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

Nadja Klein, Thomas Kneib
Chair of Statistics
Georg-August-University Göttingen

Stefan Lang
Department of Statistics
University of Innsbruck

A A Backfitting Algorithm

In this section, we summarise a backfitting algorithm for obtaining the starting values for the MCMC sampler derived in the main paper following the ideas in (Hastie and Tibshirani, 1990) for exponential family regression. Basically, we determine the maximum of the penalised log-likelihood (i.e. the posterior mode) by numerically maximising its quadratic approximation. This yields the following algorithm:

(1) *Initialisation.* Set $\beta_1^{(0)} = \dots = \beta_J^{(0)} = 0$ and $\beta_0^{(0)} = g(\hat{\vartheta})$ where g is the link function between the generic model parameter ϑ and the predictor $\boldsymbol{\eta}$ and $\hat{\vartheta}$ is a simple estimator for ϑ only depending on the responses such as the ML estimate for a model without any covariates. If, for example, ϑ stands for the average rate $\bar{\lambda} = \frac{1}{n} \sum_{i=1}^n \lambda_i$ in the ZIP model, $\hat{\vartheta}$ could be the mean of the observations $\mathbf{y} = (y_1, \dots, y_n)'$. Let T be the maximum number of iterations in the backfitting algorithm and set $t = 0$.

(2) *Estimation of $\mathbf{f}_1, \dots, \mathbf{f}_J$ and β_0 .*

(a) Set $r = 0$ and for $j = 1, \dots, J$

$$\mathbf{f}_j^{(r)} = \mathbf{f}_j^{(t)} = \mathbf{Z}_j \boldsymbol{\beta}_j^{(t)}, \quad \text{as well as} \quad \beta_0^{(r)} = \beta_0^{(t)} = g(\hat{\vartheta})$$

(b) Outer backfitting loop: Compute

$$\mathbf{z}^{(t)} = \boldsymbol{\eta}^{(t)} + \left(\mathbf{W}^{(t)} \right)^{-1} \mathbf{v}^{(t)}$$

and define $\mathbf{S}_j^{(t)} := \mathbf{Z}_j \left(\mathbf{Z}_j' \mathbf{W}^{(t)} \mathbf{Z}_j + \frac{1}{\tau_j^2} \mathbf{K}_j \right)^{-1} \mathbf{Z}_j' \mathbf{W}^{(t)}$

(c) Inner backfitting loop: Calculate for $j = 1, \dots, J$

$$\mathbf{f}_j^{(r+1)} = \mathbf{S}_j^{(t)} \left(\mathbf{z}^{(t)} - \sum_{\substack{s=1 \\ s \neq j}}^J \mathbf{f}_s^{(r)} \right)$$

(d) Centre the estimates.

(e) If for fixed $\varepsilon > 0$

$$\frac{\left| \beta_0^{(r+1)} - \beta_0^{(r)} \right| + \sum_{j=1}^J \left\| \mathbf{f}_j^{(r+1)} - \mathbf{f}_j^{(r)} \right\|}{\left| \beta_0^{(r+1)} \right| + \sum_{j=1}^J \left\| \mathbf{f}_j^{(r+1)} \right\|} < \varepsilon$$

stop the inner backfitting loop, set

$$\mathbf{Z}_j \boldsymbol{\beta}_j^{(t+1)} = \mathbf{f}_j^{(r+1)} \quad \text{as well as} \quad \beta_0^{(t+1)} = \beta_0^{(r+1)}$$

for $j = 1, \dots, J$ and go to step (f). Otherwise set $r = r + 1$ and go to (c).

(f) If $t < T$ go to (b). Otherwise stop the algorithm.

B Multilevel Framework

Recently, Lang et al. (2014) proposed a multilevel version of structured additive regression models which assume that the regression coefficients $\boldsymbol{\beta}_j$ of an effect $\mathbf{Z}_j \boldsymbol{\beta}_j$ in (2) of the main paper may themselves obey a regression model with structured additive predictor, i.e.

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \mathbf{Z}_{j1} \boldsymbol{\beta}_{j1} + \dots + \mathbf{Z}_{jJ_j} \boldsymbol{\beta}_{jJ_j} + \boldsymbol{\varepsilon}_j \quad (\text{B.1})$$

where $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \tau_j^2 \mathbf{I})$ is a vector of i.i.d. Gaussian random effects. Therefore, the regression coefficients $\boldsymbol{\beta}_j$ are related to additional nonlinear functions $f_{j1}, \dots, f_{jJ_j}$ that are again represented in terms of basis function expansions inducing the design matrices $\mathbf{Z}_{j1}, \dots, \mathbf{Z}_{jJ_j}$ and basis coefficients $\boldsymbol{\beta}_{j1}, \dots, \boldsymbol{\beta}_{jJ_j}$ which are constructed in complete analogy to the description above but with dimensions defined by the dimension of the coefficient vector $\boldsymbol{\beta}_j$.

A typical application are multilevel data where a hierarchy of units or clusters grouped at different levels is given. For the purpose of this paper, a particularly useful application is a model with spatial effects. In this case, covariate $s_i \in \{1, \dots, S\}$ is a spatial

index indicating the district observation i pertains to. Then the design matrix $\mathbf{Z}_j$ is an $n \times S$ indicator matrix with $\mathbf{Z}_j[i, s] = 1$ if the i -th observation belongs to district s and zero otherwise. The $S \times 1$ parameter vector β_j is the vector of regression parameters, i.e. the s -th element in β_j corresponds to the regression coefficient of the s -th district. Using the compound prior (B.1), we obtain an additive decomposition of the district-specific spatial effect. If no further, district-specific covariate information is available, we use the specific compound prior

$$\beta_j = \mathbf{Z}_{j1}\beta_{j1} + \epsilon_j = \mathbf{I}\beta_{j1} + \epsilon_j$$

where $\mathbf{Z}_{j1}\beta_{j1} = \mathbf{I}\beta_{j1}$ is a structured spatial effect modelled by a Markov random field prior whereas $\epsilon_j \sim N(\mathbf{0}, \tau_j^2 \mathbf{I})$ can be regarded as an additional unstructured i.i.d. random effect. The great advantage of the multilevel approach is that the full conditionals of the Markov random field become Gaussian making IWLS proposals unnecessary. Hence, problems with too low acceptance rates in applications with a large number of spatial units can be avoided. Another important advantage is the reduction in computing time as the “number of observations” relevant for updating the second level regression coefficients β_{j1} reduces to the number of districts. This is typically much less than the actual number of observations. For instance, in the insurance data set we have 162,548 observations but only 589 districts. The paper by Lang et al. (2014) also proposes highly efficient updating of the remaining terms in the level one equation (2) of the main paper by utilising the fact that for most covariates the number of different observations is far less than the actual number of observations. Although details are beyond the scope of this paper, we point out that our software is fully capable of the multilevel framework outlined in Lang et al. (2014) and makes use of the numerical efficient updating schemes described therein.

C Working Weights

C.1 Computation of the Working Weights

In principle, the working weights shown in Section 3.1 are determined as follows: For the score vector $\mathbf{v}$, the first derivatives of the log-likelihood with respect to each predictor have to be computed. The working weights are then derived by taking the expectation of the negative second derivative of the log-likelihood, compare Section 3 for more detailed explanations and formulas. We start with the ZIP model and make

use of the following equations:

$$l = \sum_{y_i=0} \log(\pi_i + (1 - \pi_i) \exp(-\lambda_i)) + \sum_{y_i>0} (\log(1 - \pi_i) + y_i \log(\lambda_i) - \lambda_i - \log(y_i!))$$

$$\begin{aligned} \frac{\partial \pi_i}{\partial \eta_i^\pi} &= \pi_i(1 - \pi_i) \\ \frac{\partial \lambda_i}{\partial \eta_i^\lambda} &= \lambda_i \\ \mathbb{E}(\mathbb{1}_{\{0\}}(y_i)) &= \mathbb{P}(y_i = 0) \end{aligned}$$

This yields the first derivatives

$$\begin{aligned} v_i^\lambda = \frac{\partial l}{\partial \eta_i^\lambda} &= \frac{-(1 - \pi_i)\lambda_i \exp(-\lambda_i)}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} \mathbb{1}_{\{0\}}(y_i) + (y_i - \lambda_i)(1 - \mathbb{1}_{\{0\}}(y_i)) \\ &= \frac{\pi_i \lambda_i}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} \mathbb{1}_{\{0\}}(y_i) + (y_i - \lambda_i) \\ v_i^\pi = \frac{\partial l}{\partial \eta_i^\pi} &= \frac{\pi_i(1 - \pi_i)(1 - \exp(-\lambda_i))}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} \mathbb{1}_{\{0\}}(y_i) - \pi_i(1 - \mathbb{1}_{\{0\}}(y_i)) \\ &= \frac{\pi_i}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} \mathbb{1}_{\{0\}}(y_i) - \pi_i \end{aligned}$$

and the working weights

$$\begin{aligned} w_i^\lambda &= \mathbb{E} \left(-\frac{\partial^2 l}{(\partial \eta_i^\lambda)^2} \right) \\ &= \mathbb{E} \left(-\frac{\pi_i \lambda_i}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} \mathbb{1}_{\{0\}}(y_i) - \frac{\pi_i(1 - \pi_i)\lambda_i^2 \exp(-\lambda_i)}{(\pi_i + (1 - \pi_i) \exp(-\lambda_i))^2} \mathbb{1}_{\{0\}}(y_i) + \lambda_i \right) \\ &= \frac{\lambda_i(1 - \pi_i)(\pi_i + (1 - \pi_i) \exp(-\lambda_i) - \exp(-\lambda_i)\lambda_i\pi_i)}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} \\ w_i^\pi &= \mathbb{E} \left(-\frac{\partial^2 l}{(\partial \eta_i^\pi)^2} \right) \\ &= \mathbb{E} \left(-\frac{\pi_i(1 - \pi_i)}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} \mathbb{1}_{\{0\}}(y_i) + \frac{\pi_i^2(1 - \pi_i)(1 - \exp(-\lambda_i))}{(\pi_i + (1 - \pi_i) \exp(-\lambda_i))^2} \mathbb{1}_{\{0\}}(y_i) + \pi_i(1 - \pi_i) \right) \\ &= \frac{\pi_i^2(1 - \pi_i)(1 - \exp(-\lambda_i))}{\pi_i + (1 - \pi_i) \exp(-\lambda_i)} \end{aligned}$$

