = 1, \ldots, K$, the responses are conditionally independent.

(a.2) Priors $p(\boldsymbol{\beta}_{j,k}|\tau^2_{j,k})$, $j = 0, \ldots, J_k$, are conditionally independent.

(a.3) Priors $p(\beta_{0,k})$ and hyperpriors $p(\tau^2_{j,k})$ are mutually independent.

Consequently, the posterior

$$p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\tau}^2_1, \ldots, \boldsymbol{\tau}^2_K|\boldsymbol{y})$$

with $\boldsymbol{\beta}_k = (\beta_{0,k}, \boldsymbol{\beta}_{1,k}{}', \ldots, \boldsymbol{\beta}_{J_k,k}{}')'$, $\boldsymbol{\tau}^2_k = (\tau^2_{1,k}, \ldots, \tau^2_{J_k,k})'$ and $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is up to a normalising constant proportional to

$$\prod_{i=1}^{n} p(y_i|\eta_{i1}, \ldots, \eta_{iK}) \prod_{k=1}^{K} \left[p(\beta_{0,k}) \prod_{j=1}^{J_k} \left[p(\boldsymbol{\beta}_{j,k}|\tau^2_{j,k}) p(\tau^2_{j,k}) \right] \right].$$

Note that assumption (a.2) may be critical in some cases. For example, one often observes empirically that random intercepts and random slopes tend to be negatively correlated unless the covariate associated with the random slope is centred. In such situations, independence between the vectors of regression coefficients comprising the random intercepts and the random slopes, respectively, would be questionable. In fact, in such cases the separate specification of user-defined scaling for the random intercepts and the random slopes may itself be questionable. However, in general prior independence should be a reasonable working assumption which also does not rule out posterior dependence. Moreover, specifying the amount of scaling separately for each effect certainly facilitates prior elicitability.

3.2 Gaussian Mean Regression

Assume in this section a Gaussian mean regression model for $\boldsymbol{y} = (y_1, \ldots, y_n)'$, i.e.

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \sum_{j=1}^{J} \boldsymbol{Z}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \tau^2_\varepsilon \boldsymbol{I}_n) \tag{12}$$

with inverse gamma prior

$$p(\tau^2_\varepsilon) \propto \frac{1}{(\tau^2_\varepsilon)^{a_\varepsilon + 1}} \exp\left(\frac{\theta_\varepsilon}{\tau^2_\varepsilon}\right)$$

for the error variance and remaining prior specifications as defined before. Note that $k = 1$ in this section and that J_k is replaced by J. Applying a mixed model representation as introduced in Section 2 allows us writing (12) as

$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{\beta}_{unpen} + \boldsymbol{V}\boldsymbol{\beta}_{pen} + \boldsymbol{\varepsilon}. \tag{13}$$

We furthermore assume that $\boldsymbol{U}$ has full column rank $\text{rk}(\boldsymbol{U}) = r$, see Remark 2(iii) for details on how this is achieved. Define for $\boldsymbol{\xi} = (\boldsymbol{\beta}'_{unpen}, \boldsymbol{\beta}'_{pen})'$ and $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{V})$ the projection on the orthogonal complement of $\boldsymbol{U}$ as

$$\boldsymbol{R}_1 = \boldsymbol{I}_n - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}', \tag{14}$$

the residual sum of squares

$$\text{SSE} = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}')\boldsymbol{y} \tag{15}$$

and for $t = \text{rk}(\boldsymbol{X}) - \text{rk}(\boldsymbol{U}) \leq \dim(\boldsymbol{\beta}_{pen})$

$$\text{rk}(\boldsymbol{X}) = \text{rk}(\boldsymbol{U}, \boldsymbol{V}) = r + t. \tag{16}$$

Remark 2.

(i) All rank conditions can be formulated directly for the reparameterised model (13) and do not have to be traced back to the original parameterisation. To see this, define $\boldsymbol{Z}^* = \boldsymbol{Z}\boldsymbol{S} = \boldsymbol{Z}\,\text{diag}(\boldsymbol{S}_j)$ with

$$\boldsymbol{\beta}_j = \boldsymbol{S}_j\boldsymbol{\xi}_j = (\tilde{\boldsymbol{U}}_j, \tilde{\boldsymbol{V}}_j)\boldsymbol{\xi}_j.$$

The matrix $\boldsymbol{S}$ has full rank such that $\text{rk}(\boldsymbol{Z}^*) = \text{rk}(\boldsymbol{Z})$ and $\text{rk}(\boldsymbol{R}_1\boldsymbol{Z}^*) = \text{rk}(\boldsymbol{R}_1\boldsymbol{Z})$ hold and finally

$$\boldsymbol{R}_1\boldsymbol{Z}^* = \boldsymbol{R}_1[(\boldsymbol{U}, \boldsymbol{0}) + (\boldsymbol{0}, \boldsymbol{V})] = (\boldsymbol{R}_1\boldsymbol{U}, \boldsymbol{0}) + (\boldsymbol{0}, \boldsymbol{R}_1\boldsymbol{V}) = (\boldsymbol{0}, \boldsymbol{R}_1\boldsymbol{V})$$

which proves

$$\text{rk}(\boldsymbol{R}_1\boldsymbol{Z}) = \text{rk}(\boldsymbol{R}_1\boldsymbol{V}) = t.$$

(ii) With $\text{rk}(\boldsymbol{U}) = r$ it can be shown that

$$\text{rk}(\boldsymbol{U}, \boldsymbol{V}) = r + t \iff \text{rk}(\boldsymbol{R}_1\boldsymbol{V}) = \text{rk}(\boldsymbol{V}'\boldsymbol{R}_1\boldsymbol{V}) = t.$$

The proof is in Supplement A.2 with Proposition 1.

(iii) In order to obtain a full column rank matrix of unpenalised effects in the mixed model representation (13), all superfluous columns have to be deleted. In particular, duplicated constant columns representing the levels of the functions are deleted which is a simple way to include the centring restrictions and is equivalent to the centring of functions that we include in our MCMC algorithm. Furthermore, using the one-to-one relationship between original parameterisation (11) and the reparameterised model (13) the restrictions for one presentation can be deduced from the other one. Hence, proving propriety of the posterior in the original model can directly be based on the mixed model which will be done in the following.

b. Conditions for Gaussian Mean Regression

(b.1) $\kappa_j - \sum_{j=1}^{J} \kappa_j + t - 1 > 0$, $j = 1, \ldots, J$.

(b.2) $n - r - J + 2a_\varepsilon > 0$.

(b.3) $\text{SSE} + 2\theta_\varepsilon > 0$.

Condition (b.1) relates the rank κ_j of the prior precision matrix of one effect with $\boldsymbol{K}_j \neq \mathbf{0}$ to the rank of all prior precision matrices. Condition (b.2) restricts the number of all effects to be smaller or equal to the number of observations but can be relaxed by increasing the hyperparameter value a_ε. Condition (b.3) is always fulfilled for $\theta_\varepsilon > 0$. In case of an improper prior for τ_ε^2, i.e. $a_\varepsilon < 0$, $\theta_\varepsilon = 0$, $\text{SSE} > 0$ has to be assured. If the number of parameters is equal or larger to the number of observations such that the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ can be interpolated, $\theta_\varepsilon > 0$ becomes necessary. In all other cases $\text{SSE} > 0$ holds almost surely such that $\theta_\varepsilon = 0$ can be chosen in most situations.

Theorem 2. *Consider the Gaussian mean regression model* (12) *with mixed model representation* (13) *and rank conditions from* (16)*. Then, condition* (b.3) *is necessary for the propriety of the joint posterior while conditions* (b.1), (b.2) *and* (b.3) *are sufficient for the propriety of the joint posterior*

The proof of Theorem 2 is in Supplement A.3 using Lemma 2 and Proposition 1 of Supplement A.2.

Remark 3.

(i) In Gaussian mean regression with inverse gamma priors for τ_j^2, additional conditions on the ranks κ_j and the number of effects compared to the shape parameters a_j of the priors are required. Consequently, one has to consider the cases $t = \kappa$ or $J = 1$ as well as $t < \kappa$ and $J > 1$ separately.

(ii) Compared to the inverse gamma case, the sufficient conditions with scale-dependent priors are stronger. This intuitively makes sense since the scale-dependent prior places additional prior mass close to the base model which is the improper part of the partially improper normal priors. On the other hand, the necessary conditions are somewhat weaker with scale-dependent priors.

