Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and Overdispersed Count Data

Nadja KLEIN, Thomas KNEIB, and Stefan LANG

Frequent problems in applied research preventing the application of the classical Poisson log-linear model for analyzing count data include overdispersion, an excess of zeros compared to the Poisson distribution, correlated responses, as well as complex predictor structures comprising nonlinear effects of continuous covariates, interactions or spatial effects. We propose a general class of Bayesian generalized additive models for zero-inflated and overdispersed count data within the framework of generalized additive models for location, scale, and shape where semiparametric predictors can be specified for several parameters of a count data distribution. As standard options for applied work we consider the zero-inflated Poisson, the negative binomial and the zero-inflated negative binomial distribution. The additive predictor specifications rely on basis function approximations for the different types of effects in combination with Gaussian smoothness priors. We develop Bayesian inference based on Markov chain Monte Carlo simulation techniques where suitable proposal densities are constructed based on iteratively weighted least squares approximations to the full conditionals. To ensure practicability of the inference, we consider theoretical properties like the involved question whether the joint posterior is proper. The proposed approach is evaluated in simulation studies and applied to count data arising from patent citations and claim frequencies in car insurances. For the comparison of models with respect to the distribution, we consider quantile residuals as an effective graphical device and scoring rules that allow us to quantify the predictive ability of the models. The deviance information criterion is used to select appropriate predictor specifications once a response distribution has been chosen. Supplementary materials for this article are available online.

KEY WORDS: Iteratively weighted least squares; Markov chain Monte Carlo; Penalized splines; Zero-inflated negative binomial; Zero-inflated Poisson.

1. INTRODUCTION

For analyzing count data responses with regression models, the log-linear Poisson model embedded in the exponential family regression framework provided by generalized linear or generalized additive models is still the standard approach. In many applied examples, however, we face one or several of the following problems:

- An excess of zeros as compared to the number of zeros expected from the corresponding Poisson fit. For example, in an application on citations of patents considered later in this article, a large fraction of patents is never cited and this fraction seems to be considerably larger than expected with a Poisson distribution.
- Overdispersion, in which case the assumption of equal expectation and variance inherent in the Poisson distribution has to be replaced by variances exceeding the expectation. While it is common practice to introduce a single, scalar overdispersion parameter to inflate the expectation (Fahrmeir and Tutz, 2001), more complex forms of overdispersion where the amount of overdispersion depends on covariates and varies over the observations are often more adequate.
- A simple linear predictor is not sufficient to capture all covariate effects. For example, the number of claims arising in car insurance for a policyholder requires both spatial effects to capture the strong underlying spatial correlation and flexible nonlinear effects to model the effects of age of the car and age of the policyholder. Further extensions may be required to include complex interaction effects or random effects in case of grouped or multilevel data.

To overcome these problems, a number of extended count data regression variants have been developed. To deal with an excess of zeros, zero-inflated count data regression models assume that the data are generated by a two-stage process in which a binary process decides between observations that are always zero and observations that will, therefore, never be cited from those that are relevant and will, therefore, never be cited from those that are relevant and for which the number of citations follows, for example, a Poisson distribution. Both the probability for the binary decision and the Poisson rate may then be characterized in terms of covariates.

To deal with overdispersion, the negative binomial distribution provides a convenient framework extending the Poisson distribution by a second parameter determining the scale of the distribution; see, for example, Hilbe (2007). The negative binomial distribution can also be combined with zero-inflation.
as described in the previous paragraph; see, among others, Winkelmann (2008).

For Poisson regression and negative binomial regression with fixed scale parameter and no overdispersion, generalized additive models as developed in Hastie and Tibshirani (1990) and popularized by Wood (2006) provide a convenient framework to overcome the linearity assumptions of generalized linear models when smooth effects of continuous covariates shall be combined in an additive predictor. Inference can then be based on optimizing a generalized cross-validation criterion (Wood 2004), a mixed model representation (Ruppert, Wand, and Carroll 2003; Fahrmeir, Kneib, and Lang 2004; Wood 2008) or Markov chain Monte Carlo (MCMC) simulations (Brezger and Lang 2006; Jullion and Lambert 2007). The framework of generalized additive models for location, scale, and shape (GAMLSS) introduced by Rigby and Stasinopoulos (2005), extends generalized additive models to more complex response distributions where not only the expectation but multiple parameters are related to additive predictors via suitable link functions. In particular, zero-inflated Poisson and zero-inflated negative binomial responses can be embedded in this framework where for the former both the probability of excess zeros and the Poisson rate and for the latter the probability of excess zeros, the expectation of the count process and the scale parameter are related to regression predictors.

Predictor specifications that go beyond the generalized additive models of Hastie and Tibshirani (1990) comprising only nonlinear effects of continuous covariates have been developed within the framework of structured additive regression and allow arbitrary combinations of parametric linear effects, smooth nonlinear effects of continuous covariates, interaction effects based on varying coefficient terms or interaction surfaces, random effects, and spatial effects using either coordinate information or regional data (Fahrmeir, Kneib, and Lang 2004; Brezger and Lang 2006). Structured additive regression relies on a unifying representation of all these model terms based on nonstandard basis function specifications in combination with quadratic penalties (in a frequentist formulation) or Gaussian priors (in a Bayesian approach).

In this article, we develop Bayesian structured additive regression models for zero-inflated and overdispersed count data covering the following unique features:

- The approach supports the full flexibility of structured additive regression for specifying additive predictors for all parameters of the response distribution including the success probability of the binary process and the scale parameter of the negative binomial distribution.
- The model formulation and inference are embedded in the general framework of GAMLSS, which allows us to develop a generic approach for constructing proposal densities in an MCMC simulation algorithm based on iteratively weighted least squares approximations to the full conditions as suggested by Gamerman (1997) or Brezger and Lang (2006) for exponential family regression models.
- We provide a numerically efficient implementation including also an extension to multilevel structure that is particularly useful in spatial regression specifications or for models including random effects; see Lang et al. (2014).
- This implementation is part of the free software package BayesX (Belitz et al., 2012).
- Theoretical results on the propriety of the posterior and positive definiteness of the working weights required in the proposal densities are included.
- Especially compared to frequentist GAMLSS formulations, our approach has the advantage to include the choice of smoothing parameters directly in the estimation run and to provide valid credible intervals which are difficult to obtain based on asymptotic maximum likelihood theory.

Model choice between different types of zero-inflated and overdispersed count data models will be approached based on quantile residuals (Dunn and Smyth 1996) to evaluate the fit and proper scoring rules (Gneiting and Raftery 2007) to determine the predictive ability. The deviance information criterion (Spiegelhalter et al. 2002) will be considered to select important covariates for the different predictors once a response distribution has been fixed.

Some rare approaches developing similar types of models and inferences are already available. For example, Fahrmeir and Osuna Echavarría (2006) developed a Bayesian approach for zero-inflated count data regression with Poisson or negative binomial responses but only allow for covariate effects on the expectation of the count data part of the response distribution and not on the probability of excess zeros or the scale parameter of the negative binomial distribution. Czado et al. (2007) also developed zero-inflated generalized Poisson regression models for count data where the overdispersion and zero-inflation parameters can be fitted by maximum likelihood methods but only considered linear predictors. He, Xue, and Shi (2010) studied the length of sick leaves in Indonesia in 1997 in a zero-inflated Poisson model that allows for the inclusion of one piecewise linear effect in both the expectation of the Poisson distribution and the probability of excess zeros. Their model, relying on piecewise linear functions with four and three knots, respectively, can be reproduced with our more flexible model. Doing so, we obtain similar results but the DIC would be in favor of a model including both the age and the income nonlinearity which would not have been possible in the model by He, Xue, and Shi (2010).

There are two packages in R that provide regression for zero-inflated models. In gam1as (Rigby and Stasinopoulos 2005) maximum (penalized) likelihood inference is used to fit models within the GAMLSS framework including the zero-inflated Poisson and (zero-inflated) negative binomial distribution. A description about the implementation of GAMLSS in R and data examples was given in Stasinopoulos and Rigby (2007). We will evaluate the comparison of the proposed Bayesian approach for zero-inflated and overdispersed count data with the penalized likelihood approach in gam1as in extensive simulations in Section 4. Linear predictors for the count and zero-inflation part can be specified in the package pscl (Zeileis, Kleiber, and Jackman 2008), while the scale parameter of the negative binomial distribution is always assumed constant. Parameters are estimated with the function optim to maximize the likelihood. We will not directly compare our results with those from pscl but will consider model specifications with the same restrictions as competitors in our applications.
The rest of this article is organized as follows: Section 2 describes the model specification for Bayesian zero-inflated and overdispersed count data regression in detail including prior specifications. Section 3 develops the corresponding MCMC simulation algorithm based on iteratively weighted least squares proposals and discusses theoretical results. Section 4 evaluates the performance of the Bayesian approach compared to the penalized likelihood approach of GAMLS in within a restricted class of purely additive models and for more complex geodative models. Sections 5 and 6 provide analyses of the applications on citations of patents and claim frequencies in car insurance. The final Section 7 will summarize our findings and comments on directions of future research.