For the ZINB model, calculations we first have

$$\begin{aligned} l &= \sum_{y_i=0} \log \left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right) \\ &\quad + \sum_{y_i>0} (\log(1 - \pi_i) + \log(\Gamma(y_i + \delta_i)) - \log(\Gamma(y_i + 1)) - \log(\Gamma(\delta_i))) \\ &\quad + \sum_{y_i>0} (\delta_i \log(\delta_i) + y_i \log(\mu_i) - (\delta_i + y_i) \log(\delta_i + \mu_i)) \end{aligned}$$

With the expressions

$$\begin{aligned}
\frac{\partial \pi_i}{\partial \eta_i^\pi} &= \pi_i(1 - \pi_i) \\
\frac{\partial \mu_i}{\partial \eta_i^\mu} &= \mu_i \\
\frac{\partial \delta_i}{\partial \eta_i^\delta} &= \delta_i \\
\frac{\partial}{\partial \eta_i^\mu} \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} &= -\frac{\delta_i \mu_i}{\delta_i + \mu_i} \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \\
\frac{\partial}{\partial \eta_i^\delta} \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} &= \delta_i \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \left(\log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\mu_i}{\delta_i + \mu_i} \right) \\
\mathbb{E}(\mathbb{1}_{\{0\}}(y_i)) &= \mathbb{P}(y_i = 0)
\end{aligned}$$

we obtain the first derivatives

$$\begin{aligned}
v_i^\mu &= \frac{\partial l}{\partial \eta_i^\mu} \\
&= \frac{-(1 - \pi_i) \delta_i \mu_i \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right) (\delta_i + \mu_i)} \mathbb{1}_{\{0\}}(y_i) + \frac{y_i \delta_i - \delta_i \mu_i}{\delta_i + \mu_i} (1 - \mathbb{1}_{\{0\}}(y_i)) \\
&= \frac{\pi_i \delta_i \mu_i}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right) (\delta_i + \mu_i)} \mathbb{1}_{\{0\}}(y_i) + \frac{y_i \delta_i - \delta_i \mu_i}{\delta_i + \mu_i} \\
v_i^\pi &= \frac{\partial l}{\partial \eta_i^\pi} \\
&= \frac{\pi_i (1 - \pi_i) \left(1 - \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right)}{\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}} \mathbb{1}_{\{0\}}(y_i) + \pi_i (1 - \mathbb{1}_{\{0\}}(y_i)) \\
&= \frac{\pi_i}{\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}} \mathbb{1}_{\{0\}}(y_i) - \pi_i \\
v_i^\delta &= \frac{\partial l}{\partial \eta_i^\delta} \\
&= \frac{(1 - \pi_i) \delta_i \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \left(\log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\mu_i}{\delta_i + \mu_i} \right)}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right)} \mathbb{1}_{\{0\}}(y_i) \\
&\quad + \left(\delta_i \log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\delta_i \mu_i - y_i \delta_i}{\delta_i + \mu_i} \right) (1 - \mathbb{1}_{\{0\}}(y_i)) + \delta_i (\psi(y_i + \delta_i) - \psi(\delta_i)) \\
&= \delta_i \left(\psi(y_i + \delta_i) - \psi(\delta_i) + \log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\mu_i - y_i}{\delta_i + \mu_i} \right) - \frac{\delta_i \pi_i \left(\log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\mu_i}{\delta_i + \mu_i} \right)}{\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}} \mathbb{1}_{\{0\}}(y_i)
\end{aligned}$$

and the working weights

$$\begin{aligned}
w_i^\mu &= \mathbb{E} \left(-\frac{\partial^2 l}{(\partial \eta_i^\mu)^2} \right) \\
&= \mathbb{E} \left(-\frac{\pi_i \delta_i \mu_i}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right) (\delta_i + \mu_i)} \mathbb{1}_{\{0\}}(y_i) \right) \\
&+ \mathbb{E} \left(\frac{\pi_i \delta_i \mu_i^2}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right) (\delta_i + \mu_i)^2} \mathbb{1}_{\{0\}}(y_i) \right) + \frac{\delta_i \mu_i}{(\delta_i + \mu_i)^2} \mathbb{E}(y_i) \\
&- \mathbb{E} \left(\frac{(1 - \pi_i) \pi_i \delta_i^2 \mu_i^2 \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right)^2 (\delta_i + \mu_i)^2} \mathbb{1}_{\{0\}}(y_i) \right) + \frac{\delta_i^2 \mu_i}{(\delta_i + \mu_i)^2} \\
&= \frac{\delta_i \mu_i (1 - \pi_i)}{(\delta_i + \mu_i)} - \frac{\pi_i (1 - \pi_i) \delta_i^2 \mu_i^2 \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right) (\delta_i + \mu_i)^2} \\
\\
w_i^\pi &= \mathbb{E} \left(-\frac{\partial^2 l}{(\partial \eta_i^\pi)^2} \right) \\
&= \mathbb{E} \left(-\frac{\pi_i (1 - \pi_i)}{\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}} \mathbb{1}_{\{0\}}(y_i) + \frac{\pi_i^2 (1 - \pi_i) \left(1 - \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right)}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right)^2} \mathbb{1}_{\{0\}}(y_i) + \pi_i (1 - \pi_i) \right) \\
&= \frac{\pi_i^2 (1 - \pi_i) \left(1 - \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right)}{\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}} \\
\\
w_i^\delta &= \mathbb{E} \left(-\frac{\partial^2 l}{(\partial \eta_i^\delta)^2} \right) \\
&= \mathbb{E} \left(\frac{\delta_i \pi_i \left(\log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\mu_i}{\delta_i + \mu_i} \right)}{\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}} \mathbb{1}_{\{0\}}(y_i) + \frac{\delta_i \pi_i \left(\frac{\mu_i}{\delta_i + \mu_i} - \frac{\delta_i \mu_i}{(\delta_i + \mu_i)^2} \right)}{\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}} \mathbb{1}_{\{0\}}(y_i) \right) \\
&- \mathbb{E} \left(\frac{\delta_i^2 (1 - \pi_i) \pi_i \left(\log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\mu_i}{\delta_i + \mu_i} \right)^2}{\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right)^2} \mathbb{1}_{\{0\}}(y_i) \right) \\
&- \delta_i \left(\mathbb{E}(\psi(y_i + \delta_i)) - \psi(\delta_i) + \log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \mathbb{E} \left(\frac{\mu_i - y_i}{\delta_i + \mu_i} \right) \right) \\
&- \delta_i \left(\delta_i \mathbb{E}(\psi_1(y_i + \delta_i)) - \delta_i \psi_1(\delta_i) + \frac{\mu_i}{\delta_i + \mu_i} - \frac{\delta_i \mu_i}{\delta_i + \mu_i} + \mathbb{E} \left(\frac{\delta_i y_i}{(\delta_i + \mu_i)^2} \right) \right)
\end{aligned}$$

$$\begin{aligned}
&= -\delta_i(1 - \pi_i) \left(\log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\mu_i}{\delta_i + \mu_i} \right) - \frac{(1 - \pi_i)\pi_i\delta_i^2 \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \left(\log \left(\frac{\delta_i}{\delta_i + \mu_i} \right) + \frac{\mu_i}{\delta_i + \mu_i} \right)^2}{\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i}} \\
&\quad - \delta_i (\mathbb{E}(\psi(y_i + \delta_i)) - \psi(\delta_i)) - \delta_i^2 (\mathbb{E}(\psi_1(y_i + \delta_i)) - \psi_1(\delta_i))
\end{aligned}$$

C.2 Positive Definiteness of the Working Weights

Lemma C.1. *The diagonal matrices $\mathbf{W}^\lambda$ and $\mathbf{W}^\pi$ of working weights in the ZIP model defined in (6) and (7) are positive definite.*

Proof. As both matrices are diagonal, we only have to show that all entries on the diagonal are greater than zero. Let us start with $\mathbf{W}^\lambda$: We need to prove that

$$\pi_i + (1 - \pi_i) \exp(-\lambda_i) > \lambda_i \pi_i \exp(-\lambda_i)$$

holds since the denominator in (6) is obviously strictly positive. Due to $\lambda_i > \log(\lambda_i)$ we get

$$\lambda_i \exp(-\lambda_i) = \exp(\log(\lambda_i) - \lambda_i) < 1.$$

Together with $(1 - \pi_i) \exp(-\lambda_i) > 0$ it follows that

$$\lambda_i \pi_i \exp(-\lambda_i) < \pi_i < \pi_i + (1 - \pi_i) \exp(-\lambda_i),$$

and hence, that the diagonal elements of $\mathbf{W}^\lambda$ are greater than zero. For $\mathbf{W}^\pi$ in (7) we only need to show

$$\exp(-\lambda_i) < 1.$$

This follows directly from $\lambda_i > 0$. □

Lemma C.2. *The diagonal matrices $\mathbf{W}^\mu$ and $\mathbf{W}^\pi$ of working weights in the ZINB model defined in (8) and (9) are positive definite.*

Proof. Again, we have to show that all entries on the diagonal are greater than zero.

Let us start with $\mathbf{W}^\mu$ in (8) by reducing all terms to their common denominator

$$\left(\pi_i + (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right) (\delta_i + \mu_i)^2$$

and comparing the numerators. The whole numerator is then given by

$$\begin{aligned}
&\delta_i^2 \mu_i \pi_i (1 - \pi_i) \left(1 - \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \right) + \delta_i \mu_i^2 (1 - \pi_i)^2 \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} + \delta_i^2 \mu_i (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \\
&\quad + \delta_i \mu_i^2 \pi_i (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \left(\left(\frac{\delta_i + \mu_i}{\delta_i} \right)^{\delta_i} - \delta_i \right).
\end{aligned}$$

The first term is greater than zero since we assume $\left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} < 1$. The second and third one are obviously also greater than zero because all factors are greater than zero. It still remains the last term $\delta_i \mu_i^2 \pi_i (1 - \pi_i) \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} \left(\left(\frac{\delta_i + \mu_i}{\delta_i}\right)^{\delta_i} - \delta_i\right)$. For this, we differentiate between the two cases $\delta_i \leq \mu_i$ and $\delta_i > \mu_i$.