3.3 Distributional Regression

Assume in this section a distributional regression model as in (M1) and (M2). The basic idea to obtain sufficient conditions for the propriety is to formulate all $k = 1, \ldots, K$ predictors $\boldsymbol{\eta}_k$ in a mixed model representation for an appropriate submodel. The required assumptions are then the following:

c. Conditions for Distributional Regression.

Assume that the set of observations can (after re-ordering) be partitioned such that for $n^* \geq 1$

(c.1) $\int \cdots \int p(y_i|\eta_{i1},\ldots,\eta_{iK})\mathrm{d}\eta_{i1}\cdots\mathrm{d}\eta_{iK} < \infty$ for $i = 1,\ldots,n^*$.

(c.2) $p(y_i|\eta_{i1},\ldots,\eta_{iK}) \leq M$ for $i = n^*+1,\ldots,n$.

This implies that for at least one observation the density is integrable (with respect to the predictors) and that all remaining densities are bounded. For discrete distributions, all densities are automatically bounded by 1 so that only Condition (c.1) can be an issue in practice. Condition (c.1) is usually fulfilled if certain restrictions apply on specific parameters that exclude extreme values on the boundary of the parameter space, see Klein, Kneib and Lang (2015) for a more detailed discussion on count data distributions. This is similar to the case of logistic regression with binary responses where finite integrals can be achieved by restricting the parameter space for the success probability to either $\pi_i > 0$ for at least one observation $y_i = 0$ or $\pi_i < 1$ for one $y_i = 1$ (excluding one of the boundary cases). For continuous distributions, the densities are sometimes not bounded (e.g. for the gamma distribution). Note that this is not a problem when all observations fulfil Condition (c.1) since $n^* = n$ is allowed. Similar as for the discrete distributions, integrability of the densities can be assured by the assumption that none of the distributional parameters is on the boundary of the parameter space (an assumption that would also have to be made to apply standard maximum likelihood asymptotics).

To explicitly differentiate between model components with proper and improper prior, we apply a mixed model representation similar as before but now to all K predictor equations in (M2) in order to obtain vectors of fixed effects with flat prior and i.i.d. Gaussian random effects with proper prior and dimension $\kappa_{j,k} = \mathrm{rk}(\boldsymbol{K}_{j,k})$. Furthermore, we separate the random effect with largest dimension in each predictor such that we obtain

$$\boldsymbol{\eta}_k = \boldsymbol{U}_k\boldsymbol{\beta}_{unpen,k} + \boldsymbol{V}_k\boldsymbol{b}_k + \boldsymbol{V}_{\varepsilon,k}\boldsymbol{b}_{\varepsilon,k}$$

where $\boldsymbol{U}_k\boldsymbol{\beta}_{unpen,k}$ comprises all predictor components with a flat prior, $\boldsymbol{V}_{\varepsilon,k}\boldsymbol{b}_{\varepsilon,k}$ corresponds to the random effect with the largest dimension, $\dim(\boldsymbol{b}_{\varepsilon,k}) = \kappa_{\varepsilon,k}$, and $\boldsymbol{V}_k\boldsymbol{b}_k$ contains all remaining random effects. Note that $\boldsymbol{b}_k$ is based on $J_k^* = J_k - 1$ effects in the notation of (M2) and $\boldsymbol{\beta}_{pen}$ in the mixed model representation in (4) corresponds to the vector resulting from $\boldsymbol{b}_k$ and $\boldsymbol{b}_{\varepsilon,k}$ where, without loss of generality, we assume that the effects in the predictors are ordered such that the J_kth effects correspond to the random effects in the mixed model representation with largest dimensions. Similarly, the design matrices $\boldsymbol{V}_k$ and $\boldsymbol{V}_{\varepsilon,k}$ correspond to the design matrix of the penalised part in (4).

Let $\tilde{k}_\varepsilon = \min\{\kappa_{\varepsilon,1},\ldots,\kappa_{\varepsilon,K}\}$ and assume that we can choose $\tilde{k}_\varepsilon$ observations including at least one observation fulfilling (c.1) to define the submodel

$$\boldsymbol{\eta}_{k,s} = \boldsymbol{U}_{k,s}\boldsymbol{\beta}_{unpen,k} + \boldsymbol{V}_{k,s}\boldsymbol{b}_k + \boldsymbol{V}_{\varepsilon,k,s}\boldsymbol{b}_{\varepsilon,k} \tag{17}$$

corresponding to these observations. Then the following rank conditions have to be fulfilled:

(c.3) The design matrix $\boldsymbol{U}_{k,s}$ has full rank r_k.

(c.4) $\text{rk}(\boldsymbol{U}_k, \boldsymbol{V}_k) = \text{rk}(\boldsymbol{U}_{k,s}, \boldsymbol{V}_{k,s}) = r_k + t_k$.

(c.5) $\text{rk}(\boldsymbol{V}_{\varepsilon,k,s}) = \tilde{k}_\varepsilon$, i.e. $\boldsymbol{V}_{\varepsilon,k,s}$ is of full rank.

To ensure (c.3), superfluous columns arising from the reparameterisation have to be deleted. In particular, duplicated constant columns representing the levels of the functions are deleted, similar to Gaussian mean regression and explained in Remark 2(iii). Condition (c.4) indicates that the rank of the design matrices in the submodel is the same as in the complete model whereas (c.5) defines a similar restriction for the design matrix of the largest random effect arising from the mixed model representation.

Finally, define the normalised submodel

$$\tilde{\boldsymbol{\eta}}_{k,s} = \tilde{\boldsymbol{U}}_{k,s}\boldsymbol{\beta}_{unpen,k} + \tilde{\boldsymbol{V}}_{k,s}\boldsymbol{b}_k + \varepsilon_{k,s} \tag{18}$$

that is obtained by multiplying (18) with $\boldsymbol{M}_k = (\boldsymbol{V}_{\varepsilon,k,s}\boldsymbol{V}_{\varepsilon,k,s}{}')^{-1/2}$ such that $\tilde{\boldsymbol{\eta}}_{k,s} = \boldsymbol{M}_k\boldsymbol{\eta}_{k,s}$, $\tilde{\boldsymbol{U}}_{k,s} = \boldsymbol{M}_k\boldsymbol{U}_{k,s}$, $\tilde{\boldsymbol{V}}_{k,s} = \boldsymbol{M}_k\boldsymbol{V}_{k,s}$, and $\boldsymbol{\varepsilon}_{k,s} \sim \text{N}(\boldsymbol{0}, \tau^2_{\varepsilon,k}\boldsymbol{I}_{\tilde{k}_\varepsilon})$ represents an i.i.d. random effect since $\boldsymbol{V}_{\varepsilon,k,s}\boldsymbol{b}_{\varepsilon,k} \sim \text{N}(\boldsymbol{0}, \tau^2_{\varepsilon,k}\boldsymbol{V}_{\varepsilon,k,s}\boldsymbol{V}_{\varepsilon,k,s}{}')$.

The corresponding residual sum of squares for the normalised submodel are then defined as

$$\text{SSE}_{k,s} := \left(\tilde{\boldsymbol{\eta}}_{k,s} - \tilde{\boldsymbol{U}}_{k,s}\boldsymbol{\beta}_{unpen,k} - \tilde{\boldsymbol{V}}_{k,s}\boldsymbol{b}_k\right)' \left(\tilde{\boldsymbol{\eta}}_{k,s} - \tilde{\boldsymbol{U}}_{k,s}\boldsymbol{\beta}_{unpen,k} - \tilde{\boldsymbol{V}}_{k,s}\boldsymbol{b}_k\right). \tag{19}$$

Based on this, we finally require the following additional conditions:

(c.6) $\kappa_{j,k} - \sum_{j=1}^{J_k^*} \kappa_{j,k} + t_k - 1 > 0$, $j = 1, \ldots, J_k^*$, $k = 1, \ldots, K$.

(c.7) $\text{SSE}_{k,s} > 0$.

Condition (c.6) relates the rank of the random effects part of one individual effect to the sum of all rank deficiencies in the corresponding predictor and requires that the dimensionality is not too small. Condition (c.7) requires that there is variation in the residual sum of squares in the normalised submodel (implying that not all effects are zero).

Theorem 3. *Consider the distributional regression model from* (M1) *and* (M2)*. Then, Conditions* (c.1), (c.2) *on the densities,* (c.3) *to* (c.5) *on the ranks as well as* (c.6) *and* (c.7) *are sufficient conditions for a proper posterior.*

A proof for this theorem is in Supplement A.4.

Remark 4.