2. ZERO-INFLATED AND OVERDISPERSSED COUNT DATA REGRESSION

2.1 Observation Models

We assume that zero-inflated count data \( y_i \) as well as covariate information \( v_i \) have been collected for individuals \( i = 1, \ldots, n \). For given covariates \( v_i \), the responses \( y_i \) are conditionally independent with (conditional) density \( p(y_i|v_i) = \pi_i \mathbb{1}_{0}(y_i) + (1 - \pi_i)\tilde{p}(y_i|v_i) \). This distribution arises from the hierarchical definition of the responses as the product of two independent variables \( y_i = \kappa_i \tilde{y}_i \), where \( \kappa_i \) is a binary selection process \( \kappa_i \sim \text{Bern}(1 - \pi_i) \) and \( \tilde{y}_i \) follows one of the standard count data models, \( \tilde{y}_i \sim \tilde{p} \) such as a Poisson distribution or a negative binomial distribution. The underlying reasoning is as follows: To model the excess of zeros observed in zero-inflated count data, the response is zero if \( \tilde{y}_i \) equals zero while additional zeros arise whenever the indicator variable \( \kappa_i \) is zero. The amount of extra zeros introduced compared to the standard count data distribution of \( \tilde{y}_i \) is determined by the probability \( \pi_i \). From the definition of zero-inflated count data models, we obtain

\[
\mathbb{E}(y_i|v_i) = (1 - \pi_i)\mathbb{E}(\tilde{y}_i|v_i)
\]

and

\[
\text{Var}(y_i|v_i) = (1 - \pi_i)\text{Var}(\tilde{y}_i|v_i) + \pi_i(1 - \pi_i)(\mathbb{E}(\tilde{y}_i|v_i))^2.
\]

Our focus is on two special cases for the count data part of the distribution, namely the Poisson distribution \( \tilde{y}_i \sim \text{Po}(\lambda_i) \) with density \( \tilde{p}(y_i) = \lambda_i^y e^{-\lambda_i} / y! \) and the negative binomial distribution \( \tilde{y}_i \sim \text{NB}(\delta_i, \delta_i / (\delta_i + \mu_i)) \) with density

\[
\tilde{p}(y_i) = \Gamma(y_i + \delta_i) / (\Gamma(y_i + 1)\Gamma(\delta_i)) \delta_i^{-\delta_i} (\mu_i)^\delta_i (\delta_i + \mu_i)^{y_i}. 
\]

The latter choice is particularly suited when the count data part of the response distribution is overdispersed.

To allow maximum flexibility in the zero-inflated count data regression specifications, both the parameter for the excess of zeros and the parameters of the count data part of the distribution are related to regression predictors constructed from covariates via suitable link functions. For zero-inflated Poisson (ZIP) regression, we choose \( \eta_i^\pi = \text{logit}(\pi_i) \) and \( \eta_i^\lambda = \log(\lambda_i) \) whereas for zero-inflated negative binomial (ZINB) regression we assume \( \eta_i^\pi = \text{logit}(\pi_i), \eta_i^\delta = \log(\mu_i) \) and \( \eta_i^\delta = \log(\delta_i) \). Both specifications can be embedded in the general class of generalized additive models for location, scale, and shape proposed by Rigby and Stasinopoulos (2005). Poisson and negative binomial (NB) regression are obtained as special instances (\( \tau = 0 \)) with one and two parameters \( \lambda, \mu, \delta \), respectively. Note that in applications we may often observe that modeling either zero-inflation or overdispersion is sufficient to adequately represent the data generating mechanism. In particular, a large fraction of observed zeros can also be related to overdispersion and it is, therefore, not generally useful to consider the most complex model type for routine applications. In Sections 5 and 6 we will further comment on this issue and will also provide ways of comparing different models for zero-inflated and overdispersed count data.

2.2 Semiparametric Predictors

For each of the predictors, we assume a structured additive specification

\[
\eta_i = \beta_0 + f_1(v_i) + \cdots + f_J(v_i),
\]

where, for notational simplicity, we drop the parameter index from the predictor and the included effects. While \( \beta_0 \) is an intercept term representing the overall level of the predictor, the generic functions \( f_j(v_i), \ j = 1, \ldots, J \), relate to different types of regression effects combined in an additive fashion. In structured additive regression, each function is approximated in terms of \( D_j \) basis functions such that

\[
f_j(v_i) = \sum_{d_j=1}^{D_j} \beta_{j,d} B_{j,d}(v_i).
\]

For example, for nonlinear effects of continuous covariates, the basis functions may be B-spline bases while for spatial effects based on coordinates, the basis functions may be radial basis functions or kernels. We will give some more details on special cases later on in this section.

The basis function approximation (1) implies that each vector of function evaluations \( f = (f_1(v_1), \ldots, f_J(v_n)) \) can be written as \( Z_j \beta_j \), where \( Z_j \) is the design matrix arising from the evaluations of the basis functions, that is, \( Z_j_{ij} = B_{j,d}(v_i) \), and \( \beta_j = (\beta_{j,1}, \ldots, \beta_{j,D_j}) \) is the vector of all regression coefficients. Then the predictor vector \( \eta = (\eta_1, \ldots, \eta_n) \) can compactly be represented as

\[
\eta = \beta_0 1 + Z_1 \beta_1 + \cdots + Z_J \beta_J,
\]

where 1 is an \( n \)-dimensional vector of ones.

2.3 Prior Specifications

To enforce specific smoothness properties of the function estimates arising from the basis function approximation (1), we consider multivariate Gaussian priors

\[
p(\beta_j|\tau_j^2) \propto \left( \frac{1}{\tau_j^2} \right)^{\frac{d}{2}} \exp \left( -\frac{1}{2\tau_j^2} \beta_j^T K_j \beta_j \right)
\]

for the regression coefficients where \( \tau_j^2 \) is the smoothing variance determining our prior confidence and \( K_j \) is the prior precision matrix implementing prior assumptions about smoothness of the function. Note that \( K_j \) may not have full rank and, therefore, the Gaussian prior will usually be partially improper yielding a singular normal prior. A completely improper prior is obtained as a special case for either \( \tau_j^2 \to \infty \) or \( K_j = 0 \).
Several prior choices are available for the variance parameters $\tau_j^2$. A standard choice is inverse gamma priors $\tau_j^2 \sim \text{IG}(a_j, b_j)$ usually with small values for the hyperparameters $a_j$ and $b_j$ (our default option is $\tau_j^2 \sim \text{IG}(0.001, 0.001)$). Other proposals are uniform priors for $\tau_j^2$ or the standard deviation $\tau_j$ obtained as special cases of inverse gamma priors with $a_j = -1$, $b_j = 0$ and $a_j = -0.5$, $b_j = 0$. Half-Cauchy priors have recently been advocated by Polson and Scott (2012); see also Gelman (2006) for a general discussion of prior specifications in Bayesian hierarchical models. We will discuss sensitivity with respect to these different prior choices later in our additive and geadditive simulations (Section 4) and the applications (Sections 5 and 6).

2.4 Special Cases

To make the generic model specification introduced in the previous section more concrete, we compactly summarize some special cases:

- Linear effects $f_j(v_i) = x_i'\beta_j$, where $x_i$ is a subvector of original covariates: The design matrix is obtained by stacking the rows $x_i$ and usually a noninformative prior with $K_j = 0$ is chosen for the regression coefficients $\beta_j$. A ridge-type prior with $K_j = I$ is an alternative, especially if the dimension of the vector $\beta_j$ is large.
- P-splines for nonlinear effects $f_j(v_i) = f_j(x_i)$ of a single continuous covariate $x_i$: The design matrix comprises evaluations of B-spline basis functions defined upon an equidistant grid of knots and a given degree. The precision matrix is given by $K_j = D'D$, where $D$ is a difference of appropriate order. Usual default choices are 20 inner knots, cubic B-splines and second-order differences; see Lang and Brezger (2004) for details.
- Markov random fields $f_j(v_i) = f_j(s_i)$ for a discrete spatial variable $s_i \in \{1, \ldots, S\}$: The design matrix is an indicator matrix connecting individual observations with corresponding regions, that is, $Z[i, s]$ is one if observation $i$ belongs to region $s$ and zero otherwise. To implement spatial smoothness, $K_j$ is chosen as an adjacency matrix indicating which regions are neighbors; see Rue and Held (2005) for details.
- Random effects $f_j(v_i) = \beta_{s_i}$ based on a grouping variable $g_i \in \{1, \ldots, G\}$: The design matrix is an indicator matrix connecting individual observations with corresponding groups, that is, $Z[i, g]$ is one if observation $i$ belongs to group $g$ and zero otherwise. The precision matrix is chosen as $K_j = I$.

A more detailed exposition for structured additive regression specifications comprising also bivariate surfaces or varying coefficient terms is provided in Fahrmeir, Kneib, and Lang (2004).

3. INference

Our Bayesian approach to zero-inflated and overdispersed count data regression relies on MCMC simulation techniques. For both the ZIP and ZINB model, the full conditionals for the regression coefficients arising from the basis function expansion are not analytically accessible due to the complex structure of the likelihoods. The same remains true for the Poisson and NB model. One possibility is to develop suitable proposal densities based on iteratively weighted least squares (IWLS) approximations to the full conditionals as detailed below. Note that in contrast, the full conditionals for the smoothing variances $\tau_j^2$ can be derived in closed form

$$
\tau_j^2 \sim \text{IG}(a_j', b_j'), \quad a_j' = \frac{\text{rk}(K_j)}{2} + a_j, \quad b_j' = \frac{1}{2}b_j' K_j \beta_j + b_j.
$$

3.1 IWLS Proposals

The basic idea of IWLS proposals is to determine an approximation of the full conditional that leads to a Gaussian proposal density with expectation and covariance matrix corresponding to the mode and the curvature at the mode at the current state of the Markov chain. The resulting Metropolis–Hastings sampler described in Section 3.2 has the advantage that the IWLS proposal automatically adapts to the form of the full conditional and thereby avoids manual tuning. To make the description easier, we assume for the moment a model with only one predictor $v$ but the principle idea immediately carries over to our multipredictor framework since in the MCMC algorithm we are always working with subblocks of coefficients corresponding to one predictor component. Let $l(\eta)$ be the log-likelihood depending on the predictor $v$. It is then easy to verify that the full conditional for a parameter block $\beta_j$ is

$$
\log(p(\beta_j | \cdot)) \propto l(\eta) - \frac{1}{2\tau_j^2} \beta_j' K_j \beta_j + \alpha \log(\tau_j^2),
$$

and $\alpha$ is absorbed for equality up to additive constants. Taylor expansion leads to the Newtons method type approximation

$$
\frac{\partial l(t)}{\partial \eta_i} - \frac{\partial^2 l(t)}{\partial \eta_i^2} \cdot (\eta_i^{(t+1)} - \eta_i^{(t)}) = 0
$$

with $t$ an index of the current iteration. From this, we can deduce the working model

$$
z(t) \sim N(\eta(t), (W(t))^{-1})
$$

in which $z = \eta + W^{-1} v$ is a vector of working observations with the predictor as expectation, $v = \partial l / \partial \eta$ is the score vector and $W$ is the working weight matrix based on a Fisher-scoring approximation, with $w_i = \mathbb{E}(-\partial^2 l / \partial \eta_i^2)$ on the diagonal and zero otherwise. With the priors (3) on $\beta_j$ we finally obtain the Gaussian proposal densities $N(\mu_j, P_j^{-1})$ for $\beta_j$ with expectation and precision matrix

$$
\mu_j = P_j^{-1} Z_j' W(z - \eta_{\cdot j}), \quad P_j = Z_j' W Z_j + \frac{1}{\tau_j^2} K_j,
$$

where $\eta_{\cdot j} = \eta - Z_j \beta_j$ is the predictor without the $j$th component.