- (i) $\delta_i \leq \mu_i$: It is sufficient to show $\left(\frac{\delta_i + \mu_i}{\delta_i}\right)^{\delta_i} \geq \delta_i$ or equivalently $\delta_i \log\left(1 + \frac{\mu_i}{\delta_i}\right) \geq \log(\delta_i)$. Because of $\frac{\mu_i}{\delta_i} \geq 1$ it is enough to prove that $\frac{1}{2}\delta_i \geq \log(\delta_i)$ holds. If $0 \leq \delta_i \leq 1$ there is nothing to do. For $\delta_i > 1$, we have

$$\begin{aligned} \log(\delta_i) &= \sum_{k=1}^{\infty} \frac{1}{k} \left(\frac{\delta_i - 1}{\delta_i}\right)^k \\ &= \frac{\delta_i - 1}{\delta_i} + \frac{1}{2} \left(\frac{\delta_i - 1}{\delta_i}\right)^2 + \frac{1}{3} \left(\frac{\delta_i - 1}{\delta_i}\right)^3 + \dots \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{2\delta_i} - \frac{1}{2\delta_i} + \frac{1}{2} \left(\frac{\delta_i - 1}{\delta_i}\right)^2 + \frac{1}{3} \left(\frac{\delta_i - 1}{\delta_i}\right)^3 + \dots \\ &\leq \frac{1}{2} + \frac{1}{2} - \frac{1}{2\delta_i} + \frac{1}{2} \left(\frac{\delta_i - 1}{\delta_i}\right)^2 + \frac{1}{3} \left(\frac{\delta_i - 1}{\delta_i}\right)^3 + \dots \\ &\leq \sum_{k=0}^{\infty} \frac{1}{2} \left(\frac{\delta_i - 1}{\delta_i}\right)^k \\ &= \frac{1}{2} \frac{1}{1 - \frac{\delta_i - 1}{\delta_i}} = \frac{1}{2} \delta_i. \end{aligned}$$

Finally, we have

$$\left(\frac{\delta_i + \mu_i}{\delta_i}\right)^{\delta_i} \geq \delta_i$$

and the third term is greater than zero in case of $\delta_i \leq \mu_i$.

- (ii) $\delta_i > \mu_i$: In this case, we rearrange the terms of the nominator as follows:

$$\begin{aligned} &\delta_i^2 \mu_i (1 - \pi_i)^2 \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} + \delta_i \mu_i^2 \pi_i (1 - \pi_i) + \delta_i \mu_i^2 (1 - \pi_i)^2 \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} \\ &+ \delta_i^2 \mu_i \pi_i (1 - \pi_i) \left(1 - \mu_i \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i}\right). \end{aligned}$$

It is sufficient to prove that

$$\frac{1}{\mu_i} \left(\frac{\delta_i + \mu_i}{\delta_i}\right)^{\delta_i} \geq 1.$$

If $\delta_i \leq 1$, there is nothing to show. For $\delta_i \geq 2$, we have

$$\begin{aligned}
\frac{1}{\mu_i} \left(1 + \frac{\mu_i}{\delta_i}\right)^{\delta_i} &= \frac{1}{\mu_i} \sum_{k=0}^{\infty} \binom{\delta_i}{k} \left(\frac{\mu_i}{\delta_i}\right)^k \\
&\geq \frac{1}{\mu_i} \sum_{k=0}^{\lfloor \delta_i \rfloor} \binom{\delta_i}{k} \left(\frac{\mu_i}{\delta_i}\right)^k \\
&\geq \frac{1}{\mu_i} + \frac{1}{\mu_i} \left(\frac{\mu_i}{\delta_i}\right)^{\lfloor \delta_i \rfloor} + \frac{1}{\mu_i} \sum_{k=1}^{\lfloor \delta_i \rfloor - 1} \binom{\delta_i}{k} \left(\frac{\mu_i}{\delta_i}\right)^k \\
&\geq \frac{1}{\mu_i} + \frac{1}{\mu_i} \left(\frac{\mu_i}{\delta_i}\right)^{\lfloor \delta_i \rfloor} + \frac{1}{\mu_i} \delta_i \sum_{k=1}^{\lfloor \delta_i \rfloor - 1} \left(\frac{\mu_i}{\delta_i}\right)^k \\
&= \frac{1}{\mu_i} + \frac{1}{\mu_i} \left(\frac{\mu_i}{\delta_i}\right)^{\lfloor \delta_i \rfloor} + \frac{1}{\mu_i} \delta_i \frac{\mu_i}{\delta_i} \sum_{k=0}^{\lfloor \delta_i \rfloor - 2} \left(\frac{\mu_i}{\delta_i}\right)^k \\
&= \frac{1}{\mu_i} + \frac{1}{\mu_i} \left(\frac{\mu_i}{\delta_i}\right)^{\lfloor \delta_i \rfloor} + \frac{1 - \left(\frac{\mu_i}{\delta_i}\right)^{\lfloor \delta_i \rfloor - 1}}{1 - \frac{\mu_i}{\delta_i}} \\
&\geq 1,
\end{aligned}$$

where $\lfloor \delta_i \rfloor = \max\{j \in \mathbb{Z} | j \leq \delta_i\}$. The only remaining case is $1 < \delta_i < 2$. Then we have

$$\begin{aligned}
\frac{1}{\mu_i} \left(\frac{\delta_i + \mu_i}{\delta_i}\right)^{\delta_i} &\geq \frac{1}{\mu_i} \frac{\delta_i + \mu_i}{\delta_i} \\
&\geq \frac{1}{\mu_i} \frac{\delta_i + \mu_i}{2} \\
&\geq \frac{1}{\mu_i} \frac{2\mu_i}{2} \\
&= 1.
\end{aligned}$$

All in all, we see that the entries on the diagonal of $\mathbf{W}^\mu$ are greater than zero and therefore that the matrix of working weights $\mathbf{W}^\mu$ is positive definite.

For $\mathbf{W}^\pi$ we need to look at (9). As $\left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i}$ is smaller than one, we get

$$1 - \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} > 0,$$

so that all factors in (9) are greater than zero. In consequence, all entries on the diagonal of $\mathbf{W}^\pi$ are greater than zero as desired. $\square$

D Propriety of the Posterior Distribution

Proof of Theorem 3.1. As described in the main paper, following the steps in Fahrmeir and Kneib (2009) we finally represent (2) in terms of $\tilde{\boldsymbol{\eta}}_s = \tilde{\mathbf{X}}\boldsymbol{\xi}_s + \tilde{\mathbf{V}}_s\mathbf{b}_s +$

$\boldsymbol{\varepsilon}_{0s}, \boldsymbol{\varepsilon}_{0s} \sim N(0, \tau_0^2 \mathbf{I})$ with reduced dimension $\dim(\boldsymbol{\varepsilon}_{0s}) = \tilde{k}_0$ and for a subsample of observations. The advantage here is that this model contains random effects with proper priors and individual-specific error term $\boldsymbol{\varepsilon}_{0s}$, such that the findings of Sun et al. (2001) can be applied.

For each of the model terms $\mathbf{Z}_j \boldsymbol{\beta}_j$ of a generic predictor $\boldsymbol{\eta}$ with improper prior precision matrix $\mathbf{K}_j$ we have $\text{rk}(\mathbf{K}_j) = k_j < d_j = \dim(\boldsymbol{\beta}_j)$. Note that for model terms with proper priors nothing has to be done since in this case $\boldsymbol{\beta}_j$ can be interpreted as a random effect without reparameterisation. The result of Rue and Held (2005, p. 91) allows us to divide $\boldsymbol{\beta}_j$ into two parts, a $(d_j - k_j)$ -dimensional vector of fixed effects $\boldsymbol{\xi}_j$ with improper prior, and a k_j -dimensional vector $\mathbf{b}_j$ of random effects with proper prior such that $\mathbf{Z}_j \boldsymbol{\beta}_j = \mathbf{X}_j \boldsymbol{\xi}_j + \mathbf{V}_j \mathbf{b}_j$ holds. This leads to the submodel (12) of the main paper. Since $\mathbf{V}_{0s}^{\vartheta_l} (\mathbf{V}_{0s}^{\vartheta_l})'$ has full rank, we also have that $\mathbf{M}^{\vartheta_l}$ has full rank and hence that Condition (B.2) also holds for the normalised submodel (13).

Define for $l = 1, \dots, m$ the vectors $(\boldsymbol{\tau}^2)^{\vartheta_l} = \left((\tau_1^2)^{\vartheta_l}, \dots, (\tau_{j_{\vartheta_l}}^2)^{\vartheta_l} \right)'$. Then, showing that the posterior

$$p(\boldsymbol{\xi}^{\vartheta_1}, \dots, \boldsymbol{\xi}^{\vartheta_m}, \mathbf{b}^{\vartheta_1}, \dots, \mathbf{b}^{\vartheta_m}, \mathbf{b}_0^{\vartheta_1}, \dots, \mathbf{b}_0^{\vartheta_m}, (\boldsymbol{\tau}^2)^{\vartheta_1}, \dots, (\boldsymbol{\tau}^2)^{\vartheta_m}, (\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m} | \mathbf{y})$$

is proper is obviously equivalent to showing the propriety of

$$p(\boldsymbol{\xi}^{\vartheta_1}, \dots, \boldsymbol{\xi}^{\vartheta_m}, \mathbf{b}^{\vartheta_1}, \dots, \mathbf{b}^{\vartheta_m}, \tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m}, (\boldsymbol{\tau}^2)^{\vartheta_1}, \dots, (\boldsymbol{\tau}^2)^{\vartheta_m}, (\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m} | \mathbf{y}).$$

