Klein, Nadja; Kneib, Thomas; Lang, Stefan

Working Paper

Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

Working Papers in Economics and Statistics, No. 2013-12

Provided in Cooperation with:
Institute of Public Finance, University of Innsbruck

Suggested Citation: Klein, Nadja; Kneib, Thomas; Lang, Stefan (2013) : Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data, Working Papers in Economics and Statistics, No. 2013-12, University of Innsbruck, Research Platform Empirical and Experimental Economics (eeecon), Innsbruck

This Version is available at:
https://hdl.handle.net/10419/101068

University of Innsbruck

Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

Nadja Klein, Thomas Kneib, Stefan Lang

Working Papers in Economics and Statistics

2013-12

Research Platform
eeecon
Empirical and Experimental Economics
University of Innsbruck
http://eeecon.uibk.ac.at/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact Address:
University of Innsbruck
Department of Public Finance
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 7171
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://eeecon.uibk.ac.at/wopec/

For a list of recent papers see the backpages of this paper.

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

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

Stefan Lang
Department of Statistics
University of Innsbruck

Abstract

Frequent problems in applied research that prevent 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 special instances, we consider the zero-inflated Poisson, the negative binomial and the zero-inflated negative binomial distribution as standard options for applied work. 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 to quantify the predictive ability of the models. The deviance information criterion is used for further model specification.

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. However, in many applied examples, 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, there is a large fraction of patents that are never cited and this fraction seems to be considerably larger than expected with a Poisson distribution fitted to the data.

- Overdispersion, where 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 limitations, 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 where a binary process decides between observations that are always zero and observations that will be realized from a usual count data distribution such as the Poisson distribution. As a consequence, zeros can either arise from the binary process or from the Poisson distribution. In the application on citations of patents, the binary process distinguishes those patents that are of very little interest and will therefore never be

cited from those that are relevant and for which the number of citations follows, e.g., 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 that allows 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 et al., 2003, Fahrmeir et al., 2004, Wood, 2008] or Markov chain Monte Carlo (MCMC) simulations [Brezger and Lang, 2006, Jullion and Lambert, 2007, Lang et al., 2013]. The framework of generalized additive models for location, scale and shape (GAMLSS) introduced by Rigby and Stasinopoulos [2005] allows to extend 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 for 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 et al., 2004, Brezger and Lang, 2006]. Structured

additive regression relies on a unifying representation of all these model terms based on non-standard basis function specifications in combination with quadratic penalties (in a frequentist formulation) or Gaussian priors (in a Bayesian approach).

In this paper, 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. It therefore considerably extends the set of available predictor specifications for all parameters involved in zero-inflated and overdispersed count data regression.

- 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 a MCMC simulation algorithm based on iteratively weighted least squares approximations to the full conditionals as suggested by Gamerman [1997] or Brezger and Lang [2006] for exponential family regression models. An alternative strategy would be the consideration of random walk proposals as in Jullion and Lambert [2007].

- We provide a numerically efficient implementation comprising also an extension to multilevel structure that is particularly useful in spatial regression specifications or for models including random effects, see Lang et al. [2013]. 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 estimate run and to provide valid confidence 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, the deviance information criterion [Spiegelhalter et al., 2002] and proper scoring rules [Gneiting and Raftery, 2007] to determine the predictive ability. Some rare approaches that develop similar types of models and inferences are already available. For example, Fahrmeir and Osuna Echavarría [2006] develop 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 develop zero-inflated generalized Poisson regression models for count data where the overdispersion and zero-inflation parameters can be fitted by maximum likelihood methods.

There are two packages in R that provide regression for zero-inflated models. In `gamlss` [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 are 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 `gamlss` in extensive simulations in Section 4. Linear predictors can be specified in the package `pscl` [Zeileis et al., 2008] to fit zero-inflated regression models. The parameters are estimated with the function `optim` to maximize the likelihood.

The rest of this paper 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 GAMLSS within a restricted class of purely additive models and for more complex geoadditive models. Sections 5 and 6 provide analyses of the applications on citations of patents and claim frequencies in car insurance. The final Section 7 summarizes our findings and comments on directions of future research.

2 Zero-Inflated Count Data Regression

2.1 Observation Models

We assume that zero-inflated count data y_i as well as covariate information $\boldsymbol{\nu}_i$ have been collected for individuals $i = 1, \ldots, n$. The conditional distribution of y_i given the covariates $\boldsymbol{\nu}_i$ is then described in terms of the density

$$p(y_i|\boldsymbol{\nu}_i) = \pi_i \mathbb{1}_{\{0\}}(y_i) + (1 - \pi_i)\tilde{p}(y_i|\boldsymbol{\nu}_i)$$

that arises from the hierarchical definition of the responses $y_i = \kappa_i \tilde{y}_i$, where κ_i is a binary selection process $\kappa_i \sim B(1 - \pi_i)$ 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 but additional zeros arise whenever the indicator variable κ_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 π_i. From the definition of zero-inflated count data models, we obtain

$$\begin{aligned} \mathrm{E}(y_i|\boldsymbol{\nu}_i) &= (1 - \pi_i)\,\mathrm{E}(\tilde{y}_i|\boldsymbol{\nu}_i) \\ \mathrm{Var}(y_i|\boldsymbol{\nu}_i) &= (1 - \pi_i)\,\mathrm{Var}(\tilde{y}_i|\boldsymbol{\nu}_i) + \pi_i(1 - \pi_i)\left(\mathrm{E}(\tilde{y}_i|\boldsymbol{\nu}_i)\right)^2. \end{aligned} \tag{1}$$

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

$$\tilde{p}(\tilde{y}_i) = \frac{\Gamma(\tilde{y}_i + \delta_i)}{\Gamma(\tilde{y}_i + 1)\Gamma(\delta_i)} \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} \left(\frac{\mu_i}{\delta_i + \mu_i}\right)^{\tilde{y}_i}.$$

The latter choice is particularly suited if 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 as well as 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 $\eta_i^\pi = \mathrm{logit}(\pi_i)$ and $\eta_i^\lambda = \log(\lambda_i)$ whereas for zero-inflated negative binomial (ZINB) regression we assume $\eta_i^\pi = \mathrm{logit}(\pi_i)$, $\eta_i^\mu = \log(\mu_i)$ 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]. Note that in applications we may often observe that modelling 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 from the previous section, we assume a structured additive specification

$$\eta_i = \beta_0 + f_1(\boldsymbol{\nu}_i) + \ldots + f_p(\boldsymbol{\nu}_i)$$

where, for notational simplicity, we drop the parameter index from the predictor and the included effects. While β_0 is an intercept term representing the overall level of the predictor, the generic functions $f_j(\boldsymbol{\nu}_i)$, $j = 1, \ldots, p$, 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

$$f_j(\boldsymbol{\nu}_i) = \sum_{k=1}^{d_j} \beta_{jk} B_{jk}(\boldsymbol{\nu}_i). \tag{2}$$