To be able to apply the IWLS proposals in the context of zero-inflated count data regression, we now have to derive the required quantities, namely the score vectors $v$ and the working weights $W$. For the ZIP model, the elements of the score vectors for the zero-inflation and the Poisson parts of the model are given by

$$
v_i^z = \frac{\pi_i \lambda_i}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} 1_{(0)}(y_i) + (y_i - \lambda_i)
$$

$$
v_i^\pi = \frac{\pi_i}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} 1_{(0)}(y_i) - \pi_i
$$

$$
\lambda_i = \eta_i + Z_i' \eta_i, \quad w_i = \mathbb{E}(-\partial^2 l / \partial \eta_i^2).
$$



and the working weights can be shown to be
\[
w^{i}_{t'} = \frac{\lambda_i}{\pi_i}(1 - \pi_i) \left( \frac{1}{\pi_i} - \frac{1}{\pi_i} \right) \exp(-\lambda_i) \frac{1 - \pi_i}{\pi_i} \exp(-\lambda_i)
\]
\[
w^{i}_{t} = \frac{\pi_i^2(1 - \pi_i)}{\pi_i} \left( 1 - \exp(-\lambda_i) \right) \frac{1 - \pi_i}{\pi_i} \exp(-\lambda_i).
\]

For the ZINB model, we obtain
\[
u^{i}_{t'} = \frac{\pi_i \delta_i \mu_i}{\pi_i + (1 - \pi_i)} \left( \frac{\mu_i}{h_i + \mu_i} \right) \left( \mu_i + \delta_i \right)
\]
\[
u^{i}_{t} = \frac{\pi_i}{\pi_i + (1 - \pi_i)} \left( \frac{\mu_i}{h_i + \mu_i} \right) \left( \mu_i + \delta_i \right)
\]
\[
u^{i}_{t} = -\delta_i \left( \log \left( \frac{\mu_i}{h_i + \mu_i} \right) + \frac{\mu_i}{h_i + \mu_i} \right) \left( \frac{\delta_i}{\delta_i + \mu_i} \right) - \delta_i \left( \frac{\delta_i}{\delta_i + \mu_i} \right) \left( \frac{\delta_i}{\delta_i + \mu_i} \right) - \delta_i \left( \log \left( \frac{\mu_i}{h_i + \mu_i} \right) + \frac{\mu_i}{h_i + \mu_i} \right)^2 \left( \frac{\delta_i}{\delta_i + \mu_i} \right)
\]
\[
u^{i}_{t} = \left( \log \left( \frac{\mu_i}{h_i + \mu_i} \right) + \frac{\mu_i}{h_i + \mu_i} \right) \left( \frac{\delta_i}{\delta_i + \mu_i} \right) - \delta_i \left( \log \left( \frac{\mu_i}{h_i + \mu_i} \right) + \frac{\mu_i}{h_i + \mu_i} \right) \left( \frac{\delta_i}{\delta_i + \mu_i} \right)
\]
where \( \psi(x) = \frac{d}{dx} \log(\Gamma(x)) \) is the digamma function for \( x > 0 \), and

We choose \( m \) such that it is lower than or equal to the largest observed count and the cumulative sum \( \sum k p(k) \) of probabilities is above a certain threshold (our default is 0.999). The computing time is considerably dominated by the evaluation of the approximations above and it proved to work quite well in practice to calculate
\[
- \delta_i \left( \mathbb{E}(\psi(y_i + \delta_i)) - \psi(\delta_i) \right) - \delta_i^2 \left( \mathbb{E}(\psi(y_i + \delta_i)) - \psi(\delta_i) \right)
\]
only within the initialization period for computing the starting values (see Section 3.2 below). After that period, we keep expression (11) fixed during the MCMC iterations which reduces computing time by at least two thirds. Explicit derivations of all score vectors and working weight matrices are given in the supplement Section C.1. Further discussion and proofs for the positive definiteness of \( W \) are given in Section 3.3.

The required quantities in the Poisson and NB model can directly be obtained from those of the ZIP and ZINB model.

3.2 Metropolis–Hastings Algorithm

The resulting MCMC algorithm can compactly be summarized as follows:

1. Initialization: Let \( T \) be the number of iterations. Determine suitable starting values for all unknown parameters (e.g., using the algorithm described in Section A).
2. Loop over the iterations \( t = 1, \ldots, T \), the predictors of a given model and the components of the predictor.
   (a) Compute the working observations \( z^{(i)} = \eta^{(i)} + \left( W^{(i)} \right)^{-1} p^{(i)} \) with current values.
   (b) Update \( \beta^{(i)} \): Generate a proposal \( \beta^{(i)} \) from the density \( q(\beta^{(i)}, \beta^{(i)}) = N(\mu^{(i)}, \left( P^{(i)} \right)^{-1}) \) with expectation \( \mu^{(i)} \) and precision matrix \( P^{(i)} \) given in (5). To solve the identifiability problem inherent to additive models, the sampled effect is corrected according to Rue and Held (2005, Algorithm 2.6), such that \( A \beta^{(i)} = 0 \) holds, with appropriate matrix \( A \), such as \( A = 1' Z_j \).
   (c) Update of \( \tau^{2}_{j} \): Generate the new state from the inverse gamma distribution \( IG\left(a^{(i)}_{j}, b^{(i)}_{j}\right) \) with \( a^{(i)}_{j} \) and \( b^{(i)}_{j} \) given in (4).

By construction, the acceptance rates of the smoothing variances are 100% as the generation of random numbers is realized by a Gibbs-sampler. During several simulations and in the applications we observed acceptance rates between 70% and 90% for linear and nonlinear effects. Note that in contrast to random walk proposal Metropolis–Hastings samplers high acceptance rates are desired since they reflect situations in which the proposal density is very close to the true full conditional. In cases with high-dimensional parameter vectors, such as in spatial effects, the acceptance rates might be lower than 30%. An extension to multilevel structure can cover this problem and is explained in Section B of the online supplementary material.
3.3 Theoretical Results and Numerical Details

Propriety of the Posterior. Since our model specification includes several (partially) improper normal priors, a natural question is whether the resulting posterior is actually proper. For exponential family regression with similar predictor types, this question has been investigated, for example, in Fahrmeir and Kneib (2009) or Sun, Tsutakawa, and Zhuoqiong (2001) and we will now generalize these results to the framework of GAMLSS. Assume that conditionally independent observations \( y_i, i = 1, \ldots, n \), with densities \( p(y_i) \) corresponding to an \( m \)-parametric distribution family with parameters \( \theta_1, \ldots, \theta_m \) are given. Let \( \eta_{i1}^{\theta_1}, \ldots, \eta_{im}^{\theta_m} \) be the predictors linked to the \( m \) parameters. For a generic predictor vector, Equation (1) allows us to write \( y_i = \sum_{j=1}^{J} Z_{ij}^{\theta_j} = \sum_{j=1}^{J} Z_{ij}^{\theta_j} \beta_j \), with appropriate design matrices \( Z_j \) and vectors of regression coefficients \( \beta_j \). The basic idea to get sufficient conditions for the propriety of the posterior is to rewrite this model in a mixed model representation with iid individual specific random effects where we explicitly differentiate between effects with proper and (partially) improper priors. We now discuss the required assumptions in more detail:

A. Conditions on the densities. Assume that the set of observations can (after reordering) be partitioned such that for \( n^* \geq 1 \)

\[
\int \cdots \int p(y_i; \eta_{i1}^{\theta_1}, \ldots, \eta_{im}^{\theta_m}) \, d\eta_{i1}^{\theta_1} \cdots d\eta_{im}^{\theta_m} < \infty
\]

for \( i = 1, \ldots, n^* \).

\[
p(y_i; \eta_{i1}^{\theta_1}, \ldots, \eta_{im}^{\theta_m}) \leq M \text{ for } i = n^* + 1, \ldots, n.
\]

This implies that for at least one observation the density is integrable (with respect to the predictors) and that all remaining densities are bounded. Since we are dealing with discrete distributions, all densities are automatically bounded by 1 so that only Condition (A.1) can be an issue in practice. In our class of models, Condition (A.1) is usually fulfilled if certain restrictions apply on specific parameters that exclude extreme values on the boundary of the parameter space. 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 Poisson distributed responses, the density is integrable for all \( y_i > 0 \) such that \( n^* \) corresponds to the number of nonzero observations. For zero-inflated Poisson, we require at least one observation with \( y_i > 0 \) for which additionally the boundary case \( \pi_i = 0 \) is excluded. Note that excluding the boundary case simply implies that we exclude cases in which no information is available on the zero-inflation part of the distribution. For the negative binomial case, we similarly have to restrict the parameter space for the overdispersion parameter to exclude the boundary cases \( \delta_i = 0 \) and \( \delta_i = \infty \) together with \( y_i > 0 \) for at least one observation \( y_i \). Note that these boundary cases simply imply that one observation \( y_i \) exists for which the second moment exists and for which information on \( \delta_i \) is available. For zero-inflated negative binomial responses, the condition on the probability of excess zeros and the condition on the overdispersion parameter have to be combined. Summarizing, we only run into trouble if for all \( y_i > 0 \) we are in one of the boundary cases described before. Exploratory analyses based on simpler parametric models will usually provide helpful guidance on excluding such situations.