First of all we have

$$\begin{aligned} & p(\boldsymbol{\xi}^{\vartheta_1}, \dots, \boldsymbol{\xi}^{\vartheta_m}, \mathbf{b}^{\vartheta_1}, \dots, \mathbf{b}^{\vartheta_m}, \tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m}, (\boldsymbol{\tau}^2)^{\vartheta_1}, \dots, (\boldsymbol{\tau}^2)^{\vartheta_m}, (\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m} | \mathbf{y}) \\ & \propto \prod_{i=1}^n p(y_i | \eta_i^{\vartheta_1}, \dots, \eta_i^{\vartheta_m}) \prod_{l=1}^m p(\tilde{\boldsymbol{\eta}}_s^{\vartheta_l} | \boldsymbol{\xi}^{\vartheta_l}, \mathbf{b}^{\vartheta_l}, (\tau_0^2)^{\vartheta_l}) \prod_{l=1}^m p(\mathbf{b}^{\vartheta_l} | (\boldsymbol{\tau}^2)^{\vartheta_l}) \\ & \quad \times \prod_{l=1}^m p((\boldsymbol{\tau}^2)^{\vartheta_l}) \prod_{l=1}^m p((\tau_0^2)^{\vartheta_l}) \\ & \propto \prod_{i=1}^n p(y_i | \eta_i^{\vartheta_1}, \dots, \eta_i^{\vartheta_m}) G \end{aligned}$$

where

$$\begin{aligned} G &= \prod_{l=1}^m \left(\frac{1}{(\tau_0^2)^{\vartheta_l}} \right)^{\tilde{k}_0/2} \exp \left(-\frac{1}{2(\tau_0^2)^{\vartheta_l}} \left(\tilde{\boldsymbol{\eta}}_s^{\vartheta_l} - \tilde{\mathbf{X}}_s^{\vartheta_l} \boldsymbol{\xi}^{\vartheta_l} - \tilde{\mathbf{V}}_s^{\vartheta_l} \mathbf{b}^{\vartheta_l} \right)' \left(\tilde{\boldsymbol{\eta}}_s^{\vartheta_l} - \tilde{\mathbf{X}}_s^{\vartheta_l} \boldsymbol{\xi}^{\vartheta_l} - \tilde{\mathbf{V}}_s^{\vartheta_l} \mathbf{b}^{\vartheta_l} \right) \right) \\ & \quad \times \prod_{l=1}^m \left(\frac{1}{|\mathbf{Q}|^{\vartheta_l}} \right)^{1/2} \exp \left(-\frac{(\mathbf{b}^{\vartheta_l})' \mathbf{Q}^{\vartheta_l} \mathbf{b}^{\vartheta_l}}{2} \right) \prod_{l=1}^m p((\boldsymbol{\tau}^2)^{\vartheta_l}) \prod_{l=1}^m p((\tau_0^2)^{\vartheta_l}) \end{aligned}$$

and $\mathbf{Q}^{\vartheta_l} = \text{Cov}(\mathbf{b}^{\vartheta_l})$. From Condition (A.2) we get together with $M^* = M^{n-n^*}$

$$\begin{aligned} p(\boldsymbol{\xi}^{\vartheta_1}, \dots, \boldsymbol{\xi}^{\vartheta_m}, \mathbf{b}^{\vartheta_1}, \dots, \mathbf{b}^{\vartheta_m}, \tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m}, (\boldsymbol{\tau}^2)^{\vartheta_1}, \dots, (\boldsymbol{\tau}^2)^{\vartheta_m}, (\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m} | \mathbf{y}) \\ \leq M^* \prod_{i=1}^{n^*} p(y_i | \eta_i^{\vartheta_1}, \dots, \eta_i^{\vartheta_m}) G. \end{aligned}$$

Integrating over $\boldsymbol{\xi}^{\vartheta_1}, \dots, \boldsymbol{\xi}^{\vartheta_m}, \mathbf{b}^{\vartheta_1}, \dots, \mathbf{b}^{\vartheta_m}, (\boldsymbol{\tau}^2)^{\vartheta_1}, \dots, (\boldsymbol{\tau}^2)^{\vartheta_m}$ implies

$$\begin{aligned} p(\tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m}, (\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m} | \mathbf{y}) \\ \leq M^* \prod_{i=1}^{n^*} p(y_i | \eta_i^{\vartheta_1}, \dots, \eta_i^{\vartheta_m}) \int \dots \int G d\boldsymbol{\xi}^{\vartheta_1} \dots d\boldsymbol{\xi}^{\vartheta_m} d\mathbf{b}^{\vartheta_1} \dots d\mathbf{b}^{\vartheta_m} d(\boldsymbol{\tau}^2)^{\vartheta_1} \dots d(\boldsymbol{\tau}^2)^{\vartheta_m}. \end{aligned}$$

The integral over G corresponds to G_3 in (A.17) of Sun et al. (2001). Hence, the posterior is proper if and only if the integral over G is finite. It can be bounded by some constant $\tilde{M}$ using the expressions (A.25) and (A.27) from Sun et al. (2001). This leads to the inequality

$$\begin{aligned} p(\tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m}, (\boldsymbol{\tau}^2)^{\vartheta_1}, \dots, (\boldsymbol{\tau}^2)^{\vartheta_m}, (\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m} | \mathbf{y}) \\ \leq \tilde{M} \prod_{i=1}^{n^*} p(y_i | \eta_i^{\vartheta_1}, \dots, \eta_i^{\vartheta_m}) g((\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m}) \end{aligned}$$

with

$$g((\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m}) = \prod_{l=1}^m ((\tau_0^2)^{\vartheta_l})^{-(\tilde{k}_0 - r^{\vartheta_l})/2 - a_0^- - \sum_{j=1}^{J^{\vartheta_l}} (a_j^-)^{\vartheta_l}} \exp\left(-\frac{\text{SSE}_s^{\vartheta_l} + 2b_0^{\vartheta_l}}{2(\tau_0^2)^{\vartheta_l}}\right)$$

and $\text{SSE}_s^{\vartheta_l}$ defined in (14). Conditions (C.1)–(C.4) imply

$$\int \dots \int g((\tau_0^2)^{\vartheta_1}, \dots, (\tau_0^2)^{\vartheta_m}) d(\tau_0^2)^{\vartheta_1} \dots d(\tau_0^2)^{\vartheta_m} < \infty$$

and thus that for some constant C

$$p(\tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m} | \mathbf{y}) \leq C \prod_{i=1}^{n^*} p(y_i | \eta_i^{\vartheta_1}, \dots, \eta_i^{\vartheta_m})$$

holds. Finally, it remains to show

$$\int \dots \int p(\tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m} | \mathbf{y}) d\tilde{\boldsymbol{\eta}}_s^{\vartheta_1} \dots d\tilde{\boldsymbol{\eta}}_s^{\vartheta_m} < \infty.$$

This follows with Condition (A.1) and by using the relation $\tilde{\boldsymbol{\eta}}_s^{\vartheta_l} = \mathbf{M}^{\vartheta_l} \boldsymbol{\eta}_s^{\vartheta_l}$:

$$\begin{aligned} & \int \dots \int p(\tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m} | \mathbf{y}) d\tilde{\boldsymbol{\eta}}_s^{\vartheta_1} \dots d\tilde{\boldsymbol{\eta}}_s^{\vartheta_m} \\ &= \int \dots \int p(\tilde{\boldsymbol{\eta}}_s^{\vartheta_1}, \dots, \tilde{\boldsymbol{\eta}}_s^{\vartheta_m} | \mathbf{y}) \det(\mathbf{M}^{\vartheta_1})^{-1} d\boldsymbol{\eta}_s^{\vartheta_1} \dots \det(\mathbf{M}^{\vartheta_m})^{-1} d\boldsymbol{\eta}_s^{\vartheta_m} \\ &\leq K \int \dots \int \prod_{i=1}^{n^*} p(y_i | \eta_i^{\vartheta_1}, \dots, \eta_i^{\vartheta_m}) d\eta_i^{\vartheta_1} \dots d\eta_i^{\vartheta_m} < \infty \end{aligned}$$

for some constant $K < \infty$. □

Figure E1: Additive ZIP model. Posterior means (dot-dashed lines) and penalised ML estimates (grey solid lines) together with 95% pointwise credible intervals (dotted lines) obtained from MCMC. The black solid lines represent the true functions. All estimates are averages over 250 simulation replications.

E Additional Figures and Further Simulation Studies

E.1 Additional Figures for the Additive ZIP Model (Section 4.1)

Figure E1 shows posterior means obtained with MCMC and penalised ML estimates averaged over all 250 replications in the ZIP model considered in Section 4.1 of the main paper. Figures E2 and E3 show estimates of those replications corresponding to the 10% and 90% quantile of MSEs, i.e. the 10% best and 10% worst estimation result obtained. Each graph additionally comprises the corresponding 95% pointwise credible interval for the MCMC estimates.

Figure E2: Additive ZIP model. Posterior means (dot-dashed lines) and penalised ML estimates (grey solid lines) together with 95% pointwise credible intervals (dotted lines) obtained from MCMC. The black solid lines represent the true functions. Estimates were chosen according to the 10% quantile of the MSE distribution obtained from 250 simulation replications.

Figure E3: Additive ZIP model. Posterior means (dot-dashed lines) and penalised ML estimates (grey solid lines) together with 95% pointwise credible intervals (dotted lines) obtained from MCMC. The black solid lines represent the true functions. Estimates were chosen according to the 90% quantile of the MSE distribution obtained from 250 simulation replications.

E.2 Other Additive Models

E.2.1 Negative Binomial Model

We use the same study design as described in Section 4 of the main paper. Hence, each of the predictors η^μ, η^δ introduced in Section 2.1 and linked to the distributional parameters μ and δ of a negative binomial distribution, is written as the sum of two nonlinear functions f_1 and f_2 :

$$\begin{aligned} f_1^\mu(x_1) &= f_1^\mu(x_1) = \log(x_1), & f_2^\mu(x_2) &= f_2^\mu(x_2) = 0.3x_2 \cos(x_2) \\ f_1^\delta(x_1) &= 0.1 \exp(0.5x_1), & f_2^\delta(x_2) &= -0.5 \operatorname{arcsinh}(x_2) \end{aligned}$$

with covariates x_1, x_2 as defined in Section 4.1. We simulated 250 replications for sample size $n = 1,000$ and restrict presentations of results to our MCMC approach. In Figure E11 the mean squared errors are summarised in terms of boxplots. Figure E4 shows posterior mean estimates together with 95% posterior credible intervals obtained from averages over the 250 replications. Figures E5 and E6 visualise posterior mean estimates together with 95% posterior credible intervals of the replication corresponding to the 10% and 90% quantile of the MSEs.