(i) Compared to Gaussian mean regression in Section 3.2 (where y is directly related to the mixed model representation rather than the predictors as in distributional regression) and the distributional approach with inverse gamma priors $\boldsymbol{\varepsilon}_{k,s}$ now has a scale-dependent hyperprior with scale parameter θ_ε.

(ii) The conditions on the densities (c.1) and (c.2) as well as the conditions on the ranks in the reparameterised model (c.3) to (c.5) are independent of the hyperpriors for τ_j^2 and hence identical to the conditions in Klein, Kneib and Lang (2015). Conditions (c.6) to (c.7) in contrast are stronger with scale-dependent hyperpriors. For instance, Condition (c.7) of variation in the $\text{SSE}_{k,s}$ could be relaxed with inverse gamma hyperpriors by choosing the scale parameter θ_ε of the errors $\boldsymbol{\varepsilon}_{k,s}$ sufficiently large. With inverse gamma priors there is an additional sufficient condition relating the rank of the random effect determining the working error of the submodel to the number of coefficients with flat prior and the number of terms with Gaussian priors in one of the predictors. With scale-dependent hyperpriors, this condition is included in Condition (c.7).

(iii) Theorem 3 gives sufficient but not necessary conditions such that the posterior can in fact still be proper even if one or more conditions are violated.

4 Inference

While the often employed $\text{IG}(a, \theta)$ priors in structured additive regression for τ^2 induce full conditionals that are again inverse gamma distributions (and hence Gibbs sampling steps can be implemented) this is no longer the case for the scale-dependent priors. To derive proposal densities for a Metropolis–Hastings type update that automatically adapt to the form of the full conditional distributions without manual tuning, we locally approximate the full conditionals of the log-variances by a normal distribution matching the mode and the curvature. This is based on the idea of iteratively weighted least squares (IWLS, Gamerman, 1997) proposals originally developed for updating the vectors of regression coefficients, see, e.g. Brezger and Lang (2006) and Gamerman (1997).

4.1 Metropolis–Hastings Updating Scheme

4.1.1 Full Conditional Distributions for Regression Coefficients

IWLS proposals for different types of distributions in the distributional regression framework have been investigated in Klein, Kneib and Lang (2015); Klein, Kneib, Lang and Sohn (2015) and we only briefly recall the principles here. Let $l(\boldsymbol{\eta}_k) \propto \log(\prod_{i=1}^n p(y_i|\eta_{i1}, \ldots, \eta_{iK}))$ denote the log-likelihood part depending on the predictor with index k (where $\propto$ is abused to denote equality up to an additive constant). Then the logarithmic full conditional $\log(p(\boldsymbol{\beta}_{j,k}|\cdot))$ of $\boldsymbol{\beta}_{j,k}$ is up to additive constants equal to

$$l(\boldsymbol{\eta}_k) - \frac{1}{2\tau_{j,k}^2}\boldsymbol{\beta}_{j,k}'\boldsymbol{K}_{j,k}\boldsymbol{\beta}_{j,k}.$$

A local approximation fitting the mode and the curvature at the mode suggests to propose $\boldsymbol{\beta}_{j,k}$ from a multivariate Gaussian distribution $\text{N}(\boldsymbol{\mu}_{j,k}, \boldsymbol{P}_{j,k}^{-1})$ with distribution- and parameter-specific mean $\boldsymbol{\mu}_{j,k}$ and precision matrix $\boldsymbol{P}_{j,k}$. A proposal is then accepted with acceptance probability according to a Metropolis–Hastings type step (Hastings, 1970).

4.1.2 Full Conditional Distributions for the Smoothing Variances

We follow a similar strategy as for $\boldsymbol{\beta}_{j,k}$ to construct Gaussian proposal densities for the smoothing variances. While asymptotic theory suggests asymptotic normality of all model parameters with appropriate location and scale parameters, a Gaussian distribution as proposal density for $\tau^2_{j,k}$ may cause invalid proposals, i.e. values smaller than zero, in several MCMC steps when the variances are small. This problem can be overcome by approximating the log-full conditional $\log(p(\log(\tau^2_{j,k})|\cdot))$ of $\log(\tau^2_{j,k})$ rather than the log-full conditional $\log(p(\tau^2_{j,k}|\cdot))$ of $\tau^2_{j,k}$. Applying the change of variable theorem with transformation $u = \log(\tau^2)$ to the full conditional for $\tau^2_{j,k}$ yields

$$\log(p(u|\cdot)) \equiv l_u \propto \frac{u}{2}(1 - \text{rk}(\boldsymbol{K}_{j,k})) - \frac{1}{2\exp(u)}\boldsymbol{\beta}'_{j,k}\boldsymbol{K}_{j,k}\boldsymbol{\beta}_{j,k} - \frac{(\exp(u))^{1/2}}{\sqrt{\theta}}.$$

Approximating l_u by a second order Taylor expansion around the current state $u^{(c)}$ and taking the exponent yields the proposal density $\text{N}(\mu_u, \sigma^2_u)$ with

$$\mu_u = \sigma^2_u \frac{\partial l_u}{\partial u} + u \quad \text{and} \quad \sigma^2_u = -1/\frac{\partial^2 l_u}{\partial u^2}$$

from which we obtain a proposal for $\log(\tau^2_{j,k})$.

Remark 5.

(i) It can be shown that the variance of the proposal density arising from the quadratic approximation of the log-full conditional is always positive. This result is based on the fact that l_u is strictly concave and that $\partial^2 l_u/\partial u^2 = 0$ can be excluded under the (mild) condition $\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta} > 0$, see Supplement F.3 for details.

(ii) Although the scale-dependent prior requires a Metropolis–Hastings instead of a Gibbs update, this has virtually no impact on computation times since we are dealing with a scalar parameter only. In all our analyses, acceptance probabilities ranged from roughly 70% to 90% and the mixing behaviour did not show any conspicuous features.

(iii) Let $u^{(p)}$ be the proposal coming from the proposal density $q(u^{(p)}|u^{(c)}) = \text{N}(\mu_u^{(c)}, (\sigma^2_u)^{(c)})$ with current value $u^{(c)} = \log((\tau^2_{j,k})^{(c)})$ plugged in. The acceptance probability α_{accept} of the Metropolis–Hastings steps is then computed as

$$\alpha_{accept} = \min\left\{1, \frac{p(u^{(p)}|\cdot)q(u^{(c)}|u^{(p)})}{p(u^{(c)}|\cdot)q(u^{(p)}|u^{(c)})}\right\} \tag{20}$$

which is in distribution equivalent to the acceptance probability

$$\tilde{\alpha}_{accept} = \min\left\{1, \frac{p((\tau^2_{j,k})^{(p)}|\cdot)\tilde{q}((\tau^2_{j,k})^{(c)}|(\tau^2_{j,k})^{(p)})}{p((\tau^2_{j,k})^{(c)}|\cdot)\tilde{q}((\tau^2_{j,k})^{(p)}|(\tau^2_{j,k})^{(c)})}\right\}$$

where $\tilde{q}$ is the lognormal proposal obtained by the change of variable theorem from q, see Supplement F.2 for a detailed derivation. The proposals will even be

identical, when random number generation of lognormally distributed variates is based on taking the logarithm of a normal variate.

(iv) As an alternative to IWLS proposals, we implemented inverse gamma proposals with parameters chosen to match shape of the scale-dependent hyperpriors. As a second option, we approximated the Weibull distribution with a gamma distribution having the advantage that the resulting full conditional is a generalised inverse Gaussian distribution. Both options turned out to be less satisfactory, both due to theoretical considerations but also in their practical performance.

4.1.3 Inference for Alternative Prior Structures

In Section 2.4, we have discussed how the hyperparameters of various alternative prior structures can be chosen according to a similar scaling criterion as for the scale-dependent priors. To make these priors work in practice, we require suitable proposal densities and we follow the same construction principle as for the scale-dependent prior. More specifically, we approximate the log-full conditional quadratically to obtain a local Gaussian approximation. Details on the resulting derivatives can be found in Supplement F.1.

As for the scale-dependent prior, half-Cauchy, half-normal and the approximate proper uniform prior lead to strictly concave log-full conditionals such that the Gaussian proposal has positive variance, compare Supplement F.3.