B. Reparameterization and conditions on ranks of design matrices. To explicitly differentiate between model components with proper and improper prior, we apply a mixed model representation to all vectors of regression coefficients with partially improper priors such that

\[
Z_j^{\theta_j} \beta_j = X_j^{\theta_j} \xi_j + V_j^{\theta_j} \delta_j,
\]

where \( p(\xi_j) \propto \text{const} \) is a vector of fixed effects with flat prior and \( b_j^{\theta_j} \mid (\tau_j^{\theta_j})^2 \sim N(0, (\tau_j^{\theta_j})^2 I_{k^j}) \) are iid Gaussian random effects with proper prior and dimension \( k_j^{\theta_j} = \text{rk}(K_j^{\theta_j}) \) (see Fahrmeir and Kneib 2009, for details). The complete predictor can then be written as

\[
y^{\theta_j} = X^{\theta_j} \xi^{\theta_j} + V^{\theta_j} b^{\theta_j} + V_0^{\theta_j} b_0^{\theta_j},
\]

where \( X^{\theta_j} \xi^{\theta_j} \) comprises all predictor components with a flat prior, \( V_0^{\theta_j} b_0^{\theta_j} \) corresponds to the random effect with the largest dimension (denoted as \( k_0^{\theta_j} \)), and \( V^{\theta_j} b^{\theta_j} \) is a composition of all remaining random effects.

Let \( \hat{k}_0 = \min\{k_0^1, \ldots, k_0^m\} \) and assume that we can choose \( \hat{k}_0 \) observations including at least one observation fulfilling (A.1) to define the submodel

\[
y_\alpha^{\theta_j} = X_\alpha^{\theta_j} \xi_\alpha^{\theta_j} + V_\alpha^{\theta_j} b_\alpha^{\theta_j} + V_0^{\theta_j} b_0^{\theta_j}
\]

(12)

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

(B.1) The design matrix \( X^{\theta_j} \) has full rank \( r^{\theta_j} \).

(B.2) \( \text{rk}(X^{\theta_j}, V^{\theta_j}) = \text{rk}(X^{\theta_j}, V^{\theta_j}) = r^{\theta_j} + r_0^{\theta_j} \).

(B.3) \( \text{rk}(V_0^{\theta_j}) = \hat{k}_0 \), that is, \( V_0^{\theta_j} \) is of full rank.

To ensure (B.1), superfluous columns arising from the reparameterization have to be deleted. In particular, duplicated constant columns representing the levels of the functions are deleted which correspond to the centering of functions that we included in the MCMC algorithm. Condition (B.2) indicates that the rank of the design matrices in the submodel is the same as in the complete model whereas (B.3) defines a similar restriction for the design matrix of the largest random effect arising from the mixed model representation.

C. Assumptions on the hyperparameters. Define the normalized submodel

\[
\tilde{y}_\alpha^{\theta_j} = \tilde{X}_\alpha^{\theta_j} \tilde{\xi}_\alpha^{\theta_j} + \tilde{V}_\alpha^{\theta_j} \tilde{b}_\alpha^{\theta_j} + \tilde{e}_\alpha^{\theta_j}
\]

(13)

that is obtained by multiplying (12) with \( M^{\theta_j} = (V_0^{\theta_j} V_0^{\theta_j})^{-1/2} \) such that \( \tilde{y}_\alpha^{\theta_j} = M^{\theta_j} y_\alpha^{\theta_j}, \tilde{X}_\alpha^{\theta_j} = M^{\theta_j} X_\alpha^{\theta_j}, \tilde{V}_\alpha^{\theta_j} = M^{\theta_j} V_\alpha^{\theta_j}, \) and \( \tilde{e}_\alpha^{\theta_j} \sim N(0, (\tau^{\theta_j}_0)^2 I_{k_0^{\theta_j}}) \) represents an iid random effect since \( V_0^{\theta_j} b_0^{\theta_j} \sim N(0, (\tau^{\theta_j}_0)^2 V_0^{\theta_j} V_0^{\theta_j}). \)

Define then the residual sum of squares for the normalized submodel as

\[
\text{SSE}_\alpha^{\theta_j} := (\tilde{y}_\alpha^{\theta_j} - \tilde{X}_\alpha^{\theta_j} \tilde{\xi}_\alpha^{\theta_j} - \tilde{V}_\alpha^{\theta_j} \tilde{b}_\alpha^{\theta_j}) (\tilde{y}_\alpha^{\theta_j} - \tilde{X}_\alpha^{\theta_j} \tilde{\xi}_\alpha^{\theta_j} - \tilde{V}_\alpha^{\theta_j} \tilde{b}_\alpha^{\theta_j}).
\]

(14)
We then assume for the hyperparameters $(a_j^{(0)}, b_j^{(0)})$ of the inverse gamma prior of the $j$th model term in the $l$th predictor:

\begin{align}
(\text{C.1}) & \quad k_j^{(0)} + 2b_j^{(0)} > \sum_{j=1}^{J} k_{j}^{(0)} - r_{j}, \quad j = 1, \ldots, J^{(0)}, \\
(\text{C.2}) & \quad k_0 - r^{(0)} + 2a_0^{(0)} + 2\sum_{j=1}^{J} (a_j^{(0)}) > 0, \quad (a_j^{(0)}) = \min(0, a_j^{(0)}), \quad j = 1, \ldots, J^{(0)}. \\
(\text{C.3}) & \quad a_j^{(0)} < b_j^{(0)} = 0 \text{ or } b_j^{(0)} > 0, \quad j = 1, \ldots, J^{(0)}. \\
(\text{C.4}) & \quad SSE_{j}^{(0)} + 2b_0^{(0)} > 0.
\end{align}

Condition (C.1) 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. The condition can be ensured by increasing the value for $a_j^{(0)}$ if necessary. Condition (C.2) restricts the number of effects with flat prior to be at most equal to the dimension of the largest random effects part in the model. Again, the condition can be ensured by increasing the hyperparameter values. Condition (C.3) requires that if $b_j^{(0)}$ is set to zero, the parameter $a_j^{(0)}$ has to be negative. This includes situations corresponding to flat priors for the random effects variance ($a_j^{(0)} = -1, b_j^{(0)} = 0$) or standard deviation ($a_j^{(0)} = -0.5, b_j^{(0)} = 0$) but excludes Jeffreys’ prior ($a_j^{(0)} = b_j^{(0)} = 0$). Finally, Condition (C.4) requires that either there is variation in the residual sum of squares in the normalized submodel (implying that not all effects are zero) or $b_0^{(0)} > 0$ holds. The latter requirement can always be ensured in practice but excludes flat priors for the random effects variances or standard deviations.

This allows us to adapt the sufficient conditions for the propriety of the posterior derived in Fahrmeir and Kneib (2009), yielding the following theorem:

**Theorem 1.** Consider a structured additive regression model within the GAMLSS framework and predictors (2). The Conditions (A.1), (A.2), (B.1)–(B.3), and (C.1)–(C.4) specified above are sufficient conditions for a proper posterior.

A proof for the theorem is contained in Section D. Note that Theorem 1 gives sufficient but not necessary conditions such that the posterior can in fact still be proper even if one or more conditions are violated. Compared to the usual exponential family case, the conditions on rank deficiencies have to apply separately for each predictor in the model, so that the total requirements are in general stronger than in the generalized additive model case.

**Regularity of the posterior precision matrix.** Concerning the IWLS proposals, one requirement is that the covariance matrix of the approximating Gaussian proposal density is positive definite and, therefore, invertible. Given full column rank of the design matrix, this is ensured when the diagonal elements of the working weight matrices are all positive. Positivity of the weights is always given for zero-inflated Poisson models, as shown in Section C.2. For zero-inflated negative binomial models, the weights involved in the updates for $\pi$ and $\mu$ are always positive (see again Section C.2) while this is not necessarily the case for the weights related to $\delta$. Note, however, that this is not too problematic since positive weights are a sufficient but not necessary condition for the precision matrix to be invertible. Moreover, we have empirically observed that negative weights only occur rarely and in extreme parameter constellations. If a computed weight is exceptionally negative, we set it to a small positive value in our implementation to avoid rank deficient precision matrices.

**Implementation.** The Bayesian zero-inflated and overdispersed count data approach developed in this article is implemented in the open source software BayesX (Belitz et al. 2012). We use methods for efficient storing of large datasets and sparse matrix algorithms for sampling from multivariate Gaussian distributions; see Lang et al. (2014) for details. This also has the advantage that the multilevel framework, briefly outlined in Section B of the supplementary materials, becomes accessible for zero-inflated and overdispersed count data regression.

To compute starting values that ensure rapid convergence toward the stationary distribution, we make use of a backfitting algorithm (Hastie and Tibshirani 1990) with fixed smoothing parameters; see Section A of the online supplementary material for further details.