Similar as for the simulations conducted for the ZIP model, results for the NB model can briefly be summarised as follows:

- (1) *Bias*. On average, we obtain satisfactory estimates for all nonlinear effects. The effects on δ are more difficult to estimate compared to the effects on μ .
- (2) *MSE*. The logarithmic mean squared errors in Figure E11 of the effects on δ are higher compared to the ones of μ .
- (3) *Coverage rates (not shown)*. For the effects on δ and f_1^μ , the coverage rates tend to be higher than the required levels. This can be an indicator for slightly too conservative credible intervals.

In conclusion, it is a greater challenge to estimate the overdispersion parameter compared to the expectation of the count process. However, since the levels of stipulated coverage are achieved in most cases, we can expect that those intervals get wider in areas where effects are difficult to estimate and hence that they are a good indicator for the statistical uncertainty.

Figure E4: Additive NB model. Posterior means (dot-dashed lines) together with 95% point-wise credible intervals (dotted lines) for $n = 1,000$ observations. The black solid lines represent the true functions. Estimates are averages over 250 simulation replications.

Figure E5: Additive NB model. Posterior means (dot-dashed lines) together with 95% point-wise credible intervals (dotted lines) for $n = 1,000$ observations. The black solid lines represent the true functions. Estimates were chosen according to the 10% quantile of the MSE distribution obtained from 250 simulation replications.

Figure E6: Additive NB model. Posterior means (dot-dashed lines) together with 95% point-wise credible intervals (dotted lines) for $n = 1,000$ observations. The black solid lines represent the true functions. Estimates were chosen according to the 90% quantile of the MSE distribution obtained from 250 simulation replications.

E.2.2 Zero-inflated Negative Binomial Model

This section contains figures for the additive ZINB model introduced in the main paper Section 4.1. Figure E7 shows posterior mean estimates together with 95% posterior credible intervals obtained from averages over the 250 replications. Figures E8 and E9 visualise posterior mean estimates together with 95% posterior credible intervals of the replications corresponding to the 10% and 90% quantile of the MSEs. In Figure E12 the mean squared errors are summarised in terms of boxplots.

(1) *Bias*. Averaging all 250 replications leads to mean estimates that are very close to the true function for sample size $n = 1,000$. From Figure E9, it can be concluded that at least in 90% of the 250 replications the general form of all simulated nonlinear effects is identified quite well. However, the curvatures of the effects on π and δ cannot be captured completely.

(2) *MSE*. As expected (compare also Section E.2.3) the mean squared error is reduced by increasing the sample size. Similar to the ZIP model, it is notable that the expectation μ of the underlying count process is easier to estimate than the probability of additional zeros. The same is found for the overdispersion parameter δ .

(3) *Coverage rates (not shown)*. Coverage rates indicate reliable credible intervals.

Figure E7: Additive ZINB model. Posterior means (dot-dashed lines) together with 95% pointwise credible intervals (dotted lines) for $n = 1,000$ observations. The black solid lines represent the true functions. Estimates are averages over 250 simulation replications.

Figure E8: Additive ZINB model. Posterior means (dot-dashed lines) together with 95% pointwise credible intervals (dotted lines) for $n = 1,000$ observations. The black solid lines represent the true functions. Estimates were chosen according to the 10% quantile of the MSE distribution obtained from 250 simulation replications.

Figure E9: Additive ZINB model. Posterior means (dot-dashed lines) together with 95% pointwise credible intervals (dotted lines) for $n = 1,000$ observations. The black solid lines represent the true functions. Estimates were chosen according to the 90% quantile of the MSE distribution obtained from 250 simulation replications.

E.2.3 Impact of Small Sample Sizes

In this section, we study the effects of decreasing the sample sizes compared to the ones in the main paper Section 4 in which we used $n = 1000$. The simulation set-up is identical to the one given in Section 4.1 despite the fact that we additionally consider $n = 250, 500$ and 750 observations. Figures E10, E11 and E12 show logarithmic mean squared errors of all effects predicted on equidistant grids from the intervals $[1, 6]$ and $[-3, 3]$ for the ZIP, NB and ZINB model, respectively. While no problems for the ZIP and NB models occurred, we faced numerical problems that led to an abnormal program termination of our implementation in 75, 12 and 1 cases for the ZINB model for sample sizes $n = 250, 500$ and 750 , respectively. The numbers of parameters to be estimated is 90 in the ZIP and NB model and 135 in the ZINB model. Table E1 displays median effective sample sizes over all replications.

Important results from our simulations are:

- (1) For all three distributions, estimating complex structures of the mean parameter λ or μ is not problematic, even with a sample size of only $n = 250$.
- (2) MSE and effective MCMC sample sizes indicate that a sample size of around 500 observations is desirable in our NB study with two nonlinear effects impacting the overdispersion parameter δ .
- (3) For estimating a rather complex ZINB model, we find that 250 observations is definitely too small to achieve reliable results. In our study, we would propose to have at least 750 observations.

Figure E10: Additive ZIP model. Comparison of $\log(\text{MSE})$ for $n = 250, 500, 750$ and $1,000$ observations based on posterior mean estimates.

Figure E11: Additive NB model. Comparison of $\log(\text{MSE})$ for $n = 250, 500, 750$ and $1,000$ observations based on posterior mean estimates.

Figure E12: Additive ZINB model. Comparison of $\log(\text{MSE})$ for $n = 250, 500, 750$ and $1,000$ observations based on posterior mean estimates.

Model	$const^{\lambda/\mu}$	f_1^λ/f_1^μ	f_2^λ/f_2^μ	$const^\pi$	f_1^π	f_2^π	$const^\delta$	f_1^δ	f_2^δ
ZIP_250	814	1000	1000	814	841	777	-	-	-
ZIP_500	1000	1000	1000	1000	887	947	-	-	-
ZIP_750	1000	1000	1000	1000	1000	1000	-	-	-
ZIP_1000	1000	1000	1000	1000	1000	1000	-	-	-
NB_250	1000	1000	1000	-	-	-	332	591	586
NB_500	1000	1000	1000	-	-	-	908	862	833
NB_750	1000	1000	1000	-	-	-	1000	888	901
NB_1000	1000	1000	1000	-	-	-	1000	1000	977
ZINB_250	248	471	738	140	174	289	97	203	214
ZINB_500	373	602	877	259	417	538	176	309	465
ZINB_750	437	672	915	339	595	622	306	422	624
ZINB_1000	594	804	1000	512	737	760	497	602	745

Table E1: Effective MCMC sample sizes (12,000 iterations, thinning parameter 10, burnin 2,000) in the ZIP, NB and ZINB model for $n = 250, 500, 750$ and 1,000 observations. The documented values are rounded medians over all replications for the constants and for nonlinear effects the medians of the whole parameter vector of length $d = 22$ in all replications.

E.2.4 Identifiability

In this section, we study the problem of potential identifiability problems when trying to separate effects between the expectation λ of the count data process and the probability of structural zeros π in the ZIP model. We consider two relatively simple situations in which one covariate x was sampled uniformly from the interval $[1, 6]$ and impacts either λ or π nonlinearly, simulated as $f(x) = \sin(x)$. If λ is constant, we choose $\lambda = 1$ and if π is constant we choose $\pi = 0.5$. For the four different sample sizes $n = 250, 500, 750, 1,000$ and 250 replications each, we estimated the true model and a misspecified model where x is included in both predictors. Results are compared by means of logarithmic mean squared errors (compare Figures E13 and E14) and average posterior means together with average 95% credible intervals for sample size $n = 500$ in Figures E15 and E16.

The results can be summarised as follows:

- (1) The MSEs of the informative effects are only impaired very marginally if an irrelevant effect is included in the other predictor.
- (2) The uncertainty of estimates raises when misspecified models are estimated.
- (3) As expected, an increasing sample size reduces the MSEs, standard deviations and uncertainties of estimated effects. This can be confirmed by looking at single replications (not shown) instead of averages over all replications.
- (4) A model where x enters only π is more difficult to estimate than a model with constant π .
- (5) For all sample sizes, the differences in DIC values between the true and the misspecified model are between -10 and 10 in most of the replications. Furthermore, the irrelevant effect is insignificant (simultaneous 95% credible interval contains the zero line) in at least 246 of 250 replications for all sample sizes. Based on the model decision rule discussed in Section 5, this means that in more than 99% we would select the true model when using the DIC for model comparison.

Figure E13: ZIP model. Effect of including the irrelevant covariate x in η^π on $\log(\text{MSE})$ for $n = 250, 500, 750$ and $1,000$ observations based on posterior mean estimates. Left (white) panels show the misspecified models and right (grey) panels correspond to the true models.

Figure E14: ZIP model. Effect of including the irrelevant covariate x in η^λ on $\log(\text{MSE})$ for $n = 250, 500, 750$ and $1,000$ observations based on posterior mean estimates. Left (white) panels show the misspecified models and right (grey) panels correspond to the true models.

Figure E15: ZIP model. Effect of including the irrelevant covariate x in η^π for $n = 500$ observations based on posterior mean estimates and 95% credible intervals, averaged over 250 replications. Left panels show the misspecified model while the upper right panel corresponds to the true model.

Figure E16: ZIP model. Effect of including the irrelevant covariate x in η^λ for $n = 500$ observations based on posterior mean estimates and 95% credible intervals, averaged over 250 replications. Left panels show the misspecified model while the lower right panel corresponds to the true model.

E.2.5 Sensitivity with Respect to Prior Choices and Hyperparameters

To investigate how sensitive additive models for zero-inflated count data are with respect to the prior settings for the smoothing variances, we re-estimated the additive ZIP model (compare Section 4.1 of the main paper) with different hyperparameters of the inverse gamma distribution. Additionally to our standard choice $a_j = b_j = 0.001$, we tried $a_j = b_j = 0.01$, $a_j = b_j = 0.0001$, $a_j = b_j = 0.00001$, $a_j = b_j = 0.000001$, 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 and on the variances, respectively. As another alternative, we consider half-Cauchy priors for the standard deviations τ_j , proposed in Gelman (2006).