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 (2) implies that each vector of function evaluations $\boldsymbol{f}_j = (f_j(\boldsymbol{\nu}_1), \ldots, f_j(\boldsymbol{\nu}_n))'$ can be written as $\boldsymbol{Z}_j\boldsymbol{\beta}_j$ where $\boldsymbol{Z}_j$ is the design matrix arising from the evaluations of the basis functions, i.e. $\boldsymbol{Z}_j[i,k] = B_{jk}(\boldsymbol{\nu}_i)$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_j})'$ is the vector of all regression coefficients. Then the predictor vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ can be compactly represented as

$$\boldsymbol{\eta} = \beta_0 \mathbb{1} + \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \ldots + \boldsymbol{Z}_p\boldsymbol{\beta}_p \tag{3}$$

where $\mathbb{1}$ 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 (2), we consider multivariate Gaussian priors

$$p(\boldsymbol{\beta}_j) \propto \left(\frac{1}{\tau_j^2}\right)^{\frac{\text{rk}(\boldsymbol{K}_j)}{2}} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right) \tag{4}$$

for the regression coefficients where τ_j^2 is the smoothing variance determining our prior confidence and $\boldsymbol{K}_j$ is the prior precision matrix implementing prior assumptions about smoothness of the function. Note that $\boldsymbol{K}_j$ may not have full rank and therefore the Gaussian prior will usually be partially improper. A completely improper prior is obtained as a special case for either $\tau_j^2 \to \infty$ or $\boldsymbol{K}_j = \boldsymbol{0}$.

To obtain a data-driven amount of smoothness, we assign inverse gamma hyperpriors $\tau_j^2 \sim \text{IG}(a_j, b_j)$ to smoothing variances with $a_j = b_j = 0.001$ as a default option.

2.4 Special Cases

To make the generic model specification introduced in the previous section more concrete, we compactly summarize some special cases by specifying the basis functions and the prior precision matrices:

- Linear effects $f_j(\boldsymbol{\nu}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_j$ where $\boldsymbol{x}_i$ is a subvector of original covariates: The design matrix is obtained by stacking the rows $\boldsymbol{x}_i$ while usually a non-informative prior with $\boldsymbol{K}_j = \boldsymbol{0}$ is chosen for the regression coefficients $\boldsymbol{\beta}_j$. A ridge-type prior with $\boldsymbol{K}_j = \boldsymbol{I}$ is an alternative especially if the dimension of the vector $\boldsymbol{\beta}_j$ is large.

- P-splines for nonlinear effects $f_j(\boldsymbol{\nu}_i) = f_j(x_i)$ 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 $\boldsymbol{K}_j = \boldsymbol{D}'\boldsymbol{D}$ where $\boldsymbol{D}$ is a difference matrix of appropriate order. Usual default choices are twenty inner knots, cubic B-splines and second order differences, see Lang and Brezger [2004] for details.

- Markov random fields $f_j(\boldsymbol{\nu}_i) = f_j(s_i)$ for a discrete spatial variable $s_i \in \{1, \ldots, S\}$: The design matrix is an indicator matrix connecting individ-

ual observations with corresponding regions, i.e., $\boldsymbol{Z}[i, s]$ is one if observation i belongs to region s and zero otherwise. To implement spatial smoothness, $\boldsymbol{K}_j$ is chosen as an adjacency matrix indicating which regions are neighbors of each others, see Rue and Held [2005] for details.

- Random effects $f_j(\boldsymbol{\nu}_i) = \beta_{g_i}$ based on a grouping variable $g_i \in \{1, \ldots, G\}$: The design matrix is an indicator matrix connecting individual observations with corresponding groups, i.e., $\boldsymbol{Z}[i, g]$ is one if observation i belongs to group g and zero otherwise. To reflect the assumption of i.i.d. random effects, the precision matrix is chosen as $\boldsymbol{K}_j = \boldsymbol{I}$.

A more detailed exposition for the generic structured additive regression specification comprising also bivariate surfaces or varying coefficient terms is provided in Fahrmeir et al. [2004] and Kneib et al. [2009].

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 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 τ_j^2 can be derived in closed form:

$$\tau_j^2|\cdot \sim \text{IG}(a_j', b_j'), \quad a_j' = \frac{\text{rk}(\boldsymbol{K}_j)}{2} + a_j, \quad b_j' = \frac{1}{2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j + b_j. \tag{5}$$

3.1 IWLS Proposals

The basic idea of IWLS proposals is to determine a quadratic approximation of the full conditional that leads to a Gaussian proposal density with expectation and covariance matrix corresponding to the mode and the curvature of the quadratic approximation. To make the description easier, we assume for the moment a model with only one predictor $\boldsymbol{\eta}$ but the principle idea immediately carries over to our multi-predictor

framework since in the MCMC algorithm we are always only working with sub-blocks of coefficients corresponding to one predictor component. Let now $l(\boldsymbol{\eta})$ be the log-likelihood depending on the predictor $\boldsymbol{\eta}$. Then it is easy to verify that the full conditional for a typical parameter block $\boldsymbol{\beta}_j$ is

$$\log(p(\boldsymbol{\beta}_j|\cdot)) \propto l(\boldsymbol{\eta}) - \frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j$$

where $\propto$ is abused to denote equality up to additive constants. The quadratic approximation to this penalized log-likelihood term is then obtained by a Taylor expansion around the mode such that

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

where t indexes the iterations of a Newtons's method type approximation. From this approximation, we can deduce the working model