A challenge in working with count data models is the numerical stability of the software. Suppose for instance that we estimate a (possibly complex) ZIP regression whereas the true model is a simple Poisson regression without zero-inflation. Then $\pi$ is actually zero and the estimated predictor $\eta^\pi$ corresponding to $\pi$ will tend to be rather small so that a software crash (e.g., due to overflow errors) is very likely. The problems get even worse for the ZINB model. A similar problem occurs for $\pi \to 1$, but this would be easier to track in applications since then all observations would be equal to zero. To avoid numerical problems, we included an optional “save estimate” in our software that prevents a software crash due to numerical instability. This is obtained by updating a vector of regression parameters $\beta_j$ only if the proposed new state $\beta_j^{(0)}$ of the Markov chain ensures that the predictor vector is within a certain prespecified range (e.g., $-10 \leq \eta^\pi \leq 10$). Otherwise, the current state of the chain is kept. In the majority of applications, a predictor outside limits will not occur or only in a very few number of iterations. If it occurs frequently, then of course the estimated results will not be fully valid but rather an indicator that the specified model is too complex for the data at hand.

### 4. SIMULATIONS

With the simulations conducted in this section, we aim at studying various properties of our proposed Bayesian approach for zero-inflated and overdispersed count data regression (denoted as MCMC in the following) and at comparing it with frequentist penalized likelihood estimation (denoted as ML in the following). The computations for the Bayesian approach have been carried out in BayesX. For penalized likelihood estimation, we considered the gamlss package in R (Stasinopoulos and Rigby 2007) but observed a number of limitations in its performance so that comparisons could not be done in all cases. We, therefore, structured the rest of this section as follows: Section 4.1 considers only additive models that consist of a combination of nonlinear effects of continuous covariates. We mainly present results for ZIP regression since we observed convergence problems of the Newton–Raphson/Fisher-scoring algorithm build in the gamlss package for about 10% of the simulation replications for the ZINB model. These problems could not even be
overcome with different smoothing parameter choices (based on \texttt{find.hyper}) for the function \texttt{pb}. We also tried the \texttt{ga} function within the \texttt{gamlss.add} package for our simulated data which caused even more convergence problems than the \texttt{gamlss} package. After the detailed comparison of frequentist and Bayesian inference for ZIP regression, we summarize results from a number of additional simulations addressing, for example, the performance of Bayesian inference in NB and ZINB regression, the impact of relatively small sample sizes on estimation quality, effect separation between the zero-inflation and the count data part of the model, and the prior sensitivity of the results. In Section 4.2, we then extend the additive models by a spatial effect comprising a structured part (modeled as a Markov random field) and an unstructured part (represented by an iid random effect). Although the \texttt{gamlss.add} package provides a possibility to fit models with spatial effects based on Markov random fields, it does not support the hierarchical model specification we employed in the simulations. We also briefly comment on the performance of the DIC for model selection in Sections 5 and 6. Considerably more details on many of the simulation results including geoadditive models are provided in Sections E and F of the online supplement to this article.

4.1 Additive Models

As building blocks for additive model specifications, we consider the functions

\[
\begin{align*}
f_1^1(x_1) &= f_1^{11}(x_1) = \log(x_1), \\
f_2^1(x_2) &= f_2^{11}(x_2) = 0.3x_2 \cos(x_2) \\
f_1^2(x_1) &= \sin(x_1), \\
f_2^2(x_2) &= -0.2x_2^2 \\
f_1^3(x_1) &= 0.1 \exp(0.5x_1), \\
f_2^3(x_2) &= -0.5 \arcsin(x_2)
\end{align*}
\]

depending on which of the count data models is used, that is, each predictor is composed of the sum of two nonlinear functions \(f_1\) and \(f_2\). The covariates \(x_1\) and \(x_2\) are obtained as iid samples from equidistant grids of step size 0.01 with \(x_1 \in [1, 6]\) and \(x_2 \in [-3, 3]\). We use the sample size \(n = 1000\) as starting point and simulate 250 replications. An average amount of about 50% and 46% of zeros is observed in the generated samples for ZIP and ZINB, respectively.

For both the Bayesian and the frequentist estimates we considered cubic P-splines based on a grid of 20 equidistant knots and second order random walk prior/difference penalty. In the frequentist setting, the smoothing parameters were estimated using the function \texttt{find.hyper} with starting value 3 for all parameters and with default settings for the remaining arguments. In the Bayesian approach, we considered inverse gamma priors with parameters \(a = b = 0.001\) as the default choice. The number of iteration steps \(T\) for each simulation run \(r\) in MCMC is set to 12,000 with a burn-in phase of 2000 iterations. We store and use every 10th iterate. In total, 90 unknown parameters are included in the ZIP model whereas 135 unknown parameters have to be estimated in case of ZINB.

To compare the results of ML and MCMC, we consider logarithmic mean squared errors (MSEs) evaluated at the observed covariate values (Figure 1), as well as empirical coverage rates of pointwise 95% credible intervals in Figure 2. Coverage rates of 80% credible intervals have also been computed but showed a similar qualitative behavior and are, therefore, omitted. Note that BayesX also provides simultaneous credible bands (following the suggestion in Krivobokova, Kneib, and Claeskens, 2010) which we will use later in our application to assist the selection of an appropriate predictor structure. Since no frequentist counterpart is available, we did not include simultaneous confidence bands in the simulation results. Figure E1 of the supplement shows posterior mean estimates obtained with MCMC and ML point estimates averaged over all 250 replications. Estimates corresponding to the 10% and 90% quantile of the MSEs (i.e., to 10% best and 10% worst estimates) are given in Figures E2 and E3 together with 95% credible intervals.

In summary, we obtain the following results:

1. Bias. Averaging all replications leads to satisfactory results for ML and MCMC with only slightly too smooth mean estimates in extreme areas of effects. At the boundaries of covariate domains, MCMC tends to fit the true functions better. Figure E3 indicates potential problems with the smoothing parameter choice of ML while the choice of hyperparameters based on MCMC seems to be more appropriate.
2. **MSE.** Figure 1 confirms that both methods deliver similar mean results. In general, the nonlinear functions with effects on λ seem to be easier to estimate than the ones impacting the probability of the additional zeros π. This can be seen in the smaller values of MSEs of \( f_1^\pi \) and \( f_2^\pi \) compared to the ones of \( f_1^\tau \) and \( f_2^\tau \).

3. **Coverage rates.** Figure 2 provides evidence that the Bayesian approach provides valid uncertainty intervals of effects which cannot be obtained based on the asymptotic theory of ML. Note that a corresponding warning is already given in the manual of Stasinopoulos, Rigby, and Akantziliotou (2008, p. 51). For MCMC, the 95% level of the posterior intervals is mostly maintained.

4. **Computing time.** On a PC with Intel(R) 3.40GHz processor, the computation time of one MCMC replication is roughly 23 sec. and is pretty stable over the simulation replications. For ML, the runtime varies much more over the replications and depends mainly on the choice of smoothing parameters. To omit warnings, we applied the `find.hyper` function which can take several minutes. Given the hyperparameters, ML requires between less than 2 and more than 15 sec.

In conclusion, bias and MSEs support that results obtained with MCMC are at least as reliable as those obtained with ML.

In addition, better coverage properties of the posterior intervals obtained with MCMC render our Bayesian approach a strong competitor to existing ML estimates.

The performance of our approach in the ZIP, NB, and ZINB regression was further investigated through simulation studies documented in detail in Section E.2 of the online supplement. These additional simulations yielded the following results:

1. **Performance for NB and ZINB regression.** Sections E.2.1 and E.2.2 confirm that for samples size \( n = 1000 \) results found in the simulation of the ZIP model carry over to the NB model and the more general and complex situation of ZINB models. In fact, there is no sign of a deteriorated performance of the Bayesian estimation approach, despite the additional complexity introduced by a third distributional parameter.

2. **Small sample sizes.** MSEs and effective MCMC sample sizes (Kaas et al., 2012) indicate that the study design chosen in Section 4.1 still yields reliable estimates in the ZIP model even if the sample size is reduced to \( n = 250 \). However, with two nonlinear effects in each parameter we find that a sample size of \( n = 500 \) in the NB model and \( n = 750 \) in the ZINB model is desirable to obtain reliable results, compare to Section on sample size.

3. **Separation of effects.** Since the zero-inflation and count data process both produce observations with \( y_i = 0 \) and the expectation of the response depends on both the probability of excess zeros and the expectation of the count data process, one might be worried about the separability of effects between these processes. In Section E.2.4 of the online supplement, we find that this separation is indeed possible based on appropriate rules for model selection discussed in detail in Section 5.

4. **Prior sensitivity.** To investigate how sensitive additive models for zero-inflated count data are with respect to the prior settings for the smoothing variances, we ran the additive ZIP model with different hyperparameters of the inverse gamma distribution. In addition to our standard choice \( a_j = b_j = 0.001 \), we tried \( a_j = b_j \in \{0.01, 0.0001, 0.00001, 0.000001\} \), as well as \( a_j = -0.5, b_j = 0 \) and \( a_j = -1, b_j = 0 \) corresponding to uniform priors on the standard deviations \( \tau_j \) and on the variances \( \tau_j^2 \), respectively. As another alternative, we considered half-Cauchy priors for the standard deviations \( \tau_j \). All these choices did not cause serious problems with mixing or convergence of the Markov chains. Moreover, posterior mean estimates of the splines on \( \lambda \) are hardly influenced by these choices while the smoothness of splines impacting \( \pi \) varies slightly (compare Section E.2.5 of the online supplement).