Figure E17 shows logarithmic MSEs of the four nonlinear effects f_1^λ , f_2^λ , f_1^π and f_2^π for the different prior settings. We also look at the posterior mean estimates for f_2^λ together with 95% credible intervals in Figure E18 and for f_1^π in Figure E19, each corresponding to one specific replication selected as the median of the MSEs. From the latter one we conclude that at least in this study the prior plays a minor role for the resulting estimates. For π , we observed that differences are somewhat higher but still the qualitative conclusions remain unchanged.

Figure E17: Additive ZIP model. $\log(\text{MSE})$ of posterior mean estimates for different prior specifications. Dark grey boxes correspond to uniform priors for variances and standard deviations, the white boxes show $\log(\text{MSE})$ under half-Cauchy priors, the light grey boxes show our standard choice $IG(a_j = 0.001, b_j = 0.001)$, and the remaining grey boxes further values for a_j, b_j .

Figure E18: Additive ZIP model. Posterior mean estimates of f_2^λ together with 95% credible intervals in one replication r corresponding to the median of MSEs for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, as well as a half-Cauchy prior for the standard deviations.

Figure E19: Additive ZIP model. Posterior mean estimates of f_1^π together with 95% credible intervals in one replication r corresponding to the median of MSEs for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, as well as a half-Cauchy prior for the standard deviations..

Figure E20: Geoadditive ZIP model. Simulated complete spatial effects on the distributional parameters λ (left) and π (right).

E.3 Geoadditive Models

E.3.1 Additional Figures for the Geoadditive ZIP Model (Section 4.2)

Figure E20 depicts the true simulated complete spatial effect of the geoadditive ZIP model discussed in Section 4.2 of the main paper. We compare these true effects with the posterior mean estimates by means of kernel densities in Figure E21 for two selected sample sizes.

Figure E21: Geoadditive ZIP model. Kernel density estimates of posterior mean complete spatial effects (dashed lines) in comparison with the true effects (solid lines). The first row corresponds to λ and the second row to π . Sample sizes $n = 1,000$ and $n = 4,000$ are shown in the first and second column, respectively.

E.3.2 Negative Binomial Model

In order to extend the additive model with a spatial effect, we proceed as in the ZIP model. The nonlinear effects remain the ones of Section E.2.1. Both parameters μ and δ are simulated with an additional spatial effect, consisting of an unstructured part, generated by $\varepsilon \sim N(0, 1/16)$ and a structured part, modelled by a Markov random field and simulated as

$$\begin{aligned} f_{\text{spat}}^{\mu}(s) &= \sin(x_s^c y_s^c) + \varepsilon_s^{\mu} \\ f_{\text{spat}}^{\delta}(s) &= 0.5x_s^c y_s^c + \varepsilon_s^{\delta}. \end{aligned}$$

Again, $s \in \{1, \dots, S\}$ describes one of the $S = 327$ districts in Western Germany, compare Section 4.2 of the main paper.

In Figure E22, the results for the simulated complete spatial effects are visualised. As before, we performed studies for four different sample sizes ($n = 1,000$, $n = 2,000$ and $n = 4,000$), each consisting of 250 replications. Again, we restrict ourselves to a summary for sample sizes $n = 1,000$ and $n = 4,000$. To sum up the performed studies, we present the estimated complete spatial effect compared to the true simulated effect for the two selected sample sizes in Figure E24. Besides this, the $\log(\text{MSE})$ in Figure E23 and the kernel estimates of the complete spatial effects in Figure E25 give further information about the quality of the estimates. The conclusions from this study are comparable to the ones we already presented for the corresponding ZIP model. Effects on μ are easier to estimate with respect to the computed MSE. An increase in the sample size improves our estimates. The complete spatial effect on δ is underestimated and too smooth but with a sample size of $n = 4,000$ it gets much closer to the true effect.

Figure E22: Geoadditive NB model. Simulated complete spatial effects on the ditributional parameters μ (left) and δ (right).

Figure E23: Geoadditive NB model. $\log(\text{MSE})$ of nonlinear and spatial effects based on posterior means for sample sizes $n = 1,000$ and $n = 4,000$.

Figure E24: Geoadditive NB model. Estimated complete posterior mean spatial effects (y-axes) compared to true effects (x-axes). The first row corresponds to μ , the second row to δ . Sample sizes $n = 1,000$ and $n = 4,000$ are shown in the first and second column, respectively.

Figure E25: Geoadditive NB model. Kernel density estimates of posterior mean complete spatial effects (dashed lines) in comparison with the true effects (solid lines). The first row corresponds to μ and the second row to δ . Sample sizes $n = 1,000$ and $n = 4,000$ are shown in the first and second column, respectively.

Figure E26: Geoadditive ZINB model. $\log(\text{MSE})$ of nonlinear and spatial effects based on posterior means for sample sizes $n = 1,000$ and $n = 4,000$.

E.3.3 Zero-Inflated Negative Binomial Model

For the description of the data generating process, compare Section 4.2 of the main paper. The resulting complete spatial effects for all three model parameters are plotted in Figure E27.

Our results of the study are summarised in Figures E26, E28 and E29.

Figure E27: Geoadditive ZINB model. Simulated complete spatial effects on the ditributional parameters μ (topleft), π (topright) and δ (bottom).

Figure E28: Geoadditive ZINB model. Estimated complete posterior mean spatial effects (y-axes) compared to true effects (x-axes). The first row corresponds to μ , the second row to π and the third row to δ . Sample sizes $n = 1,000$ and $n = 4,000$ are shown in the first and second column, respectively.

Figure E29: Geoadditive ZINB model. Kernel density estimates of posterior mean complete spatial effects (dashed lines) in comparison with the true effects (solid lines). The first row corresponds to μ , the second row to π and the third row to δ . Sample sizes $n = 1,000$ and $n = 4,000$ are shown in the first and second column, respectively.

E.4 Impact of Prior Specifications in the Geoadditive Models

Similar to the additive model in Section E.2.5 we re-estimated the geoadditive ZIP model for sample size $n = 1,000$ with alternative prior specifications. Additionally to our standard choice, an inverse gamma prior with parameters $a_j = b_j = 0.001$, we tried $a_j = b_j = 0.01$, $a_j = b_j = 0.0001$, $a_j = b_j = 0.00001$, $a_j = b_j = 0.000001$, as well as $a_j = -0.5, b_j = 0$ and $a_j = -1, b_j = 0$, that is uniform priors on the standard deviations and on the variances. We also estimated the models with half-Cauchy priors for the standard deviations. Figures E30 and E31 show posterior mean estimates of complete spatial effects compared to the true counterparts.

The following findings have been made:

- (1) Similar to the additive models in Section E.2.5 the nonlinear effects are hardly affected by any of the prior choices.
- (2) For the spatial effects, similar results hold as long as a_j, b_j do not get too small. More precisely, for $a = b < 0.0001$ the MCMC iterations has to be increased to overcome convergence problems. Thus we do not recommend small values $a_j = b_j = 0.00001$ in hierarchical situations.
- (3) Uniform priors on both standard deviations and variances worked without severe mixing problems in all effects. However, under a uniform prior for the variances we find a relatively high variability between true and estimated spatial effect on π .
- (4) Under a half-cauchy prior the spatial effect on π is estimated much too small compared to the true effect.
- (5) The posterior mean spatial effect on λ in contrast is very robust against different priors.

One reason for the (limited) prior sensitivity found for the spatial effect may be the inclusion of a hierarchical i.i.d. random effect which corresponds to an information with only relatively scarce information on the effect to be estimated. Still, prior sensitivity was only an issue in very specific situations and in general the impact of the prior choice was rather limited.

Figure E30: Geoadditive ZIP model. Posterior mean complete spatial effect on λ (y-axes) compared to the true effect (x-axes) for $n = 1,000$ and for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, as well as a half-Cauchy prior for the standard deviations.

Figure E31: Geoadditive ZIP model. Posterior mean complete spatial effect on π (y-axes) compared to the true effect (x-axes) for $n = 1,000$ and different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, as well as a half-Cauchy prior for the standard deviations.

F Model Choice via the Deviance Information Criterion

The deviance information criterion (DIC) (Spiegelhalter et al., 2002) is used very often for model selection in hierarchical Bayesian models. However, it is difficult to say what would constitute an important difference in DIC. Very roughly, we make use of the rule of thumb where differences of more than 10 in DIC "definitely" rule out the model with the lower DIC, differences between 5 and 10 are "substantial" but not definite, whereas differences less than 5 support neither model. In this case, models with very different estimates, should not only be selected by choosing the one with the lowest DIC.

Of course, these settings only provide a rule of thumb and applied researchers can certainly deviate from it. Our intention is to enforce the selection of sparse and parsimonious models and therefore we require a rather large difference in DICs. In the applications, we assist the model choice by considering simultaneous credible bands when differences between two models fall within the region of indifference (difference in DICs between 5 and 10). In the applications documented in Sections 5 and 6 of the main paper, we further illustrate and discuss the application of the rule of thumb for model selection.