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

where $\boldsymbol{z} = \boldsymbol{\eta} + \boldsymbol{W}^{-1}\boldsymbol{v}$ is a vector of working observations with the predictor of the given model as expectation, $\boldsymbol{v} = \partial l/\partial \boldsymbol{\eta}$ is the score vector and $\boldsymbol{W}$ are working weight matrices based on a Fisher-scoring approximation, with $w_i = \mathrm{E}(-\partial^2 l/\partial \eta_i^2)$, on the diagonals and zero otherwise. Finally, we obtain that the IWLS proposal distribution for $\boldsymbol{\beta}_j$ is $\mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{P}_j^{-1})$ with expectation and precision matrix

$$\boldsymbol{\mu}_j = \boldsymbol{P}_j^{-1}\boldsymbol{Z}_j'\boldsymbol{W}(\boldsymbol{z} - \boldsymbol{\eta}_{-j}) \qquad \boldsymbol{P}_j = \boldsymbol{Z}_j'\boldsymbol{W}\boldsymbol{Z}_j + \frac{1}{\tau_j^2}\boldsymbol{K}_j, \tag{6}$$

where $\boldsymbol{\eta}_{-j} = \boldsymbol{\eta} - \boldsymbol{Z}_j\boldsymbol{\beta}_j$ 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 vector $\boldsymbol{v}$ and the working weights $\boldsymbol{W}$. For the ZIP model, the elements of the score vectors for the zero-inflation and the Poisson parts of the model are given by

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

and the working weights can be shown to be

$$w_i^{\lambda} = \frac{\lambda_i(1-\pi_i)\left(\pi_i + (1-\pi_i)\exp(-\lambda_i) - \exp(-\lambda_i)\lambda_i\pi_i\right)}{\pi_i + (1-\pi_i)\exp(-\lambda_i)} \tag{7}$$

$$w_i^{\pi} = \frac{\pi_i^2(1-\pi_i)\left(1 - \exp(-\lambda_i)\right)}{\pi_i + (1-\pi_i)\exp(-\lambda_i)} \tag{8}$$

For the ZINB model, we obtain

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

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log(\Gamma(x))$ is the digamma function for $x > 0$, and

$$
w_i^{\mu} = \frac{\delta_i\mu_i(1-\pi_i)}{(\delta_i+\mu_i)} - \frac{\pi_i(1-\pi_i)\delta_i^2\mu_i^2\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}}{\left(\pi_i + (1-\pi_i)\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}\right)(\delta_i+\mu_i)^2} \tag{9}
$$

$$
w_i^{\pi} = \frac{\pi_i^2(1-\pi_i)\left(1 - \left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}\right)}{\pi_i + (1-\pi_i)\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}} \tag{10}
$$

$$
\begin{aligned}
w_i^{\delta} &= -\delta_i(1-\pi_i)\left(\log\left(\frac{\delta_i}{\delta_i+\mu_i}\right) + \frac{\mu_i}{\delta_i+\mu_i}\right) - \delta_i\left(\mathrm{E}(\psi(y_i+\delta_i)) - \psi(\delta_i)\right) \qquad (11) \\
&\quad - \frac{(1-\pi_i)\pi_i\delta_i^2\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}\left(\log\left(\frac{\delta_i}{\delta_i+\mu_i}\right) + \frac{\mu_i}{\delta_i+\mu_i}\right)^2}{\pi_i + (1-\pi_i)\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}} - \delta_i^2\left(\mathrm{E}(\psi_1(y_i+\delta_i)) - \psi_1(\delta_i)\right)
\end{aligned}
$$

where $\psi_1(x) = \frac{\mathrm{d}^2}{\mathrm{d}x\mathrm{d}x'}\log(\Gamma(x))$ is the trigamma function for $x > 0$. In order to compute the expectations of the digamma and trigamma functions contained in $\boldsymbol{W}^{\delta}$, we do the following approximations:

$$
\begin{aligned}
\mathrm{E}\left(\psi(y_i+\delta_i)\right) &\approx \sum_{k=0}^{m} \psi(k+\delta_i)p(k) \\
\mathrm{E}\left(\psi_1(y_i+\delta_i)\right) &\approx \sum_{k=0}^{m} \psi_1(k+\delta_i)p(k),
\end{aligned}
$$

where 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). Unfortunately, the computing time is considerably dominated by the evaluation of the expectations above. A trick that proved to work quite well in practice is to compute the quantity

$$
-\delta_i\left(\mathrm{E}(\psi(y_i+\delta_i)) - \psi(\delta_i)\right) - \delta_i^2\left(\mathrm{E}(\psi_1(y_i+\delta_i)) - \psi_1(\delta_i)\right) \tag{12}
$$

only within the initialization period for computing starting values (see Section 3.2 below). After that period, we keep expression (12) fixed during MCMC iterations. This procedure reduces computing time at least by two thirds while high acceptance rates and good mixing properties are preserved.

The required quantities in the NB model can directly be obtained from the score vectors and working weights of the ZINB distribution with $\pi = 0$.

3.2 Metropolis-Hastings Algorithm for Zero-Inflated Count Data Regression

The resulting MCMC algorithm can now be compactly summarized as follows:

1. Initialization: Let T be the number of iterations. Set $t = 0$ and determine suitable starting values for all unknown parameters (for example utilizing the backfitting 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 $\boldsymbol{z}^{(t)} = \boldsymbol{\eta}^{(t)} + \left(\boldsymbol{W}^{(t)}\right)^{-1} \boldsymbol{v}^{(t)}$ based on the current values.

   (b) Update $\boldsymbol{\beta}_j$: Generate a proposal $\boldsymbol{\beta}_j^p$ from the density $q(\boldsymbol{\beta}_j^{(t)}, \boldsymbol{\beta}_j^p) = \mathrm{N}\left(\boldsymbol{\mu}_j^{(t)}, \left(\boldsymbol{P}_j^{(t)}\right)^{-1}\right)$ with expectation $\boldsymbol{\mu}_j$ and precision matrix $\boldsymbol{P}_j$ given in (6), and accept the proposal with probability

   $$\alpha\left(\boldsymbol{\beta}_j^{(t)}, \boldsymbol{\beta}_j^p\right) = \min\left\{\frac{p(\boldsymbol{\beta}_j^p|\cdot)q(\boldsymbol{\beta}_j^p, \boldsymbol{\beta}_j^{(t)})}{p(\boldsymbol{\beta}_j^{(t)}|\cdot)q(\boldsymbol{\beta}_j^{(t)}, \boldsymbol{\beta}_j^p)}, 1\right\}.$$

   To solve the identifiability problem inherent to additive models, the sampled effect is corrected according to Algorithm 2.6 in Rue and Held [2005] such that $\boldsymbol{A}\boldsymbol{\beta}_j = \boldsymbol{0}$ holds, with an appropriate matrix $\boldsymbol{A}$, such as $\boldsymbol{A} = \mathbb{1}'\boldsymbol{Z}_j$.