4.2 Geoadditive Models

In this section, the simulations for the ZIP, NB, and ZINB model have been extended by an additional spatial effect simulated as

\[
\begin{align*}
 f_{\text{spat}}^h(s) &= f_{\text{spat}}^\mu(s) = \sin(x_s^T x_s^T) + \varepsilon_s \\
 f_{\text{spat}}^\pi(s) &= \sin(x_s^T \cos(0.5 y_s^T)) + \varepsilon_s \\
 f_{\text{spat}}^\tau(s) &= 0.5 x_s^T y_s^T + \varepsilon_s
\end{align*}
\]

The structured part of the spatial effect \( f_{\text{spat}} \) is assigned a Markov random field prior and simulated on the basis of centroids with standardized coordinates \((x_s^T, y_s^T)\), \( s \in \{1, \ldots, S\} \) of the \( S = 327 \) regions in Western Germany. The unstructured part is described by a random effect \( \varepsilon_s \sim N(0, 1/16) \) for each of the regions. In Figure E20 of the supplement, the simulated complete spatial effects for the rate \( \lambda \) of the count process and the probability of the additional zeros \( \pi \) in case of a ZIP model are visualized. The model for a generic predictor \( \eta \) can now be written as

\[ \eta = f_1 + f_2 + f_{\text{spat}} = Z_1 \beta_1 + Z_2 \beta_2 + Z_{\text{spat}} \beta_{\text{spat}}. \]

Estimates are obtained from a two-level model in which the total spatial effect is decomposed in a structured part and an unstructured effect \( \varepsilon \); see Section B of the online supplement and Lang et al. (2014).

Since the mixing of the Markov chains in a geoadditive model is in general less satisfactory than in additive models, the number of iterations is increased to 55,000 with a burn-in phase of 5000. We store each 50th iterate so that the final sample size of 1000 is retained. To find a desirable sample size for which satisfactory estimate results can be achieved, we performed estimates for \( n = 1000, 2000, \) and 4000 observations. Note that in the following we restrict ourselves to the presentation of results in the ZIP model. Outcomes for the NB and ZINB model are summarized at the end of this section and illustrated in Sections E.3.2 and E.3.3, respectively.

As shown in the supplement of Fahrmeir and Lang (2001), the unstructured and the structured spatial effect can generally not be separated and are often estimated with bias. Only the sum of both effects is estimated satisfactorily. This means in practice that only the complete spatial effect should be interpreted and nothing (or not much) can be said about the relative importance of both effects. Exceptions are cases where one of both effects (either the unstructured or the structured effect) is estimated
practically zero and the other effect clearly dominates. Hence, we present the estimated complete spatial effect compared to the true simulated effect for two selected sample sizes \( n = 1000 \) and 4000 in Figure 4. Beside that, the log(MSE) in Figure 3 and the kernel densities of complete posterior mean spatial effects in Figure E21 give further information about the quality of the inference.

Important results obtained from these simulations are:

1. **MSE.** The spatial effect has higher MSEs compared to the nonlinear effects but we observe that for greater sample sizes the MSEs can be reduced in all effects. Comparing Figure 1 with 3, it is positive to note that for sample size \( n = 1000 \) an additional spatial effect does not impair the MSEs of the nonlinear effects.

2. **Bias.** Figure 4 shows that extreme values of the spatial effect are the most difficult to estimate since both large negative and large positive effects tend to be underestimated. Together with Figure E21, it can be concluded that the spatial effect for \( \pi \) is estimated too smooth at the boundaries in comparison with the true effect.

3. **On a PC with Intel(R) 3.40GHz processor, the computing time for one replication is around 4 min.** We also assessed prior sensitivity as in the case of additive models presented in the previous section. Similar to additive models, nonlinear effects are quite robust even for very small values \( a_j, b_j \) without any convergence problems. Furthermore, under uniform priors or inverse gamma priors with \( a_j = b_j > 0.00001 \) we also did not observe serious mixing problems in the samples of the Markov random field. However, for smaller hyperparameters \( a_j = b_j = 0.00001 \) the number of iterations and the thinning has to be increased considerably, such that we do not recommend these specifications given the theoretical results on the propriety discussed in Section 3.3. On average, posterior mean spatial effects on \( \pi \) are worse under a half-Cauchy and \( \text{IG}(a_j = -1, b_j = 0) \) prior, compare Section E.4.

The simulations show that with a sample size of \( n = 4000 \) the estimated complete spatial effect is similar to the simulated, true one. The quality of mean estimates of nonlinear effects remains as in the previous section even when adding an additional spatial effect. Hence, it can be said that both nonlinear and spatial effects are well identified in the estimates, especially when taking the complexity of the models into account. Similar basic outcomes are obtained for the NB and ZINB models (documented in the supplement). For the performance of the deviance information criterion (DIC, Spiegelhalter et al. 2002) in geoadditive models to select relevant effects we refer the reader to a study documented in the online supplement Section F. There we also provide further details about how estimates suffer due to model misspecification.

5. **PATENT CITATIONS**

In our first application, we analyze the number of citations of patents granted by the European Patent Office (EPO). An inventor who applies for a patent has to cite all related, already existing patents his patent is based on. The data have originally been collected to study the occurrence of objections against patents for 4866 patents; see Jerak and Wagner (2006). Details about the dataset including summary statistics and a discussion about outlier removal can be found in (Fahrmeir et al. 2013).

A raw descriptive analysis of the response variable (number of citations, ncit) gives an average of 1.64 and a variance of 7.53. Roughly 46% of the observations are zeros, the smallest and largest observed values are 0 and 40. While these summary statistics do not take the potential covariate effects into account, they already provide evidence that overdispersion and zero-inflation may be relevant to obtain a realistic model for the conditional distribution of the number of citations.

We consider the four candidates Poisson, ZIP, NB, and ZINB as possible distributions for the response. Available explanatory variables are the grant year (year), the number of designated states (ncountry), the number of claims against the patents (nclaims), and further binary and categorical covariates collected in \( x \) (compare Table G4 of the online supplement). Our model class allows the inclusion of all covariates in each distribution parameter, but model selection shows that this
usually does not lead to an optimal model. To determine such models we proceeded as follows:

Given one of the four distribution candidates, we used the DIC to select appropriate predictor structures for all distribution parameters to obtain a good compromise between fit and model complexity. Starting from the full model with predictor structure $\eta = f_1(\text{ncountry}) + f_2(\text{year}) + f_3(\text{nclaims}) + x'\beta$ for all relevant model parameters, we performed a stepwise model selection and decided for the model with the lowest DIC based on the rule of thumb discussed in Section F. When comparing two models, this rule indicates a clear preference for one of the two models under consideration if the DIC difference is larger than 10 points whereas it is indecisive otherwise. In the latter case, we assist the model selection by looking at the significances of the estimated effects, that is, we exclude a parametric effect from the optimal model if the 95% credible interval contains zero. For nonlinear effects, neither conducting separate tests for each knot nor pointwise credible intervals are desirable approaches to assess the uncertainty of the complete estimated nonlinear effect. We, therefore, base our decision on simultaneous credible bands (Krivolobokova, Kneib, and Claeskens 2010) and exclude nonlinear effects if such a 95% credible band contains zero for all observed covariate values. Overall, this procedure aims at favoring sparse and parsimonious models since the considerable flexibility offered by GAMLSS may otherwise tempt researchers to specify overly complex models. Of course, the rule of thumb may (and should) be adapted to the specific requirements of an analysis. To study how estimates suffer when fitting misspecified models, we conducted additional simulations described in Sections E.2.4 and F.3 of the online supplement. The basic outcome of these studies is that point estimates of influential effects are rather insensitive to the inclusion of noise variables that do not have an effect on the response distribution.

In the following, optimal models obtained from our stepwise search are denoted by (M1). The predictors of ZINB_M1, for example, are given by

$$\eta^\mu = \beta_0^\mu + \beta_1^\mu \text{ncountry} + \beta_2^\mu \text{patgsgr} + f_1^\mu(\text{year}) + f_2^\mu(\text{nclaims})$$

$$\eta^\pi = \beta_0^\pi + \beta_1^\pi \text{ncountry} + f_1^\pi(\text{year}) + f_2^\pi(\text{nclaims})$$

$$\eta^\delta = \beta_0^\delta + \beta_1^\delta \text{patgsgr} + f_2^\delta(\text{year} - 1991)$$

For the other distributions, the predictors for the location parameter $\lambda$ or $\mu$ are identical to that of ZINB. For the parameters $\pi$ or $\delta$ the selected covariate effects in the ZIP and NB model differ compared to ZINB, see Table G5 of the online supplement for a full documentation. To achieve some evidence that our flexible model class offers an improved fit for the patent data, we also considered models in which

(M2) the predictor of $\lambda$ (or $\mu$) and $\pi$ is a linear version of M1 (i.e., all spline-based estimates are replaced by linear effects) and where $\delta$ is estimated as a constant not depending on covariates. These models provide a Bayesian analogue to the results obtained with the R package psc1 (Zeileis, Kleiber, and Jackman, 2008) for ZIP or ZINB, as well as to a usual generalized linear model for Poisson or NB.

(M3) the predictors for the mean parameter of the count process are identical to those of M1 but $\pi$ and $\delta$ are considered as constants not depending on any covariates as proposed in Fahrmeir and Osuna Echavarría (2006). While estimating a constant $\pi$ in the ZIP model and a constant $\delta$ in the NB model did not cause any problems, a ZINB model with flexible predictor for $\mu$ and constant $\pi$ and $\delta$ was somewhat problematic in our implementation but also in the original implementation of Fahrmeir and Osuna Echavarría (2006). Also, when estimating the model with a linear predictor for $\mu$ in psc1, we obtained exceptionally high standard errors. Consequently, we consider ZINB_M3 as the model with $\eta^\mu$ and $\eta^\pi$ from ZINB_M1 and $\delta$ restricted to be constant.