F.1 Simulation Setup

To study the performance of our rule of thumb for choosing an appropriate model based on DIC, the true models are defined as in Section 4. This means that a generic predictor η linked to any parameter of the ZIP or ZINB distribution is given by

$$\begin{aligned}\boldsymbol{\eta} &= \boldsymbol{f}_1 + \boldsymbol{f}_2 + \boldsymbol{f}_{\text{spat}} \\ &= \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}_{\text{spat}}\boldsymbol{\beta}_{\text{spat}}.\end{aligned}$$

Depending on the chosen model, we used the functions

$$\begin{aligned}f_1^\lambda(x_1) &= f_1^\mu(x_1) = \log(x_1), & f_2^\lambda(x_2) &= f_2^\mu(x_2) = 0.3x_2 \cos(x_2) \\ f_1^\pi(x_1) &= \sin(x_1), & f_2^\pi(x_2) &= -0.2x_2^2 \\ f_1^\delta(x_1) &= 0.1 \exp(0.5x_1), & f_2^\delta(x_2) &= 0.5 \operatorname{arcsinh}(x_2)\end{aligned}$$

Model	Description
ZIP_M1	simulated model
ZIP_M2	including an irrelevant linear effect in η^λ
ZIP_M3	including an irrelevant nonlinear effect in η^π
ZIP_M4	excluding the relevant complete spatial effect in η^π
ZIP_M5	excluding the relevant complete spatial effect in η^λ
ZIP_M6	including the relevant nonlinear effect of x_1 in η^π linearly
NB_M1	simulated model
NB_M2	including an irrelevant linear effect in η^μ
NB_M3	including an irrelevant nonlinear effect in η^δ
NB_M4	excluding the relevant complete spatial effect in η^δ
NB_M5	excluding the relevant complete spatial effect in η^μ
NB_M6	including the relevant nonlinear effect of x_1 in η^δ linearly
ZINB_M1	simulated model
ZINB_M2	including an irrelevant linear effect in η^μ
ZINB_M3	including an irrelevant nonlinear effect in η^δ
ZINB_M4	excluding the relevant complete spatial effect in η^μ
ZINB_M5	excluding the relevant complete spatial effect in η^δ
ZINB_M6	including the relevant nonlinear effect of x_1 in η^μ linearly

Table F2: Model choice based on DIC. Description of estimated models.

where the covariates x_1 and x_2 are obtained as i.i.d. samples, such that for $i = 1, \dots, n$, we have $x_{i1} \in [1, 6]$ and $x_{i2} \in [-3, 3]$. The complete spatial effect is obtained as

$$\begin{aligned}
f_{\text{spat}}^\lambda(s) &= f_{\text{spat}}^\mu(l) = \sin(x_s^c y_s^c) + \varepsilon_s \\
f_{\text{spat}}^\pi(s) &= \sin(x_s^c) \cos(0.5 y_s^c) + \varepsilon_s^\pi \\
f_{\text{spat}}^\delta(s) &= 0.5 x_s^c y_s^c + \varepsilon_s^\delta.
\end{aligned}$$

and ε^λ , ε^μ , ε^π , ε^δ are Gaussian with mean 0 and variance 1/16.

To validate the ability of the DIC to find the best model in geoaddditive regression specifications for responses with ZIP, NB or ZINB distribution, we re-estimated different misspecified versions of each of the three simulated models, compare Table F2 for a scheme of all tested candidates. All models are replicated 250 times for sample sizes $n = 1,000$ and $n = 4,000$.

Model comparison	$\Delta > 10$	$ \Delta < 10$	$\Delta < -10$
ZIP_M1 vs ZIP_M2	0%/0%	100%/100%	0%/0%
ZIP_M1 vs ZIP_M3	0%/0%	100%/100%	0%/0%
ZIP_M1 vs ZIP_M4	64%/100%	36%/0%	0%/0%
ZIP_M1 vs ZIP_M5	100%/100%	0%/0%	0%/0%
ZIP_M1 vs ZIP_M6	19.2%/98.8%	80.8%/1.2%	0%/0%
NB_M1 vs NB_M2	0%/0%	100%/100%	0%/0%
NB_M1 vs NB_M3	0%/0%	99.6%/100%	0.4%/0%
NB_M1 vs NB_M4	62.8%/100%	37.2%/0%	0%/0%
NB_M1 vs NB_M5	100%/100%	0%/0%	0%/0%
NB_M1 vs NB_M6	8.8%/62%	91.2%/38%	0%/0%
ZINB_M1 vs ZINB_M2	2.8%/0%	92.8%/100%	4.4%/0%
ZINB_M1 vs ZINB_M3	0%/0%	100%/100%	0%/0%
ZINB_M1 vs ZINB_M4	98.4%/100%	1.6%/0%	0%/0%
ZINB_M1 vs ZINB_M5	20.4%/75.6%	79.6%/24.4%	0%/0%
ZINB_M1 vs ZINB_M6	98.8%/100%	1.2%/0%	0%/0%

Table F3: Model choice based on DIC. Percentage of replications with certain DIC differences for $n = 1,000$ and $4,000$ in the ZIP, NB and ZINB model. Δ is defined as the difference between the DIC of the true model and the DIC of the misspecified model.

F.2 Results

Table F3 displays the percentage of replications in which differences between the DIC of the true and the misspecified models (Δ) are greater than 10 (true model would be selected), smaller than -10 (misspecified model would be selected) and replications in which the DIC would be indecisive.

For all ZIP models, the false model is never selected by DIC with the selection criterion defined before. In case of the NB model, there is one of 250 replications in which for sample size $n = 1,000$ the difference in DIC is in the range of 15 and in favour of the false model if an irrelevant nonlinear effect is added. In case of the ZINB model, there are 11 such cases. One explanation is that the sample size $n = 1,000$ is relatively small compared to the number of parameters that have to be estimated. Another reason might be that the additional nonlinear effect was simulated in the interval $[-1, 1]$ which is similar to the true covariate values.

Furthermore, we can draw the following conclusions:

(1) *Including irrelevant effects (models M2 and M3).* In all models for the different distributions (except of the cases mentioned above) adding an additional, irrelevant

effect in the predictor of one of the model parameters yields no DIC differences greater than 10 such that the DIC is indecisive between the two models. However, one can argue that in such a case one would either decide for the sparse model which is the true model or one would usually try to take another criterion into account. In particular, a covariate can be seen as irrelevant or not significant if the corresponding Bayesian credible interval contains zero. For nonlinear effects we choose the simultaneous 95% credible interval constructed according to the proposal of Krivobokova et al. (2010). Applying this rule, the irrelevant effect would be excluded (and hence a decision in favour of the true model would be made) in at least 92% of all tested models and for all distributions where an irrelevant variable was included.

(2) *Omitted relevant effects (models M_4 and M_5).* Here we have to distinguish between the different sample sizes $n = 1,000$ and $n = 4,000$. In the latter case, the DICs have a difference greater than 10 in favour of the true model in at least 75.6% of the replications (such that at least in these we would choose the correct model). In the ZIP and NB model, the true model would even be selected in 100%. For a smaller sample size of $n = 1,000$, it depends on the parameter where an effect is omitted. Again we observe that estimating the mean parameter of the count process is easier compared to possible further distribution parameters. But in at least 20.4% of all these models we would prefer the true model whereas in the remaining cases the DIC is indecisive.

(3) *Wrong specification (model M_6).* We also simulated one model for each distribution where in one of the parameter specifications a nonlinear effect was included linearly. Depending on the parameter and the distribution in some or most cases, the true model would be selected (but never the misspecified one).

Figure F32: Geoadditive ZIP model. Posterior means (dotdashed lines) for f_1^λ together 95% credible intervals (dotted lines) for $n = 1,000$ observations in models M1 to M6. The solid lines are the true function. Estimates are averages over 250 simulation replications.

F.3 Effects of Misspecified Models on Estimates

In this section, we consider how effects suffer in the models of Table F2 and restrict our presentation to the ZIP model with sample size $n = 1,000$. Figures F32 to F35 show posterior mean estimates together with 95% posterior probabilities of the four nonlinear effects on λ and π in models M1 to M6. The estimated complete spatial effect on λ and π in the true and all misspecified models can be compared to the true effect in Figures F36 and F37. From these estimates, we can conclude the following:

- (1) Estimates of nonlinear effects are well identified, even in misspecified models.
- (2) The most difficult parameter in our models is the spatial effect on π when the relevant spatial effect on λ is excluded (compare M5 in Figure F36).

Figure F33: Geoadditive ZIP model. Posterior means (dotdashed lines) for f_2^λ together 95% credible intervals (dotted lines) for $n = 1,000$ observations in models M1 to M6. The solid lines are the true function. Estimates are averages over 250 simulation replications.

Figure F34: Geoadditive ZIP model. Posterior means (dotdashed lines) for f_1^π together 95% credible intervals (dotted lines) for $n = 1,000$ observations in models M1 to M6. The solid lines are the true function. Estimates are averages over 250 simulation replications.

Figure F35: Geoadditive ZIP model. Posterior means (dotdashed lines) for f_2^π together 95% credible intervals (dotted lines) for $n = 1,000$ observations in models M1 to M6. The solid lines are the true function. Estimates are averages over 250 simulation replications.

Figure F36: Geoadditive ZIP model. Posterior mean complete spatial effect on λ (y-axes) compared to true effect (x-axes) for $n = 1,000$ observations in models M1 to M6.

Figure F37: Geoaddditive ZIP model. Posterior mean complete spatial effect on π (y-axes) compared to true effect (x-axes) for $n = 1,000$ observations in models M1 to M6.

	Continuous covariates			
Variable	Description	Mean	Std-dev.	Min/max
<i>year</i>	grant year	1991	3.94	1980/1997
<i>ncountry</i>	number of designated states	7.77	4.12	1/17
<i>nclaims</i>	number of claims	12.33	8.13	1/50
	Binary covariates			
Variable	Description	Frequency in %		
<i>opp</i>	patent opposition	yes = 1/no = -1	41.2/58.9	
<i>biopharm</i>	patent from biotech/pharma sector	yes = 1/no = -1	43.9/56.1	
<i>patus</i>	patent holder from US	yes = 1/no = -1	33.3/66.7	
<i>patgsgr</i>	patent holder from Germany, Switzerland or Great Britain	yes = 1/no = -1	23.7/76.3	

Table G4: Patent citations. Description of explanatory variables in the data set of 4,805 observations.

G Supplementary Material: Patent Citations (Section 5)

Table G4 gives a summary of continuous, binary and categorical variables in the patent data set. Figure G38 shows normalised quantile residuals of the remaining models described in Section 5 of the main paper. Predictor specifications for all models of type M1 can be found in Table G5. Table G6 summarises the estimated linear effects in the best model ZINB_M1.