4.2 Implementation

The new prior structure developed in this paper as well as the alternative prior structures (half-normal, half-Cauchy, approximate uniform for τ) are integrated into the open source software BayesX (Belitz et al., 2015), version 3.0.2. A generic MCMC algorithm is provided in Supplement F.4. We utilise methods for efficient storing of large data sets and sparse matrix algorithms for sampling from multivariate Gaussian distributions (George and Liu, 1981; Rue, 2001) and profit from existing procedures for computing simultaneous confidence bands for nonparametric effects as developed in Krivobokova et al. (2010) or to specify multilevel models (Lang et al., 2014) that can immediately be combined with the new prior structures.

5 Empirical Evaluation

We conducted several simulations in which we compared the performance of scale-dependent priors for varying levels of the probability statement α with the one of inverse gamma priors with both parameters equal and small, flat priors for τ and τ^2, inverse gamma priors with shape parameter $a = 1$ and scale parameter optimised according to the scaling criterion, as well as the other alternative priors discussed in Section 2.4 (again with optimised parameter values). We consider three distinct scenarios:

Scenario 1 Effects close to the base model, i.e. effects that are close to linear in a penalised spline specification.

Scenario 2 A flat likelihood due to a small amount of information in the data.

Scenario 3 Strong deviations from the base model, i.e. nonlinear effects with strong curvature.

In the following, we restrict the presentation to basic settings and core results while the simulations are documented in detail in Supplement G.

5.1 Simulation Settings

In all the scenarios considered, we keep simulation settings fixed to allow for a consistent comparison across the settings:

- *Sample sizes* n are chosen from $\{50, 100, 250, 500, 1000\}$.
- *Covariates* are restricted to one single scalar covariate $x \sim \mathrm{U}(-1, 1)$.
- *Simulated effects* visualised in Figure 2 are

  Scenario 1 $f(x) = \sin(x)$.
  Scenario 2 $f(x) = 7\exp(-\exp(5x))$.
  Scenario 3 $f(x) = 1.5\sin(1.25\pi x + 0.5)/\exp(x)$.

- *Responses*
  - $y \sim \mathrm{N}(f(x), 1)$ in Scenario 1 and 3.
  - $y \sim \mathrm{Be}(\pi)$, $\pi = \exp(f(x))/(1 + \exp(f(x)))$ in Scenario 2.
- *Hyperprior specifications*
  - (ϵ, ϵ)-inverse gamma priors ($\mathrm{IG}(\epsilon, \epsilon)$) with $\epsilon \in \{0.001, 0.01\}$ for τ^2.
  - $\mathrm{IG}(1, \epsilon)$, $\epsilon = 0.005$ as used frequently in the literature as well as $\mathrm{IG}(1, \theta)$ (denoted as $\mathrm{IG}(\alpha)$ in the following) for τ^2.
  - Flat priors for standard deviations τ and variances τ^2.
  - Half-normal ($\mathrm{HN}(\alpha)$), half-Cauchy ($\mathrm{HC}(\alpha)$) and approximate uniform ($\mathrm{U}(\alpha)$) prior with scale parameters θ for τ.
  - Scale-dependent priors ($\mathrm{SD}(\alpha)$) with scale parameters θ for τ^2.
- *Scale parameters* θ are computed with fixed values $P = 1$, $c = 3$ and $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.5, 0.7\}$.

For each scenario $R = 1000$ replications are used to compute mean squared errors (MSEs), bias, credible intervals and coverage rates.

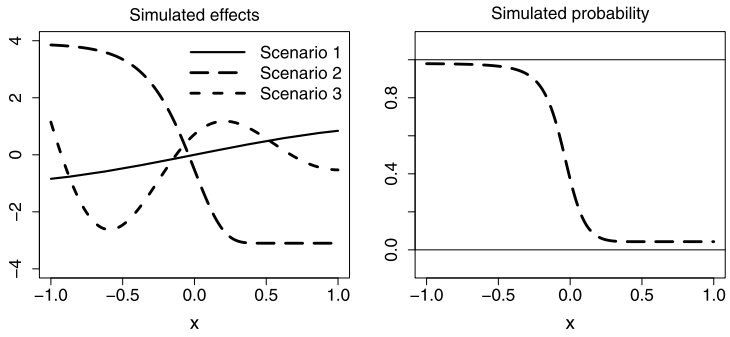

Figure 2: Empirical Evaluation. Centred simulated functions $f(x)$ in Scenarios 1 to 3 (left) and success probability π with inverse logit link in Scenario 2.

5.2 Results

Scenario 1: Close to the Base Model Estimates for f are less wiggly under scale-dependent priors as compared to inverse gamma priors (Figures G13, G14) in the sense that the latter tend to overestimate the sigmoid shape. For scale-dependent, half-normal and uniform priors, MSEs and widths of credible intervals turn out to be pretty insensitive with respect to the different values of α while variation for the inverse gamma and half-Cauchy priors is larger, compare Figures 3 and 4. However, the width of credible intervals is larger with inverse gamma priors as compared to the ones of scale-dependent priors for $\alpha < 0.7$. While the bias turns out to be similar for all priors (Figure G12), scale-dependent priors deliver narrower credible intervals which still maintain the desired coverage levels on average as shown in Figures G10, G11. Across the sample sizes, scale-dependent priors show a very positive overall performance as compared to all alternative priors, in particular to inverse gamma as well as flat priors which have higher MSEs and larger widths of credible intervals. Note that the smallest log(MSE) and widths of credible intervals of half-Cauchy priors for $\alpha = 0.01$ are at least partly misleading since Figures G10 to G14 give evidence for problems in maintaining the desired coverage levels. Basically the half-Cauchy prior induces a very strong preference for variances close to zero that allow to identify linear effects fairly well but also bears the risk of undercoverage.

Scenario 2: Flat Likelihood Due to the model specification, the success probability π is likely to be close to one or close to zero for most observations which leads to a rather low level of information on the regression effects and therefore a flat likelihood. During the MCMC run, proposals for the regression coefficients yielding either highly negative or highly positive predictor values occur frequently (in up to 26% of the replications for n small) with $\text{IG}(\epsilon, \epsilon)$ and flat priors. The consequence are numerical instabilities or posterior mean function estimates with huge MSEs and wide credible intervals when the sample size is small. In our simulation, these problems are less frequent (less than 1% software crashes) when scale-dependent and the other alternative priors are em-

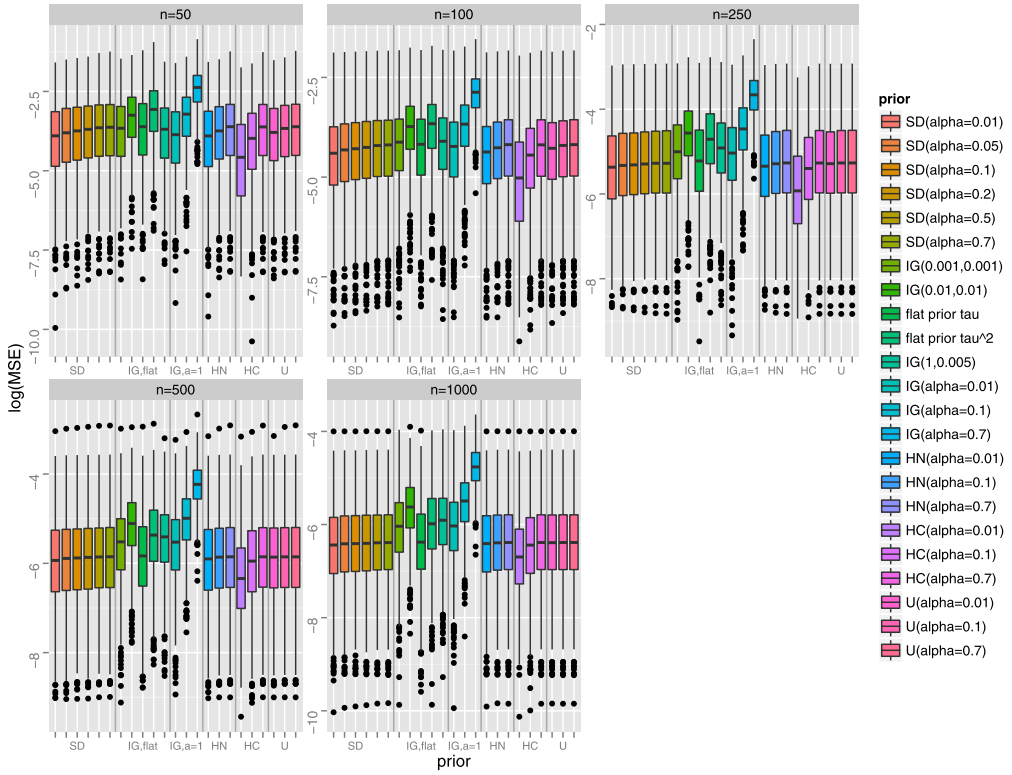

Figure 3: Empirical Evaluation, scenario 1. Shown are boxplots of log(MSE) for different sample sizes n and the different prior settings.

ployed. This beneficial behaviour results from user-defined scaling that helps to avoid the very extreme and unlikely predictor constellations. Consequently, MSEs and credible intervals of estimates in small data sets are better for priors with parameters chosen according to the scaling criterion, compare Figures 5 and G15. While all priors cannot maintain the coverage levels for small sample sizes, increasing the sample size mostly solves this problem except for half-normal and uniform priors, see Figures G16, G17. Overall, again the scale-dependent prior seems to be pretty robust. Note that the extremely small MSEs observed for half-Cauchy priors for $\alpha = 0.01$ as seen in Scenario 1 cannot be found in this setting since the performance does not gain from the strong preference of small variances.