The results of all 11 models were compared in terms of normalized (randomized) quantile residuals as a graphical device suggested by Stasinopoulos, Rigby, and Akantziliotou (2008): For an observation $y_i$, the residual is given by $\hat{r}_i = \Phi^{-1}(u_i)$ where $\Phi^{-1}$ is the inverse cumulative distribution function of a standard normal distribution, $u_i$ is a random value from the uniform distribution on the interval $[F(y_i - 1|\hat{\theta}), F(y_i|\hat{\theta})]$. This comprises all estimated model parameters and $F(\cdot|\hat{\theta})$ is the cumulative distribution function obtained by plugging in those estimated parameters. If the residuals are evaluated for the true model, they follow a standard normal distribution (Dunn and Smyth 1996) and, therefore, models can be checked by quantile-quantile-plots. Since the residuals are random, several randomized sets of residuals have to be studied before a decision about the adequacy of the model can be made. Figure 5 shows one realization for the flexible models Poisson_M1, ZIP_M1, NB_M1, and ZINB_M1. The residuals of the remaining, simpler models are given in the supplement Figure G38. Figure 5 clearly indicates a preference for the NB or ZINB model that provide a considerably better fit for estimating the distribution of patent citations in general. Although the residuals of the Poisson model can be improved by applying the ZIP model, the sample quantiles greater than 2 are too high compared to the true quantiles. Both the NB and ZINB model seem to overcome this problem. Furthermore, comparing Figures 5 and G38 indicates an improvement in both cases when allowing for nonlinear effects and estimating $\pi$ and $\delta$ as functions of relevant covariates.

In a second step, we applied proper scoring rules proposed by Gneiting and Raftery (2007) to confirm the findings assessed by the residuals: Let $y_1, \ldots, y_K$ be data in a hold out sample and $\hat{p}_r = (\hat{p}_{r0}, \hat{p}_{r1}, \ldots)$ the estimated probabilities of a predictive distribution with $\hat{p}_{rk} = P(y_r = k)$. A score is obtained by summing up individual score contributions, that is,

$$S = \sum_{r=1}^{R} S(\hat{p}_r, y_r).$$

Let $p_0$ be the true distribution, then Gneiting and Raftery (2007) took the expected value of the score under $p_0$ to compare different scoring rules. A scoring rule is called proper if $S(p_0, p_0) \geq S(\hat{p}, p_0)$ for any predictive distribution $\hat{p}$ and it is strictly proper if equality holds if and only if $\hat{p} = p_0$. We consider three common scores: the Brier (or quadratic) score, $S(\hat{p}_r, y_r) = - \sum_k (1(y_r = k) - \hat{p}_{rk})^2$, the logarithmic score, $S(\hat{p}_r, y_r) = \log(\hat{p}_{yr})$, and the spherical score $S(\hat{p}_r, y_r) = \frac{\hat{p}_{yr}}{\sqrt{\sum_k \hat{p}_{rk}^2}}$. All these scoring rules are strictly proper but the logarithmic score has the drawback that it only takes into account one single probability of the predictive distribution and is, therefore, susceptible to extreme observations. This property is shared with the DIC since both are based on the log-likelihood. In our application, the predictive distribution is assessed by a 10-fold cross-validation. Table 1 summarizes scores and DIC.
for all 11 candidate models. Similar to the residuals, the scores indicate that a Poisson distribution is the worst assumption. The scores of ZIP are higher compared to Poisson but the best scores are obtained from ZINB. Including relevant effects in all distribution parameters does not only improve the fit to the data but also the predictive ability since scores of M1 models are higher compared to M3 models. We find similar results for flexible nonlinear specifications compared to linear predictors (M1 vs. M2). Comparing all 11 models, scores and DIC would be in favor of the most flexible model ZINB M1 such that we choose this model as our final model for illustration purposes.

By checking MCMC samples graphically, we found that 12,000 MCMC iterations with a burnin of 2000 and a thinning of 10 yield convincing results in all Poisson, ZIP, NB, and ZINB models listed in Table 1. The computing time of ZINB M1 is about 4 min on a PC with Intel(R) Core i7 3.40 GHz processor. Since the nonlinear functions \( f_1 \) to \( f_3 \) are modeled by cubic P-splines with 20 inner knots and second-order random walk prior, the ZINB M1 model contains 99 unknown parameters (compare Table G5). The effective number of parameters is estimated to be approximately 36. For all three distribution parameters, the effective MCMC sample sizes are on average greater than 410 and show only rather small autocorrelations. Similar to the simulation studies, the posterior distributions of effects are very robust with respect to the prior choice and hyperparameter setting. For all considered choices, we did not face mixing problems and obtain very similar effects. The effects of year and nclaims on \( \mu \) are slightly rougher for an IG\((a_j, b_j)\) prior with \( a_j = b_j = 0.01 \). Further evidence and detailed results are given in the supplement G.1.

Figure 6 displays posterior mean estimates of ncountry, year, and nclaims on \( \mu \) (first row) and of ncountry on \( \pi \) (second row) together with pointwise 80% and 95% credible intervals. Vertical stripes indicate the relative amount of observations relating to the different covariate values (the darker the stripes, the more data). For the variable year, we found a linear impact on \( \pi \). A summary of the posterior distributions of all included linear effects can be found in the supplement, Table G6.

The following observations and interpretations on selected effects can be made: Looking at patents with grant year later than 1985, we estimate that patents are cited the less the newer they are. The effect of ncountry on \( \pi \) is the highest at around 10 designated states while the effect of year rises the probability of never being cited steadily. For an adequate interpretation of the expectation \((1 - \pi)\mu\) of the response, it is important to see that an increase of the effects on \( \mu \) and a decline of the function estimates on \( \pi \) result in a growing estimated expectation and vice versa. For the patent data, we find that the influence of year and ncountry on the total expectation behaves similar as when considering only the expectation of the count data part \( \mu \).

---

Table 1. Patent citations: Summarized scores in all models obtained from 10-fold cross-validation and DIC based on estimates of the whole dataset

| Model   | Quadratic score | Log score  | Spherical score | DIC   |
|---------|-----------------|------------|-----------------|-------|
| Poisson_M1 | -3803.74       | -10632.76  | 2139.66         | 19,876|
| Poisson_M2 | -3864.39       | -10488.12  | 2091.64         | 20,430|
| ZIP_M1   | -3465.55       | -8666.79   | 2500.36         | 17,119|
| ZIP_M2   | -3463.55       | -8166.37   | 2500.02         | 17,470|
| ZIP_M3   | -3516.25       | -9019.12   | 2472.58         | 17,420|
| NB_M1    | -3408.13       | -7996.85   | 2557.12         | 15,746|
| NB_M2    | -3456.03       | -8119.29   | 2526.99         | 16,078|
| NB_M3    | -3441.16       | -8140.00   | 2534.55         | 15,966|
| ZINB_M1  | -3378.04       | -7939.16   | 2576.78         | 15,644|
| ZINB_M2  | -3386.04       | -7953.84   | 2566.39         | 15,804|
| ZINB_M3  | -3383.05       | -7962.77   | 2571.01         | 15,679|
6. MODELING CLAIM FREQUENCIES

We also apply the developed methods to a dataset with \( n = 162,548 \) policyholders from a car insurance in Belgium of the year 1997. The insurance premium in car insurances is based on detailed statistical analyses of the policyholder’s risk structure. One important step is to model the loss frequency which usually depends on characteristics of the policyholder and the vehicle. Typical covariates are the age of the policyholder (ageph), age of the car (agec), the engine power (power), and the previous claim experience. In Belgium, the claim experience is measured by a 23-level bonus-malus-score (bm). The higher the score, the worse the claim history of the policyholder. Furthermore, the geographical information in which of the 589 districts (distr) in Belgium the policyholder’s car is registered is provided.

The dataset has already been treated in Denuit and Lang (2004) who applied geoadditive Poisson models. A detailed analysis based on both count data regression for claim frequencies and zero adjusted models (as introduced in Heller, Stasinopoulos, and Rigby 2006) for claim sizes in the framework of GAMLSS is provided in Klein et al. (2014). Here we build upon these more detailed treatments to illustrate the application of zero-inflated and overdispersed models for claim frequencies. We, therefore, consider the predictor

\[
\eta = f_1(\text{ageph}) + \text{sex} f_2(\text{ageph}) + f_3(\text{agec}) + f_4(\text{bm})
\]

\[
+ f_5(\text{power}) + f_6(\text{distr}) + x' \beta
\]

for the mean parameter in the count process, that is, \( \lambda \) in case of ZIP and \( \mu \) in case of NB or ZINB. The spatial effect has been modeled by a Markov random field and the term \( x' \beta \) contains additional linear effects of dummy variables that will not be discussed here. Model comparison is provided in the online supplement Section H.1. In summary, there is no clear evidence in favor of ZIP or NB and both models seem to provide a more reasonable fit compared to the Poisson model. In the following, we present results for the ZIP model to illustrate the interpretation of estimated effects. The selected model for \( \pi \) has

\[
\eta^\pi = f_1'(\text{ageph}) + f_2'(\text{agec}) + f_3'(\text{distr}) + (x')' \beta^\pi.
\]

Based on our selection rule discussed in the previous section, the spatial effect could also be excluded from the predictor as the DIC is indecisive and the simultaneous credible intervals fully cover the zero line. However, the estimated spatial effect displays an interesting cluster structure with districts of consistently low effect in the north around Antwerp. Replacing the smooth spatial effect with a dummy for this low-risk region also yielded a significant effect so that we decided to include the spatial effect for illustration purposes.