   (c) Update of τ_j^2: Generate the new state from the inverse Gamma distribution $\mathrm{IG}\left(a_j', (b_j')^{(t)}\right)$ with a_j' and b_j' given in (5).

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. In cases with high-dimensional parameter vectors such as in spatial effects acceptance rates might be lower than 30%. An extension to multilevel structure can cover this problem and is explained in the following section.

3.3 Multilevel Framework

Recently, Lang et al. [2013] proposed a multilevel version of structured additive regression models where it is assumed that the regression coefficients $\boldsymbol{\beta}_j$ of a term f_j in (3) may themselves obey a regression model with structured additive predictor, i.e.

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1} + \ldots + \boldsymbol{Z}_{jp_j}\boldsymbol{\beta}_{jp_j} + \boldsymbol{\varepsilon}_j. \tag{13}$$

Here the terms $\boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1}, \ldots, \boldsymbol{Z}_{jp_j}\boldsymbol{\beta}_{jp_j}$ correspond to additional nonlinear functions $f_{j1}, \ldots, f_{jp_j}$ and $\boldsymbol{\varepsilon}_j \sim N(\boldsymbol{0}, \tau_j^2\boldsymbol{I})$ is a vector of i.i.d Gaussian random effects. 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 are models with spatial effects. In this case, covariate $z_j \in \{1, \ldots, S\}$ is a spatial index and $z_{ij} = s_i$ indicates the district observation i pertains to. Then the design matrix $\boldsymbol{Z}_j$ is an $n \times S$ indicator matrix with $\boldsymbol{Z}_j[i, s] = 1$ if the i-th observation belongs to district s and zero otherwise. The $S \times 1$ parameter vector $\boldsymbol{\beta}_j$ is the vector of regression parameters, i.e. the s-th element in $\boldsymbol{\beta}_j$ corresponds to the regression coefficient of the s-th district. Using the compound prior (13), 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

$$\boldsymbol{\beta}_j = \boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1} + \boldsymbol{\varepsilon}_j = \boldsymbol{I}\boldsymbol{\beta}_{j1} + \boldsymbol{\varepsilon}_j$$

where $\boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1} = \boldsymbol{I}\boldsymbol{\beta}_{j1}$ is a structured spatial effect modeled by a Markov random field prior whereas $\boldsymbol{\varepsilon}_j \sim N(\boldsymbol{0}, \tau_j^2\boldsymbol{I})$ 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 $\boldsymbol{\beta}_{j1}$ reduces to the number of districts which 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. [2013] also proposes highly efficient updating of the remaining terms in the level one equation (3) by utilizing 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. [2013] and makes use of the numerical efficient updating schemes described therein.

3.4 Theoretical Results & 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 et al. [2001] and we will now generalize these results to the GAMLSS framework. Assume therefore conditionally independent observations y_i, $i = 1, \ldots, n$, and density $f_i(y_i)$ belonging to an m-parametric distribution family with parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ such that the first and second derivative of the log-likelihood exist. Let $\boldsymbol{\eta}^{\boldsymbol{\theta}_1}, \ldots, \boldsymbol{\eta}^{\boldsymbol{\theta}_m}$ be the predictors linked to the m parameters of the underlying distribution. For each predictor, equation (2) allows us to write $\boldsymbol{\eta} = \sum_{j=1}^{p} \boldsymbol{Z}_j \boldsymbol{\beta}_j$ with appropriate design matrices $\boldsymbol{Z}_j$ and regression vectors $\boldsymbol{\beta}_j$. The basic idea to get sufficient conditions for the propriety of the posterior is to rewrite this model in a mixed model representation with i.i.d. individual specific random effects where we explicitly differ between effects with proper and (partially) improper priors. 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 3.1. *Consider a structured additive regression model within the GAMLSS framework and predictors* (2)*. Assume that conditions 1.–6. specified in Section C*

hold and assume that for $j = 1, \ldots, p$ *and* $l = 1, \ldots, m$ *either* $a_j^{\theta_l} < b_j^{\theta_l} = 0$ *or* $b_j^{\theta_l} > 0$ *hold, where* $a_j^{\theta_l}, b_j^{\theta_l}$ *are the parameters of the inverse gamma prior for* $(\tau^2)^{\theta_l}$. *If the residual sum of squares defined in* (C.6) *for the predictors in the normalized submodel* (C.5) *is greater than* $-2b_0^{\theta_l}$, *then the joint posterior is proper.*

A proof for the theorem is contained in Section C. The technical conditions 1. – 6. given there can be very briefly summarized as the requirement that the sample size should not be too small compared to the total rank deficiency in the Gaussian priors. 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 that in the generalized additive model case.

Regularity of the posterior precision matrix

Concerning the IWLS proposals, a requirement is that the covariance matrix of the approximating Gaussian proposal density is positive definite and therefore invertible. This is ensured if the working weights are all positive. Given full column rank of the design matrix, positivity of the weights is always given for zero-inflated Poisson models as shown in Section B.2. For zero-inflated negative binomial models, the weights involved in the updates for π and μ are always positive (see again Section B.2) while this is not necessarily the case for the weights related to δ. 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 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 paper is implemented in the free, open source software package BayesX [Belitz et al., 2012]. The implementation makes routine use of efficient storing even for large data sets and sparse matrix algorithms for sampling from multivariate Gaussian distributions, see Lang et al. [2013] for details. The implementation in this framework

also has the advantage that the multilevel framework briefly outlined in Section 3.3 becomes accessible for zero-inflated and overdispersed count data regression.

To compute starting values for the MCMC algorithm that ensure rapid convergence towards the stationary distribution, we make use of a backfitting algorithm [Hastie and Tibshirani, 1990] with fixed smoothing parameters. The idea of the algorithm is to approximate the mode of the log-likelihood function and is part of the procedure in BayesX, see Section A for further details.

A challenge when 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 π is actually zero and the estimated predictor η^π corresponding to π will tend to be rather small such that a software crash (e.g. due to overflow errors) is very likely. The problems become even worse for the ZINB model. We therefore included in our software a "save estimate" option that prevents a software crash due to numerical instability. This is obtained by updating a vector of regression parameters, $\boldsymbol{\beta}_j$ say, only if the proposed new state $\boldsymbol{\beta}_j^p$ 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 occur only in a very view number of iterations. If it occurs frequently, then of course the estimated results are not fully valid but rather an indicator that the specified model is too complex for the data at hand.