Scenario 3: Strongly Nonlinear Effect In this setting, all priors perform similarly well with respect to MSEs, coverage rates and widths of credible intervals except the IG(0.7) and U(0.01) specifications, compare Figures G21 to G22. While IG(0.7) priors have markedly higher MSEs, width of credible intervals and coverage rates, U(0.01) priors yield higher MSEs but obviously much to narrow pointwise credible intervals in

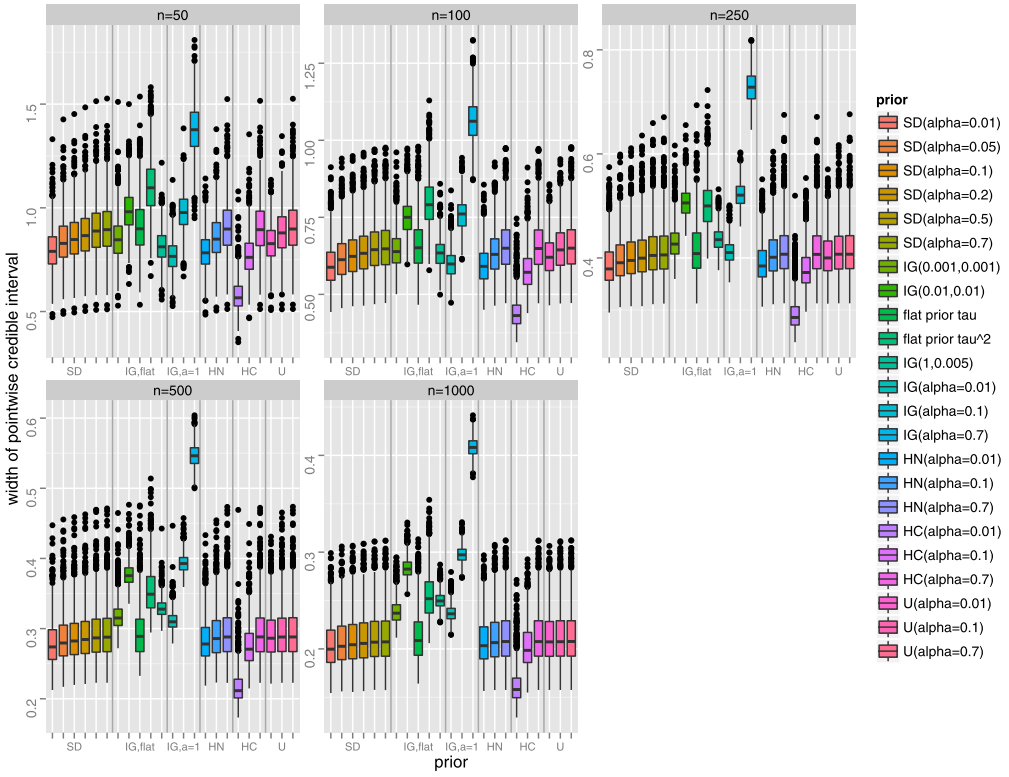

Figure 4: Empirical Evaluation, scenario 1. Shown are widths of pointwise credible intervals for different sample sizes n and the different prior settings.

the inner range of x, compare Figure G24. The figure also indicates that half-normal priors tend to underestimate the desired pointwise coverage rates for $x \leq -0.5$. The remaining results are reasonable since the effect signal is moderately large and far from a linear effect. Basically (and as expected) differences between the prior specifications tend to be small when enough information is contained in the data.

Overall Conclusion In summary, inverse gamma as well as flat priors turn out to be rather questionable in situations that require a preference for the base model or that carry only weak information about the effects of interest. The recently proposed half-normal, half-Cauchy and uniform priors for τ are then often a better choice. However, depending on the data, we found that in terms of credible intervals, robustness or accuracy of point estimates, scale-dependent priors can yield slightly better results in some situations and, more importantly, appear to perform robustly well across all considered aspects. Hence, the new class of scale-dependent priors can be considered a reasonable default option.

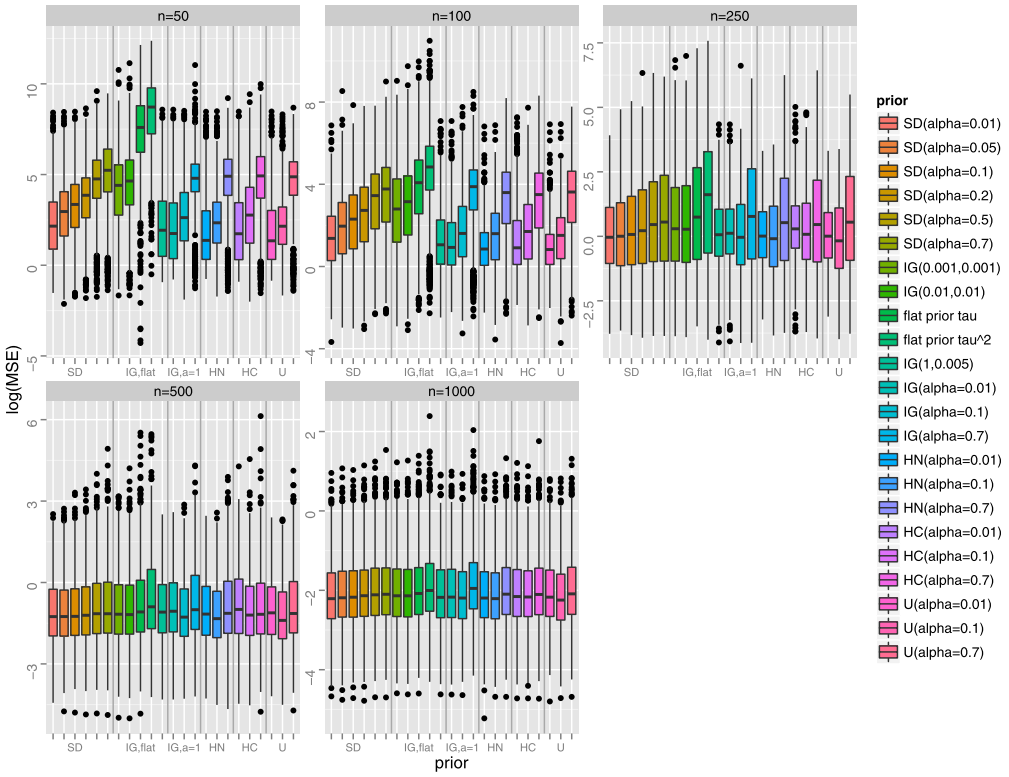

Figure 5: Empirical Evaluation, scenario 2. Shown are boxplots of log(MSE) for different sample sizes n and the different prior settings.

6 Applications

6.1 Patent Citations

Klein, Kneib and Lang (2015) analysed the number of citations (*ncit*) of patents granted by the European Patent Office (EPO) comparing different distributional regression models for zero-inflated and overdispersed count data. Following previous analyses by Jerak and Wagner (2006) and Klein, Kneib and Lang (2015), we removed roughly 1% of observations with extreme values for some of the covariates such that the data set consists of $n = 4805$ patents. Information on several covariates includes the continuous variables grant year (*year*), number of designated states (*ncountry*) and number of claims (*nclaims*), see, e.g. Jerak and Wagner (2006) for more details on the data.