The computing time was less than 450 min on a PC with Intel(R) 3.40 GHz processor and a total number of 55,000 MCMC iterations. In Figure 7, the estimated nonlinear effects on \( \pi \) are plotted together with 80% and 95% pointwise credible intervals. Again, vertical stripes indicate the relative amount of data of the corresponding covariate values. Figure 8 depicts the estimated spatial effects on \( \lambda \) and \( \pi \). We also reestimated the models with different hyperparameters for the inverse gamma priors, including flat priors for variances or standard deviations, as well as half-Cauchy priors for the standard deviations. Compared to the simulation studies in hierarchical geoadditive models (where an additional iid random effect was simulated, which was neither significant nor selected by the DIC in this study),
the mixing of the spatial effect on $\lambda$ for values of $\tau_j^2$ as small as $a_j = b_j = 0.000001$ was less problematic. Nonlinear and linear effects, as well as the spatial effect on $\lambda$ were estimated similarly in all cases, compare Section H.2 for further illustrations. Some smaller differences occurred for the spatial effect on the probability of excess zeros but these are likely to be caused by the small values that are estimated for this effect anyway.

The spatial effect in Figure 8 clearly indicates a large number of expected claims in urban areas like Brussels, Antwerp, or Liège. For $\pi$ the monotonically increasing effect of age-c can be seen as an indication for an excess of zero claims for older cars. The estimated effects for $\lambda$ (shown in Figure H43 of the online supplement) are generally close to those found in Denuit and Lang (2004).

7. SUMMARY AND CONCLUSIONS

In this article, we developed numerically efficient Bayesian zero-inflated and overdispersed count data regression with semiparametric predictors as special cases of GAMLSS relying on iteratively weighted least squares proposals. A particular focus has been laid on the ZIP, NB, and ZINB distribution as standard choices for applied work. Our framework goes far beyond the model flexibility in the R package gamlasso (Stasinopoulos and Rigby 2007) since our predictors may include complex, hierarchical spatial effects and may in general cope with hierarchical data situations as described in Lang et al. (2014). Moreover, simulation studies revealed that the Bayesian approach yields reliable credible intervals in situations where the asymptotic likelihood theory fails while at the same time giving point estimates of at least similar quality. For model choice, we considered quantile residuals as a possibility to evaluate the general potential of a given model to fit the data. The deviance information criterion takes the complexity of an estimated model into account and can, therefore, be seen as a valuable tool both in comparing response distributions and predictor specifications. Proper scoring rules evaluated on hold out samples allow to assess the predictive ability of estimated models. Nevertheless, model choice and variable selection remain relatively tedious in particular due to multiple predictors involved. In the future it would, therefore, be desirable to develop automatic model choice and variable selection strategies in the spirit of Belitz and Lang (2008) in a frequentist setting or Scheipl, Fahrmeir, and Kneib (2012) in a Bayesian approach via spike and slab priors.

The Bayesian formulation of GAMLSS also provides the possibility to include modified/extended prior structures without major changes of the basic algorithm. For example, truncated normal priors may be considered to further improve the numerical efficiency or Dirichlet process mixture priors could be included to facilitate the inclusion of nonnormal random effects distributions. It will also be of interest to extend the Bayesian treatment of GAMLSS to further classes of discrete and continuous distributions or even combinations of both. A first attempt in the direction of the latter has been made in Klein et al. (2014) in the context of zero-adjusted models as introduced in a frequentist setting by Heller, Stasinopoulos, and Rigby (2006).

SUPPLEMENTARY MATERIALS

The online supplement for this article includes further details on the backfitting algorithm for finding the starting values of the MCMC algorithm and additional information about the multilevel framework; the explicit derivation of all working weights for the proposal densities and proofs for their positivity; the proof of Theorem 1; additional simulation studies dealing with additive and geoadditive NB and ZINB models; the impact of small sample sizes; identifiability and separability of effects; prior sensitivity; model choice based on the DIC; and additional results for the applications on patent citations and claim frequencies in car insurances.

[Received June 2013. Revised March 2014.]

REFERENCES

Belitz, C., and Lang, S. (2008), “Simultaneous Selection of Variables and Smoothing Parameters in Structured Additive Regression Models,” Computational Statistics and Data Analysis, 53, 61–81. Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2012), “Bayesx – Software for Bayesian Inference in Structured Additive Regression Models,” Version 2.1. Available at http://www.bayesx.org [406,411]
Breger, A., and Lang, S. (2006), “Generalized Structured Additive Regression Based on Bayesian P-Splines,” Computational Statistics & Data Analysis, 50, 967–991. [406]

Czado, C., Erhardt, V., Min, A., and Wagner, S. (2007), “Zero-Inflated Generalized Poisson Models With Regression Effects on the Mean, Dispersion and Zero-Inflation Level Applied to Patent Outsourcing Rates,” Statistical Modelling, 7, 125–153. [406]

Denuit, M., and Lang, S. (2004), “Non-Life Rate-Making With Bayesian GAMs,” Insurance: Mathematics and Economics, 35, 627–647. [417,418]

Dunn, P. K., and Smyth, G. K. (1996), “Randomized Quantile Residuals,” Journal of Computational and Graphical Statistics, 5, 236–245. [406,415]

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. [410,411]

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. [413]

Fahrmeir, L., and Ousma Echavarría, L. (2006), “Structured Additive Regression for Overdispersed and Zero-Inflated Count Data,” Applied Stochastic Models in Business and Industry, 22, 351–369. [406,415]

Fahrmeir, L., and Tutz, G. (2001), Multivariate Statistical Modelling Based on Generalized Linear Models, New York: Springer. [405]

Fahrmeir, L., Kneib, T., and Lang, S. (2004), “Penalized Structured Additive Regression for Space-Time Data: A Bayesian Perspective,” Statistica Sinica, 14, 731–761. [406,408]

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013), Regression—Models, Methods and Applications, New York: Springer. [414]

Gamerman, D. (1997), “Sampling From the Posterior Distribution in Generalized Linear Mixed Models,” Statistics and Computing, 7, 57–68. [406]

Gelman, A. (2006), “Prior Distributions for Variance Parameters in Hierarchical Models,” Bayesian Analysis, 1, 515–533. [408]

Gneiting, T., and Raftery, A. E. (2007), “Strictly Proper Scoring Rules, Prediction, and Estimation,” Journal of the American Statistical Association, 102, 359–378. [406,415]

Hastie, T. J., and Tibshirani, R. J. (1990), Generalized Additive Models, London: Chapman & Hall. [406,411]

He, X., Xue, H., and Shi, N. Z. (2010), “Sieve Maximum Likelihood Estimation for Doubly Semiparametric Zero-Inflated Poisson Models,” Journal of Multivariate Analysis, 101, 2026–2038. [406]

Heller, G., Stasinopoulos, D. M., and Rigby, R. A. (2006), “The Zero-Adjusted Inverse Gaussian Distribution as a Model for Insurance Data,” in Proceedings of the 21st International Workshop on Statistical Modelling, eds. J. Newell, J. Hinde, and J. Einbeck, pp. 226–233. [417,418]

Hilbe, J. M. (2007), Negative Binomial Regression, Cambridge, UK: Cambridge University Press. [405]

Jullion, A., and Lambert, P. (2007), “Robust Specification of the Roughness Penalty Prior Distribution in Spatially Adaptive Bayesian P-Splines Models,” Computational Statistics & Data Analysis, 51, 2542–2558. [406]

Kaas, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (2012), “Markov Chain Monte Carlo in Practice: A Roundable Discussion,” The American Statistician, 52, 93–100. [413]

Klein, N., Denuit, M., Kneib, T., and Lang, S. (2014), “Nonlife Ratemaking and Risk Management With Bayesian Generalized Additive Models for Location, Scale and Shape,” Insurance: Mathematics and Economics, 55, 225–249. [417,418]

Krivobokova, T., Kneib, T., and Claeskens, G. (2010), “Simultaneous Confidence Bands for Penalized Spline Estimators,” Journal of the American Statistical Association, 105, 852–863. [412,415]

Lang, S., and Breger, A. (2004), “Bayesian P-Splines,” Journal of Computational and Graphical Statistics, 13, 183–212. [408]

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K., and Kneib, T. (2014), “Multilevel Structured Additive Regression,” Statistics and Computing, 24, 223–238. [406,414,413,418]

Polson, N. G., and Scott, J. G. (2012), “On the Half-Cauchy Prior for a Global Scale Parameter,” Bayesian Analysis, 7, 887–902. [408]

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, 54, 507–554. [406,407]

Rue, H., and Held, L. (2005), Gaussian Markov Random Fields, London: Chapman & Hall / CRC. [408,409]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), Semiparametric Regression, Cambridge: Cambridge University Press. [406]

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. [418]

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), “Bayesian Measures of Model Complexity and Fit,” Journal of the Royal Statistical Society, Series B, 64, 583–639. [406,414]

Stasinopoulos, D. M., and Rigby, R. A. (2007), “Generalized Additive Models for Location Scale and Shape (GAMLSS) in R,” Journal of Statistical Software, 23, [406,411,418]

Stasinopoulos, D. M., Rigby, B., and Akantziliotou, C. (2008), Instructions on How to Use the Gamlss Package in R (2nd ed.), available at www.gamlss.org. [413,415]

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. [410]

Winkelmann, R. (2008), Econometric Analysis of Count Data, New York: Springer. [406]

Wood, S. N. (2004), “Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models,” Journal of the American Statistical Association, 99, 673–686. [406]

Wood, S. N. (2006), Generalized Additive Models : An Introduction With R, London: Chapman & Hall. [406]

——— (2008), “Fast Stable Direct Fitting and Smoothness Selection for Generalized Additive Models,” Journal of the Royal Statistical Society, Series B, 70, 495–518. [406]

Zeileis, A., Kleiber, C., and Jackman, S. (2008), “Regression Models for Count Data in R,” Journal of Statistical Software, 27. [406,415]