4 Simulations

This section has two central and simulation based aims to show empirically the performance of the two proposed theoretical models: First, we compare Bayesian inference in additive models with maximum likelihood estimates where the former one is realized in BayesX and for the latter one we use the `gamlss` package in R [Stasinopoulos and Rigby, 2007]. Note, that for the ZINB model we observed convergence problems of the Newton-Raphson/Fisher-scoring algorithm build in the `gamlss` package for about 10% of the simulation replications despite several trials with different hyperparameter settings for the function `pb` that is used to determine smoothing

parameters in `gamlss`. We also tried the `ga` function within the `gamlss.add` package for our simulated data which caused even more convergence problems than with the `gamlss` package. Section 4.1 is therefore organized as follows: First, we present results of the ZIP model for both methods and proceed then in presenting the outcomes of our Bayesian approach in the ZINB model. In the course of this section, frequentist estimates based on the `gamlss` package will be denoted by ML.

In Section 4.2 we look at more complex models that allow to capture unobserved heterogeneity and spatial correlations. The simulation studies presented in Section 4.1 are extended by a spatial effect comprising a structured part based on regions in Germany and modeled by a Markov random field and an unstructured part simulated by a random effect. Although the `gamlss.add` package also provides a possibility to fit models comprising spatial effects based on Markov random fields, it does not support the hierarchical model specification we employed in the simulations. All corresponding studies for the negative binomial distribution can be found in Section E.1.

4.1 Additive Models

In order to compare the ZIP model based on inference described in Section 3 with the frequentist version by Stasinopoulos and Rigby [2007] and to show that the ZINB model can be estimated reliably in the Bayesian framework, we consider the functions

$$f_1^\lambda(\boldsymbol{x}_1) = f_1^\mu(\boldsymbol{x}_1) = \log(\boldsymbol{x}_1), \quad f_2^\lambda(\boldsymbol{x}_2) = f_2^\mu(\boldsymbol{x}_2) = 0.3\boldsymbol{x}_2 \cos(\boldsymbol{x}_2)$$

$$f_1^\pi(\boldsymbol{x}_1) = \sin(\boldsymbol{x}_1), \quad f_2^\pi(\boldsymbol{x}_2) = -0.2\boldsymbol{x}_2^2$$

$$f_1^\delta(\boldsymbol{x}_1) = 0.1\exp(0.5\boldsymbol{x}_1), \quad f_2^\delta(\boldsymbol{x}_2) = -0.5\,\text{arcsinh}(\boldsymbol{x}_2),$$

depending on which of the two models is considered. Each of the predictors introduced in Section 2.1 is written as the sum of two nonlinear functions f_1 and f_2 where the covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained as i.i.d. samples from equidistant grids of step size 0.01, such that for $i = 1, \ldots, n$, we have $x_{i1} \in [1, 6]$ and $x_{i2} \in [-3, 3]$. We use the sample size $n = 1,000$ and simulate 250 replications. An averaged amount of about 50% and 46% of zeros is observed in the generated samples for ZIP and ZINB, respectively. For MCMC inference, posterior mean and quantiles can be computed for each replication using the samples obtained in the MCMC iterations. From the

simulation runs, we also obtain overall empirical bias and MSE for the estimates of all functions as well as pointwise coverage rates. In addition, BayesX provides simultaneous credibility bands which are not discussed here, see Krivobokova et al. [2010] for theoretical details. In the ZIP model, the corresponding quantities are also calculated for ML.

The design matrices in ML and MCMC inferences are induced by cubic B-spline basis functions constructed based on a grid of 20 equidistant knots within the range of the covariates. In ML estimates of the ZIP model, the smoothing parameters $\frac{1}{\tau^2}$ were estimated by using the function `find.hyper` with starting value 3 for all parameters and with default settings for the remaining arguments of the function. The priors for regression coefficients and smoothing variances of the MCMC approach are chosen as presented in Section 2.3. The number of iteration steps K for each simulation run r in MCMC is set to $12,000$ with a burn-in phase of 2,000 iterations. We store and use every 10-th iterate for inference.

Figure D1 (compare supplement Section D) shows the mean over all replications achieved in the ZIP model of ML and MCMC compared to the true simulated functions. In Figure 1, the logarithmic mean squared errors for both approaches are plotted in form of boxplots. Finally, we look at pointwise 95% coverage rates for the

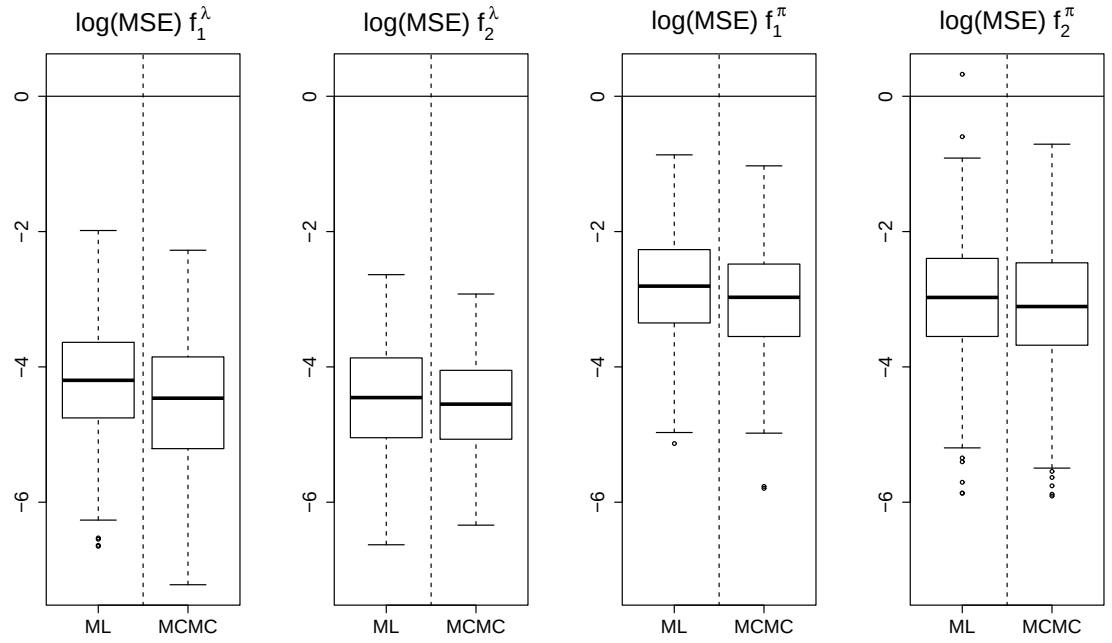

Figure 1: ZIP additive model. log(MSE) of ML and MCMC estimates

ZIP model in Figure 2. 80% coverage rates have also been computed but showed a similar qualitative behaviour and are therefore omitted. The following findings can