The results of Klein, Kneib and Lang (2015) show that (conditionally) the number of citations are highly overdispersed while there is only small evidence for zero-inflation. In addition, Klein, Kneib and Lang (2015) had problems in estimating a 'full' zero-inflated negative binomial model, i.e. a model with conditional density

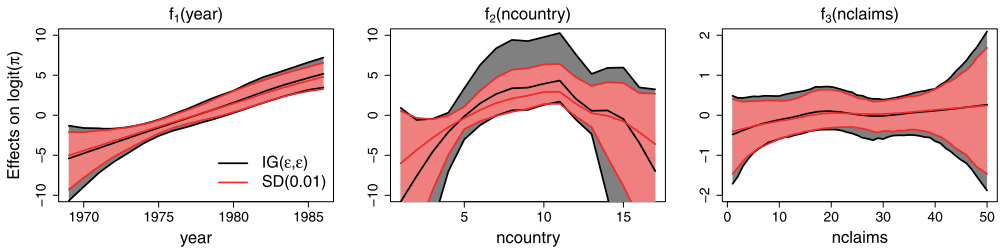

Figure 6: Patent citations. Comparison of estimated nonlinear effects of continuous covariates *year*, *ncountry* and *nclaims* on logit(π). Shown are posterior means and 95% credible intervals with scale-dependent priors, $\alpha = 0.01$, $c = 1$, (red) and inverse gamma priors IG(ϵ, ϵ), $\epsilon = 0.001$, (black).

$$p(ncit_i|\mu_i, \delta_i, \pi_i) = \pi_i \mathbb{1}_{\{ncit_i=0\}} + (1-\pi_i)\frac{\Gamma(ncit_i + \delta_i)}{\Gamma(ncit_i + 1)\Gamma(\delta_i)} \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} \left(\frac{\mu_i}{\delta_i + \mu_i}\right)^{ncit_i}$$

where each of the distribution parameters μ_i, δ_i and π_i, $i = 1, \ldots, n$ is linked to a generic predictor of the form

$$\eta_i = f_1(ncountry_i) + f_2(year_i) + f_3(nclaims_i) + \boldsymbol{x}_i'\boldsymbol{\beta}.$$

In particular, partly non-stationary MCMC paths for the constant in π_i have been observed which are related to the small evidence for zero-inflation, i.e. the 'true' π_i is rather small for some covariate combinations which induces a flat likelihood. Replacing inverse gamma priors IG(ϵ, ϵ) ($\epsilon = 0.001$) for the spline coefficients of f_1 to f_3 with the scale-dependent priors ($\alpha = 0.01$, $c = 3$, $P = 1$) reduces the convergence issues and avoids the negatively affected relative effect estimates as shown in Figures 6 and 7 (first and second column).

While exemplary sample paths of one coefficient for the effects of *year* and *nclaims* in Figure 7 do not indicate any convergence problems, the constant β_0 and the coefficients for the effect of *ncountry* show a non-stationary pattern with inverse gamma priors (second and third column). With the scale-dependent prior (first column), half-normal (third column) and half-Cauchy prior (fourth column) these problems can be reduced without requiring manual tuning. Approximate uniform priors have the worst convergence behaviour of β_0.

The effect of *year* on logit (π) is close to linear and the scale-dependent prior delivers less conservative credible intervals, similar as for the effect of *nclaims*. However, this effect is not significant. At the boundary values of *ncountry*, the inverse gamma prior yields strongly negative predictor values causing numerical instabilities. The scale-dependent prior seems to penalise against these extreme values (which is reasonable due to the construction Principle 4 in Section 2.3) and enhances numerical stability in this situation. Figures H28 to H31 (Supplement H) indicate narrower credible intervals for all effects when half-normal, half-Cauchy or uniform priors are used. However, from simulations this has to be interpreted with caution due to the fact that for small α these two priors tend to underestimate the desired coverage level.

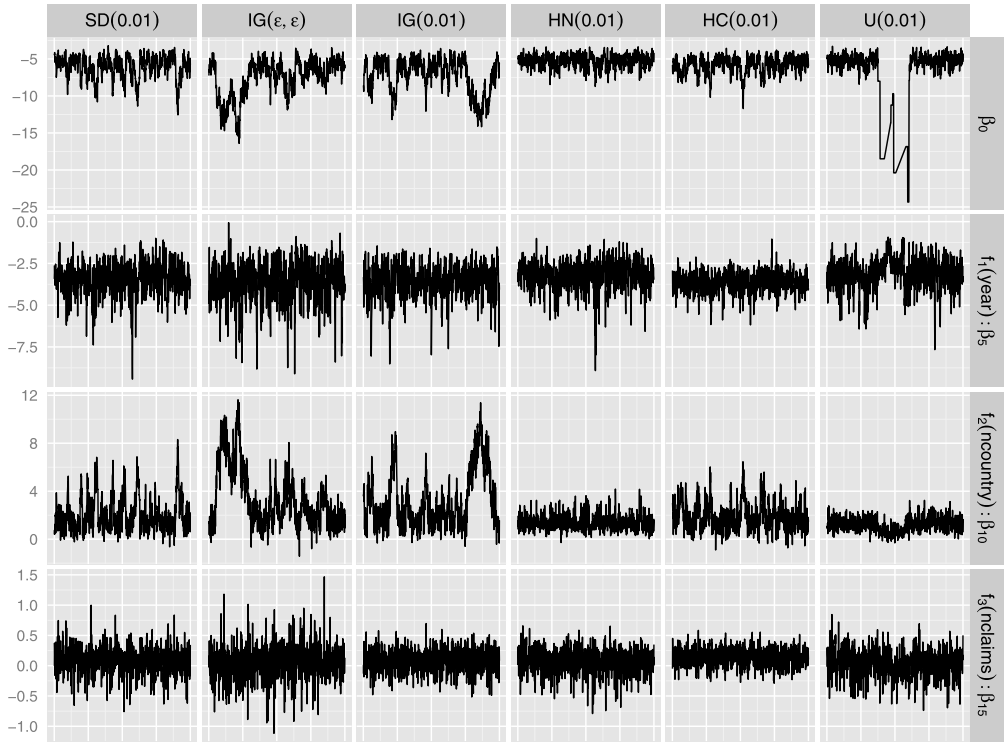

Figure 7: Patent citations. Sample paths (row wise) of the coefficient for the constant β_0 and one exemplary coefficient for the estimated effects of *year*, *ncountry* and *nclaims* each. Columns correspond to the different hyperpriors $\mathrm{SD}(\alpha), \mathrm{IG}(\alpha), \mathrm{HN}(\alpha), \mathrm{HC}(\alpha), \mathrm{U}(\alpha)$ with $\alpha = 0.01$, $c = 3$ and $\epsilon = 0.001$ for the inverse gamma prior $\mathrm{IG}(\epsilon, \epsilon)$. Results are based on 55000 MCMC iterations with burnin of 5000 and thinning parameter equal to 50.

6.2 Childhood Undernutrition in Zambia

As a second illustration, we use observations on malnutrition of 4421 children from Zambia in the year 1992. The data have been collected as part of nationally representative demographic and health surveys which are freely available at `www.measuredhs.com` and are described in more detail, e.g. in Fahrmeir et al. (2013).

Childhood undernutrition is usually determined by a Z-score

$$zscore_i = \frac{h_i - m}{s}$$

reflecting the nutritional status of child i with height h_i in the population of interest. The values m and s correspond to the mean height of children and their standard deviation in a suitable reference population of the same age group and gender.

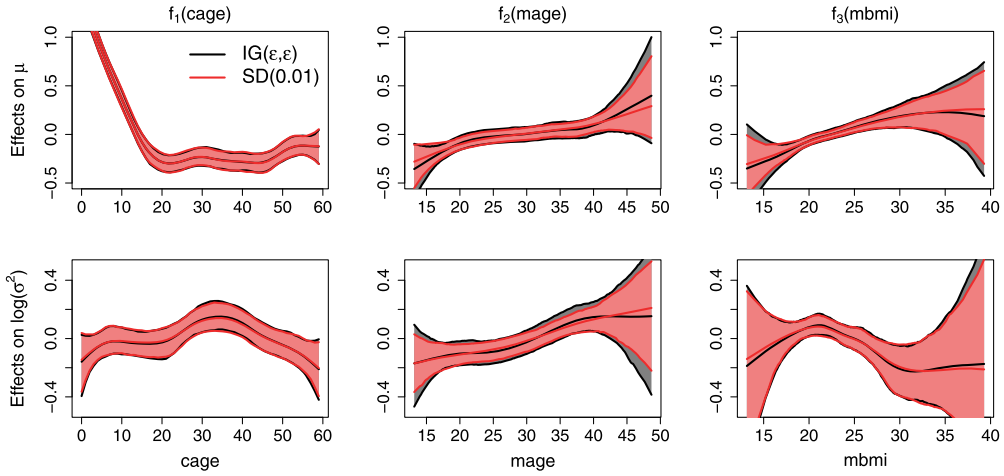

Figure 8: Childhood undernutrition. Comparison of estimated nonlinear effects of continuous covariates *cage*, *mage* and *mbmi* with scale-dependent priors, $\alpha = 0.01$, $c = 3$, (red) and inverse gamma priors $\text{IG}(\epsilon, \epsilon)$, $\epsilon = 0.001$, (black). Shown are posterior means and 95% credible intervals on $\mathbb{E}(zscore)$ (top) and on $\log(\sigma^2)$ (bottom).