Model	Predictors
Poisson_M1	$\eta^\lambda = \beta_0^\lambda + \beta_1^\lambda opp + \beta_2^\lambda biopharm + \beta_3^\lambda patus + \beta_4^\lambda patgsgr + f_1^\lambda(ncountry)$ $+ f_2^\lambda(year) + f_3^\lambda(nclaims)$
ZIP_M1	$\eta^\lambda = \beta_0^\lambda + \beta_1^\lambda opp + \beta_2^\lambda biopharm + \beta_3^\lambda patus + \beta_4^\lambda patgsgr + f_1^\lambda(ncountry)$ $+ f_2^\lambda(year) + f_3^\lambda(nclaims)$ $\eta^\pi = \beta_0^\pi + \beta_1^\pi opp + \beta_2^\pi biopharm + \beta_3^\pi patus + \beta_4^\pi patgsgr + f_1^\pi(ncountry)$ $+ f_2^\pi(year)$
NB_M1	$\eta^\mu = \beta_0^\mu + \beta_1^\mu opp + \beta_2^\mu biopharm + \beta_3^\mu patus + \beta_4^\mu patgsgr + f_1^\mu(ncountry)$ $+ f_2^\mu(year) + f_3^\mu(nclaims)$ $\eta^\delta = \beta_0^\delta + \beta_1^\delta opp + \beta_2^\delta biopharm + \beta_3^\delta patus + \beta_4^\delta patgsgr + f_1^\delta(ncountry)$ $+ f_2^\delta(year) + f_3^\delta(nclaims)$
ZINB_M1	$\eta^\mu = \beta_0^\mu + \beta_1^\mu opp + \beta_2^\mu biopharm + \beta_3^\mu patus + \beta_4^\mu patgsgr + f_1^\mu(ncountry)$ $+ f_2^\mu(year) + f_3^\mu(nclaims)$ $\eta^\pi = \beta_0^\pi + \beta_1^\pi biopharm + f_1^\pi(ncountry) + \beta_2^\pi(year - 1991)$ $\eta^\delta = \beta_0^\delta + \beta_1^\delta patus + \beta_2^\delta patgsgr$

Table G5: Patent citations. Predictor structures of the models Poisson_M1, ZIP_M1, NB_M1 and ZINB_M1.

varname	pmean	pstd	pqu2p5	pqu10	pmed	pqu90	pqu97p5
$\mu: const$	0.79	0.07	0.66	0.71	0.79	0.88	0.93
$\mu: opp$	0.20	0.02	0.16	0.17	0.20	0.23	0.24
$\mu: biopharm$	0.15	0.03	0.09	0.11	0.15	0.19	0.21
$\mu: patus$	0.002	0.02	-0.05	-0.03	0.00	0.03	0.05
$\mu: patgsgr$	-0.12	0.03	-0.17	-0.16	-0.12	-0.09	-0.07
$\pi: const$	-3.44	0.39	-4.44	-3.92	-3.39	-3.03	-2.82
$\pi: biopharm$	0.67	0.16	0.37	0.48	0.67	0.87	1.01
$\pi: year - 1991$	0.64	0.06	0.52	0.56	0.64	0.72	0.77
$\delta: const$	-0.09	0.07	-0.22	-0.17	-0.09	0.00	0.05
$\delta: patus$	-0.32	0.05	-0.42	-0.38	-0.32	-0.26	-0.23
$\delta: patgsgr$	-0.10	0.06	-0.22	-0.18	-0.10	-0.02	0.03

Table G6: Patent citations. Estimated linear effects in the best performing model ZINB_M1.

Figure G38: Patent citations. Comparison of normalised quantile residuals in linear models (M2) and models with constant π and δ (M3).

prior	pmean	pstd	pqu2p5	pqu10	pmed	pqu90	pqu97p5
a=-1, b=0	0.659	0.067	0.534	0.573	0.658	0.746	0.796
a=-0.5, b=0	0.654	0.067	0.527	0.571	0.650	0.742	0.786
a=0.01, b=0.01	0.654	0.069	0.524	0.566	0.652	0.743	0.801
a=0.001, b=0.001	0.641	0.064	0.519	0.560	0.638	0.724	0.775
a=0.0001, b=0.0001	0.639	0.067	0.511	0.558	0.637	0.724	0.772
a=0.00001, b=0.00001	0.639	0.065	0.513	0.553	0.639	0.724	0.762
a=0.000001, b=0.000001	0.641	0.066	0.520	0.555	0.637	0.728	0.784
half-Cauchy	0.645	0.069	0.516	0.558	0.645	0.738	0.784

Table G7: Patent citations. Estimated linear effect of *year* on π in the best performing model ZINB_M1 for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, as well as a half-Cauchy prior for the standard deviations. Our standard choice is printed in bold.

G.1 Sensitivity Analysis

As in our simulations, we checked how estimates of the final ZINB_M1 model differ if the prior specifications are modified. Figure G39 shows the posterior mean effects of *ncountry* on π together with 80% and 95% posterior intervals under uniform priors for the standard deviations and variances, i.e. $IG(a_j = -0.5, b_j = 0)$, $IG(a_j = -1, b_j = 0)$, as well as for hyperparameters $a_j = b_j = 0.01, \dots, 0.000001$. Additionally, we consider a half-Cauchy prior for the standard deviations. Sampling paths of parameter β_5 are given in Figure G41. For the same priors, Figure G40 shows the posterior mean effects of *ncountry*, *year* and *nclaims* on μ . Table G7 compares the posterior of the linear effect of *year* on π .

Figure G41 gives some exemplary evidence that no mixing problems occurred in all settings. Furthermore, Figures G39 and G40, as well as Table G7 highlight that estimates are very similar irrespective of the chosen prior structure. Two observations attract attention: First, the effect of *nclaims* on μ which is estimated smoother under $IG(a_j = 0.000001, b_j = 0.000001)$ and the half-Cauchy and less smooth under $IG(a_j = 0.01, b_j = 0.01)$ and $a_j = -1, b_j = 0$ compared to the remaining prior choices. Second, the effect of *year* on μ is estimated slightly rougher under a uniform prior for the variances. However, the general results drawn from the estimates would be the same independently of the chosen prior.

Figure G39: Patent citations. Posterior mean estimate of $ncountry$ on π together with point-wise 80% and 95% credible intervals for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, as well as a half-Cauchy prior for the standard deviations.

Figure G40: Patent citations. Posterior mean estimates of $ncountry$, $year$ and $nclaims$ on μ for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, as well as a half-Cauchy prior for the standard deviations.

Figure G41: Patent citations. Sampling paths corresponding to the regression coefficient β_5 of the nonlinear effect *year* on μ for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, as well as a half-Cauchy prior for the standard deviations.

H Supplementary Material: Modelling Claim Frequencies (Section 6)

H.1 Model Comparison

Since the response variable contains a lot of zeros and a limited number of observations with more than one claim, estimating full models with all potential covariates for the remaining parameters (π and/or δ) causes problems in the mixing behaviour, especially in case of the ZINB model. We therefore performed a preliminary variable selection, starting from very simple predictor specifications for π or δ and including step by step effects on the basis of the DIC. In Section F of the supplement, we investigated the performance of the DIC for selecting predictors in zero-inflated and overdispersed count data regression and we basically found that the DIC provides suitable guidance also in this extended model class. Based on results obtained for the ZIP and the NB model, both of which indicate a very good fit for the data as shown by the quantile residuals visualised in Figure H42, we refrained from searching for (even) more complex ZINB models.

Table H8 shows the calculated scores (which have been introduced in Section 5 of the main paper) for the Poisson, ZIP and NB distribution obtained by a ten-fold cross validation. In general, differences are smaller than for the patent application but there is still an indication for additional zero-inflation or overdispersion since the Poisson distribution always yields the smallest score. For the Brier and spherical score, there is some evidence in favour of the ZIP model while the logarithmic score would prefer the NB model. The quantile residuals depicted in Figure H42 tell a similar story and indicate that the Poisson distribution is not able to adequately represent the claim frequency distribution. Both ZIP and NB yield residuals that are very close to the diagonal and therefore provide a very similar fit. For ZIP, there are some deviations from the diagonal line for larger residuals which may hint at additional overdispersion. These deviations may also be responsible for the fact that the logarithmic score favours the NB model, since this score reacts particularly sensitive to predictive problems of extreme (in our case large) observations.

Figure H42: Insurance claims. Comparison of quantile residuals in the Poisson, ZIP, NB model.

Model	Brier Score	Logarithmic Score	Spherical Score
Poisson	-32,260.2	-62,129.3	145,404.0
ZIP	-32,247.4	-61,999.7	145,410.7
NB	-32,252.1	-61,979.2	145,409.2

Table H8: Insurance claims. Summarised scores the three competing models Poisson, ZIP and NB obtained from ten-fold cross validation.

Figure H43: Insurance claims. Posterior mean estimates on λ (centred around zero) together with 80% and 95% pointwise credible intervals. Axes have different scales in order to enhance visibility of estimates.

Figure H44: Insurance claims. Posterior mean estimates of nonlinear effects on λ and π in the geoadditive ZIP model for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, and the half-Cauchy prior on standard deviations.

H.2 Sensitivity Analysis

Figures H44, H45, and H46 show posterior mean estimates of spatial effects and nonlinear effects included in the ZIP model of the main paper. We find that the posterior mean spatial effect on λ is very similar independent of the prior choice. For the effect on π , we observe a higher variability and smoother effects for smaller hyperparameter values a_j and b_j , as well as the half-Cauchy prior. This can be explained by the fact that the effect in general is rather small. The nonlinear effects are very robust against different priors.

Figure H45: Insurance claims. Posterior mean spatial effects on λ in the geoadditive ZIP model for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, and the half-Cauchy prior on standard deviations.

Figure H46: Insurance claims. Posterior mean spatial effects on π in the geoadditive ZIP model for different parameters of the inverse gamma hyperpriors $IG(a_j, b_j)$, uniform priors for variances and standard deviations, and the half-Cauchy prior on standard deviations.

References

- 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.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis* **1**: 515–533.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman & Hall/CRC, New York/Boca Raton.
- Krivobokova, T., Kneib, T. and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators, *Journal of the American Statistical Association* **105**: 852–863.
- Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K. and Kneib, T. (2014). Multilevel structured additive regression, *Statistics and Computing* **24**: 223–238.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*, Chapman & Hall/CRC, New York/Boca Raton.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **64**: 583–639.
- 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.