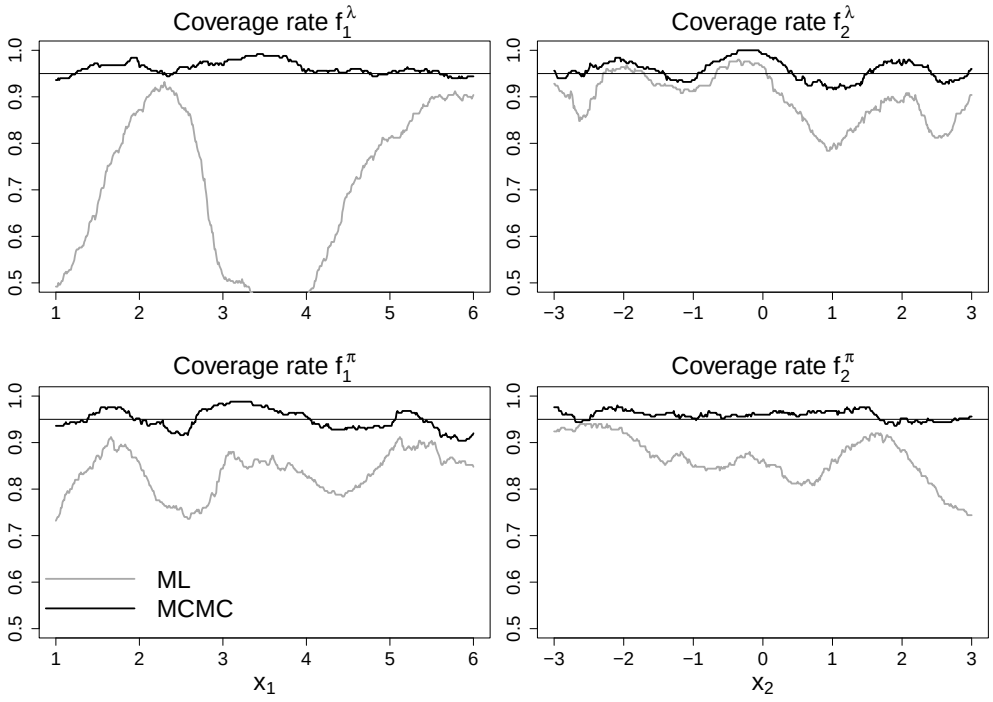

Figure 2: ZIP additive model. Pointwise 95% coverage rates of ML and MCMC estimates

be obtained from the described study for the ZIP model:

- Bias: Averaging all replications leads to satisfactory results for ML and MCMC with only slightly too smooth mean estimates in extreme areas of effects. On the boundary of covariates, MCMC tends to fit the true functions better.

- MSE: Figure 1 confirms the observation that both methods deliver similar mean results since the boxplots of the logarithmic mean squared errors resemble each other summarized over all replications. In general, the nonlinear functions with effects on rate 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 the mean squared errors of f_1^λ and f_2^λ compared to the ones of f_1^π and f_2^π.

- Pointwise coverage rates: Figure 2 provides evidence that the Bayesian approach provides valid confidence intervals which cannot be obtained based on the asymptotic theory of ML. Note, that a corresponding warning is already given in the manual of Stasinopoulos et al. [2008, p.51]. There it is said that

standard errors for fitted distribution parameters might be unreliable if the link function is not the identity function. For MCMC, the 95% level of the credible intervals is mostly maintained.

In conclusion, bias and MSE support that results obtained with MCMC are at least as reliable as those obtained with ML. In addition, the better coverage properties of the credible intervals obtained with MCMC render our Bayesian approach a strong competitor to existing ML estimates.

As stated earlier, a similar simulation study was performed for the ZINB model but no reliable results could be achieved with ML. We therefore only discuss results for MCMC estimates. To have a comparative component we repeated the simulation study with the same simulated effects but doubled the sample size to $n = 2,000$ observations and plotted the mean over all mean estimates for both sample sizes in Figure D2 of the supplement. All corresponding logarithmic mean squared errors of the 250 replications computed from MCMC estimates are given in Figure 3, as well as

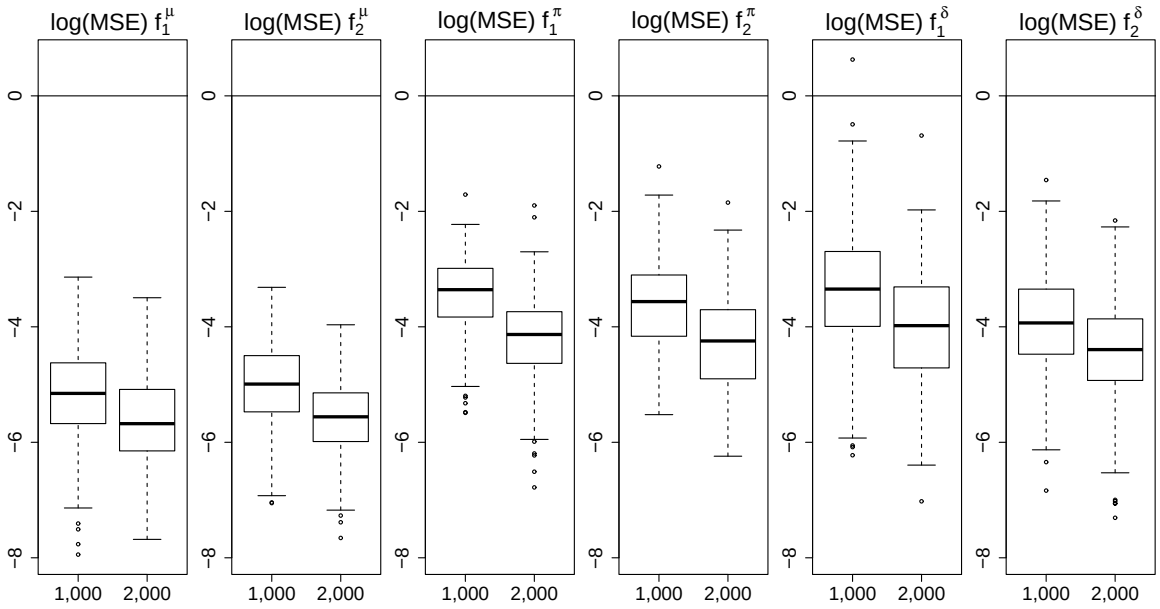

Figure 3: ZINB additive model. log(MSE) of MCMC estimates

95% pointwise credible intervals in Figure 4. Results can be summarized as follows:

- Bias: Averaging all 250 replications leads to mean estimates that are very close to the true function.
- MSE: As expected, the mean squared error is reduced by increasing the sample

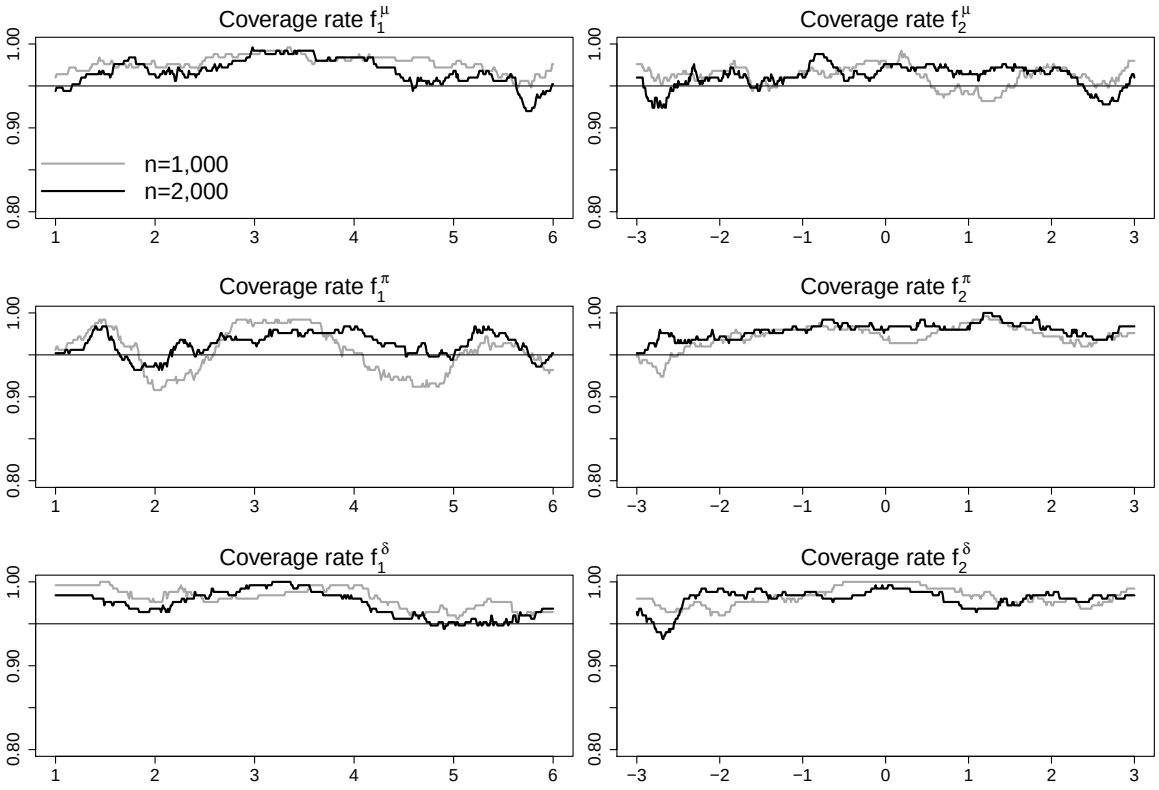

Figure 4: ZINB additive model. Pointwise 95% coverage rates of MCMC estimates

size. Similar to the ZIP model, is is notable that the expectation μ of the underlying count process is easier to estimate than the probability of additional zeros. The same is observable here for the overdispersion parameter δ. The decline in quadratic deviations from the true function by increasing the sample size has its greatest effect in f_1^{δ}, such that the outliers with an MSE greater than one vanish.

- Pointwise coverage rates: The pointwise coverage rates in Figure 4 indicate reliable credible intervals for both sample sizes.

In a nutshell, the positive results found in the simulation on ZIP data carry over to the more general and complex situation of ZINB data. 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.

4.2 Geoadditive Models

In a second step, the simulation studies for all three, the ZIP, ZINB and NB model have been extended where an additional spatial effect on the Western part of Germany was simulated as follows

$$\begin{aligned} f^{\lambda}_{\text{spat}}(l) = f^{\mu}_{\text{spat}}(l) &= \sin(x^c_l y^c_l) + \epsilon_l \\ f^{\pi}_{\text{spat}}(l) &= \sin(x^c_l)\cos(0.5 y^c_l) + \epsilon^{\pi}_l \\ f^{\delta}_{\text{spat}}(l) &= 0.5 x^c_l y^c_l + \epsilon^{\delta}_l. \end{aligned}$$

The structured part of the spatial effect f_{spat} is estimated by a Markov random field and is simulated on the basis of centroids c_s with standardized coordinates (x^c_s, y^c_s), $s \in \{1, \ldots, S\}$ of the $S = 327$ regions in Western Germany. The unstructured part is described by an additional random effect $\epsilon_s \sim \text{N}(0, 1/16)$ for each of the regions. In Figure D3 of the supplement, the two simulated complete spatial effects for the rate λ of the count process as well as for the probability of the additional zeros π in case of a ZIP model are visualized. The model for a generic predictor $\boldsymbol{\eta}$ can now be written as

$$\boldsymbol{\eta} = f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) + \boldsymbol{f}_{\text{spat}} + \boldsymbol{\epsilon} = \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}_{\text{spat}}\boldsymbol{\beta}_{\text{spat}} + \boldsymbol{\epsilon}.$$

Estimates are based on a two-level structured additive regression where the total spatial effect is decomposed in a structured part $\boldsymbol{f}_{\text{spat}}$ and an unstructured effect $\boldsymbol{\epsilon}$. The basic idea of the framework was introduced in Section 3.3.

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 5,000. We store each 50-th iterate so that the final sample size of 1,000 is retained. To find a desirable sample size for which satisfactory estimate results can be achieved, we performed estimates for $n = 1,000,\ 2,000,\ 4,000$ and $16,000$ observations. Note, that in the following we restrict to the presentation of results in the ZIP model. Results for the ZINB model are summarized at the end of this section. An illustration of results for this model are shown in Section E.2 as well as in E.1.2 for the the NB model.