For convenience, we assume a Gaussian location-scale model, that is, the Z-scores are conditionally normally distributed with

$$\begin{aligned} zscore_i &= \beta_0 + f_1(cage_i) + f_2(mage_i) + f_3(mbmi_i) + f_{spat}(district_i) \\ &\quad + \varepsilon_i,\ \varepsilon_i \sim \text{N}(0, \sigma_i^2), \\ \log(\sigma_i^2) &= \tilde{\beta}_0 + \tilde{f}_1(cage_i) + \tilde{f}_2(mage_i) + \tilde{f}_3(mbmi_i) + \tilde{f}_{spat}(district_i). \end{aligned}$$

In the equations above, f_1 to f_3 ($\tilde{f}_1$ to $\tilde{f}_3$) are smooth functions of the continuous covariates *cage* (child's age), *mage* (mother's age at birth) and *mbmi* (mother's body mass index), while f_{spat} ($\tilde{f}_{spat}$) represents the spatial effect that was assigned a Markov random field prior, and β_0 ($\tilde{\beta}_0$) is the usual overall intercept. For all nonparametric effects, we use $\text{IG}(\epsilon, \epsilon)$ priors with $\epsilon = 0.001$ and compare the results to scale-dependent priors ($\alpha = 0.01$, $c = 3$, $P = 1$ for splines and $P = n$ for spatial effects) in Figures 8 and 9. Comparisons with the other alternative priors discussed in Section 2.4 are shown in Supplement I. In addition, Figure I36 shows spatial effects of all other priors against the ones of scale-dependent priors per region.

The estimated effects of *mage* on the conditional expectation $\mathbb{E}(zscore)$ and on $\log(\sigma^2)$ as well as the effect of *mbmi* on $\mathbb{E}(zscore)$ are in accordance with the construction principles of the scale-dependent prior: Being close to a linear effect, this model is preferred by the scale-dependent prior and allows for narrower credible intervals. The effects of *cage* on both parameters are clearly nonlinear and both hyperpriors deliver very similar estimated curves. Comparable results can be observed for the spatial effect which is reasonable since the variable *district* has significant impact on both distribution parameters (based on a 95% credible interval). The effect of *mbmi* on $\log(\sigma^2)$ turns out

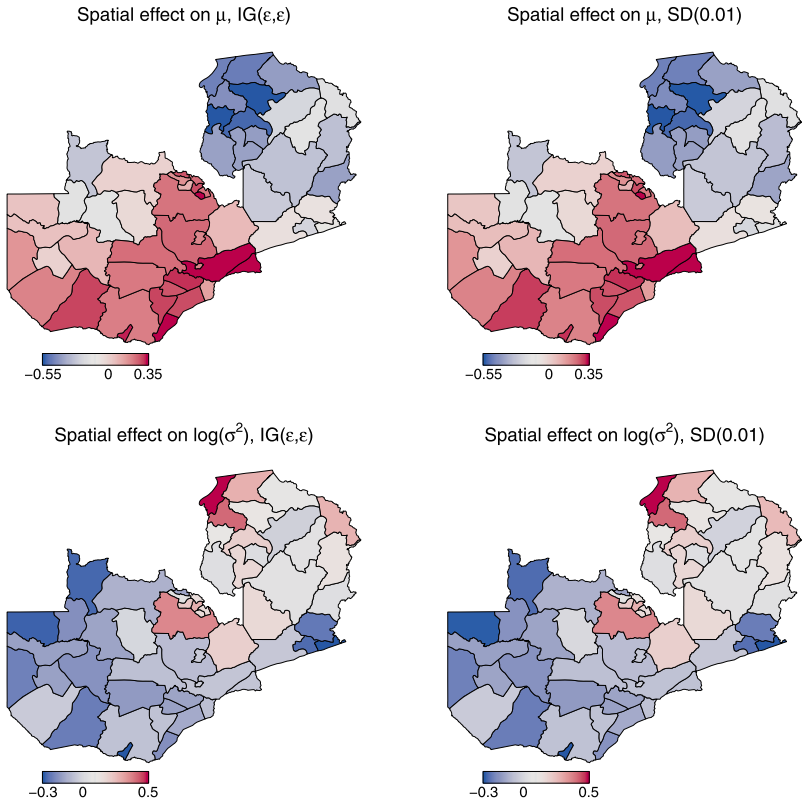

Figure 9: Childhood undernutrition. Comparison of estimated spatial effects with inverse gamma priors IG(ϵ, ϵ), $\epsilon = 0.001$ (left) and scale-dependent priors, $\alpha = 0.01$, $c = 3$ (right). Shown are posterior estimates on $\mathbb{E}(zscore)$ (top) and on $\log(\sigma^2)$ (bottom).

to be insignificant. Differences to the other hyperpriors are rather small, see Figures I32 to I35. One exception are the spatial effects with approximate uniform priors for τ, which are estimated on a smaller scale as compared to the remaining hyperpriors, see Figure I36.

7 Summary and Discussion

As Simpson et al. (2014) point out, 'priors are the Bayesian's greatest tool, but they are also the greatest point for criticism'. In the context of Bayesian inference in structured additive regression as it has originally been proposed in Fahrmeir and Lang (2001) or Fahrmeir et al. (2004) for exponential families, prior elicitation is in particular an issue when it comes to choosing hyperpriors for the smoothing variances. For convenience, basically all previous applications of structured additive regression focus on inverse gamma priors that are considered 'vaguely informative' but as Simpson et al. (2014)

show, these priors typically favour more complex specifications over the base model implied by zero variances.

To overcome this limitation, this paper makes a concrete suggestion for alternative hyperprior specifications in Bayesian structured additive regression relying on assumptions and principles that can be linked directly to the individual predictor components and that can easily be modified due to the principle of user-defined scaling. This notion of scale enhances interpretability and justification for hyperparameters (via the scale parameter θ) as it is related to the magnitude of the effects which are the quantities of interest. Based on Gaussian approximations of the full conditional distributions for $\log(\tau^2)$, we implemented an MCMC sampler as a part of a numerically stable and efficient implementation that provides an attractive option for applied researchers. The usage of MCMC is only meaningful when the posterior for parameters of interest has a proper density. We therefore derived sufficient and sometimes necessary conditions for the propriety and discussed the relevance of the assumptions in practice. Finally, both empirical studies and applications indicated that our proposed hyperprior is a promising alternative to the classical inverse gamma priors and also performs competitive with several alternative suggested prior specifications.

In general, the principle of user-defined scaling can of course be applied beyond the types of models considered in this paper. For example, the parameters of spike and slab priors for variable or function selection (George and Mc Culloch, 1993; Ishwaran and Rao, 2005; Scheipl et al., 2012) could also be based on analogous considerations for the scaling of effects under both the spike and the slab component. Note that, as discussed in Section 3.5 of Simpson et al. (2014), the penalised complexity prior itself does not make a good shrinkage prior due to the basic problem that the base model has been misspecified: The light tails of an exponential distribution will shrink large effects too strong towards zero. For spike and slab priors the situation is similar since employing the scale-dependent priors would worsen effect separability between spike and slab. The reason is the third principle of constant rate penalisation. This problem, when both priors have their mode at zero, is similar to the one of a double exponential distribution discussed in Frühwirth-Schnatter and Wagner (2011). The proper application of scale-dependent prior specifications in variable and function selection therefore promises to be a valuable field for future research.

Supplementary Material

Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression: Supplement (DOI: 10.1214/15-BA983SUPP; .pdf).

References

Bayarri, M. J. and García-Donato, G. (2008). "Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70: 981–1003. MR2530326. doi: http://dx.doi.org/10.1111/j.1467-9868.2008.00667.x. 1073

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. 1092

Berger, J. O. (2006). "The case for objective Bayesian analysis (with discussion)." *Bayesian Analysis* 1: 385–402. MR2221271. doi: http://dx.doi.org/10.1214/06-BA115. 1072

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). "The formal definition of reference priors." *The Annals of Statistics* 37: 905–938. MR2502655. doi: http://dx.doi.org/10.1214/07-AOS587. 1072

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 41: 113–147. MR0547240. 1072

Brezger, A. and Lang, S. (2006). "Generalized structured additive regression based on Bayesian P-splines." *Computational Statistics & Data Analysis* 50: 967–991. MR2210741. doi: http://dx.doi.org/10.1016/j.csda.2004.10.011. 1072, 1090

Eilers, P. H. and Marx, B. D. (1996). "Flexible smoothing using B-splines and penalized likelihood." *Statistical Science* 11: 89–121. MR1435485. doi: http://dx.doi.org/10.1214/ss/1038425655. 1078

Fahrmeir, L. and Kneib, T. (2009). "Propriety of posteriors in structured additive regression models: Theory and empirical evidence." *Journal of Statistical Planning and Inference* 139: 843–859. MR2479832. doi: http://dx.doi.org/10.1016/j.jspi.2008.05.036. 1072

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. 1071, 1075, 1076, 1101

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). *Regression – Models, Methods and Applications*. Springer, Berlin. MR3075546. doi: http://dx.doi.org/10.1007/978-3-642-34333-9. 1071, 1075, 1099

Fahrmeir, L. and Lang, S. (2001). "Bayesian semiparametric regression analysis of multicategorical time–space data." *Annals of the Institute of Statistical Mathematics* 53: 11–30. MR1820949. doi: http://dx.doi.org/10.1023/A:1017904118167. 1101

Frühwirth-Schnatter, S. and Wagner, H. (2010). "Stochastic model specification search for Gaussian and partial non-Gaussian state space models." *Journal of Econometrics* 154: 85–100. MR2558953. doi: http://dx.doi.org/10.1016/j.jeconom.2009.07.003. 1072

Frühwirth-Schnatter, S. and Wagner, H. (2011). "Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data." In: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (eds), *Bayesian Statistics 9*, Oxford, pp. 165–200. MR3204006. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0006. 1072, 1102

Gamerman, D. (1997). "Sampling from the posterior distribution in generalized linear mixed models." *Statistics and Computing* 7: 57–68. doi: http://dx.doi.org/10.1023/A:1018509429360. 1090

García-Donato, G. and Sun, D. (2007). "Objective priors for hypothesis testing in one-way random effects models." *The Canadian Journal of Statistics* 35: 303–320. MR2393611. doi: http://dx.doi.org/10.1002/cjs.5550350207. 1073

Gelman, A. (2005). "Analysis of variance: why it is more important than ever (with discussion)." *The Annals of Statistics* 33: 1–53. MR2157795. doi: http://dx.doi.org/10.1214/009053604000001048. 1072, 1074, 1080

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis* 1: 515–533. MR2221284. doi: http://dx.doi.org/10.1214/06-BA107A. 1072, 1074, 1080

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y. S. (2008). "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics* 2: 1360–1383. MR2655663. doi: http://dx.doi.org/10.1214/08-AOAS191. 1072

George, A. and Liu, J. W. (1981). *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall, Englewood Cliffs. MR0646786. 1092

George, E. and Mc Culloch, R. (1993). "Variable selection via Gibbs sampling,." *Journal of the American Statistical Association* 88: 881–889. doi: http://dx.doi.org/10.1080/01621459.1993.10476353. 1102

Ghosh, M. (2011). "Objective priors: An introduction for frequentists (with discussion)." *Statistical Science* 26: 187–202. MR2858380. doi: http://dx.doi.org/10.1214/10-STS338. 1072

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC, New York/Boca Raton. MR1082147. 1071

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57: 97–109. doi: http://dx.doi.org/10.1093/biomet/57.1.97. 1090

Hodges, J. S. (2013). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. Chapman & Hall/CRC. MR3289097. 1072, 1074, 1080

Ishwaran, H. and Rao, S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *Annals of Statistics* 33: 730–773. MR2163158. doi: http://dx.doi.org/10.1214/009053604000001147. 1102

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford. MR0187257. 1073

Jerak, A. and Wagner, S. (2006). "Modeling probabilities of patent oppositions in a Bayesian semiparametric regression framework." *Empirical Economics* 31: 513–533. doi: http://dx.doi.org/10.1007/s00181-005-0047-0. 1097

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: http://dx.doi.org/10.1111/1467-9876.00385. 1071

Klein, N. (2015). *sdPrior: Scale-Dependent Hyperpriors in Structured Additive Distributional Regression*. R package version 0.3. 1078

Klein, N. and Kneib, T. (2015). "Scale-dependent priors for variance parameters in structured additive distributional regression: Supplement." *Bayesian Analysis*. doi: http://dx.doi.org/10,124/15-BA983SUPP. 1074

Klein, N., Kneib, T., Klasen, S. and Lang, S. (2015). "Bayesian structured additive distributional regression for multivariate responses." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64: 569–591. MR3367789. doi: http://dx.doi.org/10.1111/rssc.12090. 1074

Klein, N., Kneib, T. and Lang, S. (2015). "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: http://dx.doi.org/10.1080/01621459.2014.912955. 1073, 1084, 1088, 1090, 1097

Klein, N., Kneib, T., Lang, S. and Sohn, A. (2015). "Bayesian structured additive distributional regression with an application to regional income inequality in germany." *Annals of Applied Statistics* 9: 1024–1052. MR3371346. doi: http://dx.doi.org/10.1214/15-AOAS823. 1073, 1074, 1090

Kneib, T., Hothorn, T. and Tutz, G. (2009). "Variable selection and model choice in geoadditive regression models." *Biometrics* 65: 626–634. MR2751488. doi: http://dx.doi.org/10.1111/j.1541-0420.2008.01112.x. 1071

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: http://dx.doi.org/10.1198/jasa.2010.tm09165. 1092

Lang, S. and Brezger, A. (2004). "Bayesian P-splines." *Journal of Computational and Graphical Statistics* 13: 183–212. MR2044877. doi: http://dx.doi.org/10.1198/1061860043010. 1078

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K. and Kneib, T. (2014). "Multilevel structured additive regression." *Statistics and Computing* 24: 223–238. MR3165550. doi: http://dx.doi.org/10.1007/s11222-012-9366-0. 1092

Lindgren, F., Rue, H. and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73: 423–498. MR2853727. doi: http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x. 1073

Polson, N. G. and Scott, J. G. (2012). "On the half-Cauchy prior for a global scale parameter." *Bayesian Analysis* 7: 887–902. MR3000018. doi: http://dx.doi.org/10.1214/12-BA730. 1072

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: http://dx.doi.org/10.1111/j.1467-9876.2005.00510.x. 1073

Rue, H. (2001). "Fast sampling of Gaussian Markov random fields with applications." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63: 325–338. MR1841418. doi: http://dx.doi.org/10.1111/1467-9868.00288. 1092

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman & Hall/CRC, New York/Boca Raton. MR2130347. doi: http://dx.doi.org/10.1201/9780203492024. 1072

Rue, H., Martino, S. and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71: 319–392. MR2649602. doi: http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x. 1073

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press. MR1998720. doi: http://dx.doi.org/10.1017/CBO9780511755453. 1076

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: http://dx.doi.org/10.1080/01621459.2012.737742. 1102

Simpson, D., Rue, H. Martins, T. G., Riebler, A. and Sørbye, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors, arXiv:1403.4630, Norwegian University of Sciences and Technology, Trondheim, Norway, submitted to Statistical Science. 1072, 1073, 1076, 1077, 1078, 1082, 1101, 1102

Sørbye, S. H. and Rue, H. (2014). "Scaling intrinsic Gaussian Markov random field priors in spatial modelling." *Spatial Statistics* 8: 39–51. MR3326820. doi: http://dx.doi.org/10.1016/j.spasta.2013.06.004. 1080

Sun, D., Tsutakawa, R. K. and He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models." *Statistica Sinica* 11: 77–95. MR1820002. 1084

Wand, M. P. (2000). "A comparison of regression spline smoothing procedures." *Computational Statistics* 15: 443–462. MR1818029. doi: http://dx.doi.org/10.1007/s001800000047. 1076

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC, New York/Boca Raton. MR2206355. 1071, 1075

Acknowledgments

We thank one referee, the associate editor and the editor for their careful review which was very helpful in improving upon the initial submission. Financial support by the German Research Foundation via the research training group 1644 and the projects KN 922/4-1/2 is gratefully acknowledged.