As has been shown in Lang and Fahrmeir [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. We therefore present the estimated complete spatial effect compared to the true simulated effect for two selected sample sizes $n = 1,000$ and $4,000$ in Figure 5. Beside this, the log(MSE) in Figure 6 and the kernel densities of complete spatial effects in Figure D4 give further information about the quality of the inference.

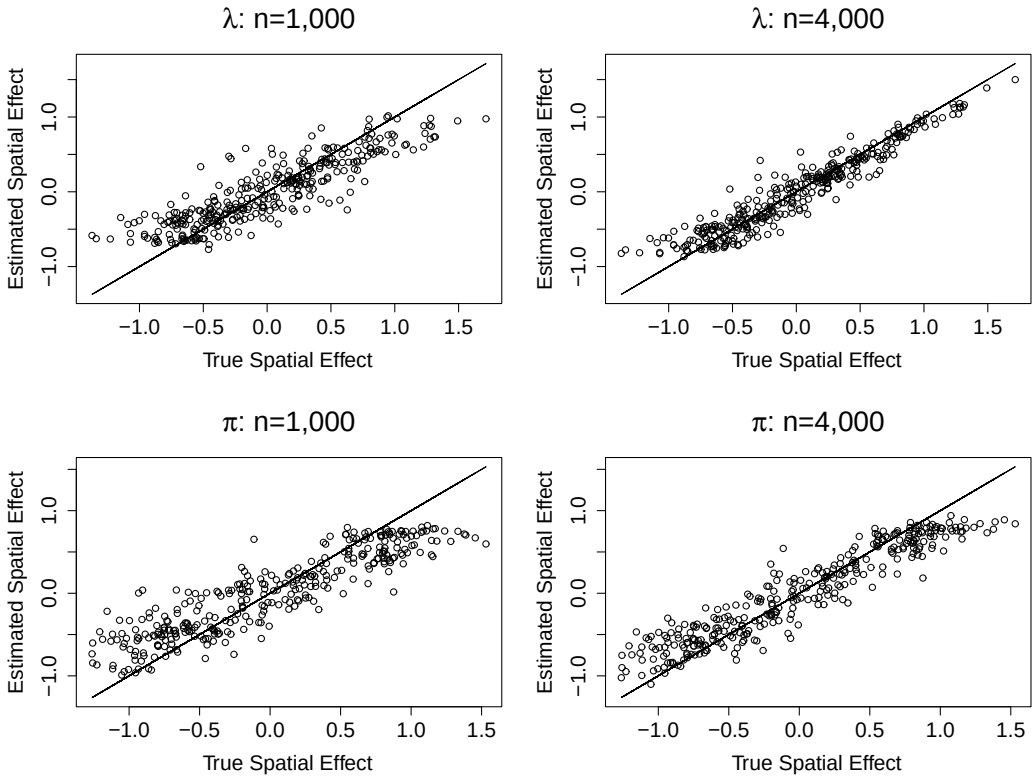

Figure 5: ZIP geoadditive model. Estimated complete spatial effects

The results visualized in these figures can be summed up as follows:

- MSE: The spatial effect has higher log(MSE) compared to the nonlinear effects but we observe that for greater sample sizes the MSE can be reduced in all effects. If one compares Figure 1 with Figure 6, it is positive to note that for sample size $n = 1,000$ an additional spatial effect does not impair the MSE of the nonlinear effects.

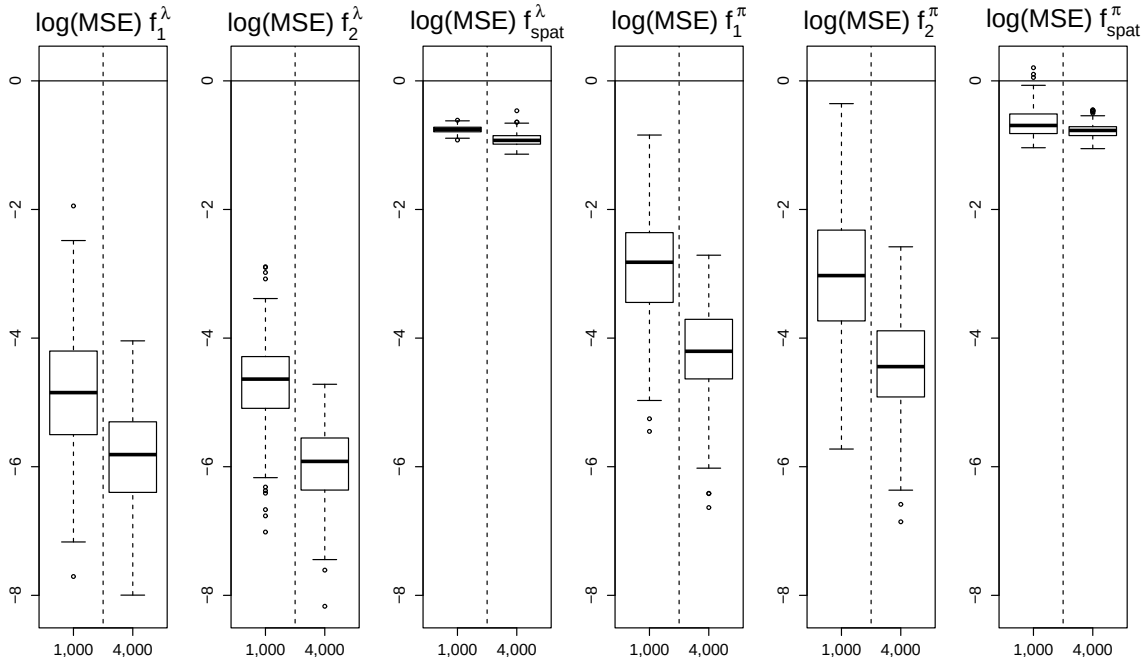

Figure 6: ZIP geoadditive model. log(MSE) of nonlinear and complete spatial effects

- Bias: Figure 5 shows that an increase of the sample size improves the estimates. Extreme values of the spatial effect are most difficult to estimate where both high negative and high positive effects are underestimated. Together with Figure D4 it can be said that the complete spatial effect tends to be rated too smooth in comparison with the true effect.

Results showed that with a sample size of $n = 4,000$ 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.

5 Application: Patent Citations

In our first application we will 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 on the

number of citations for 4,866 patents, see [Graham et al., 2002, Jerak and Wagner, 2006]. Details about data set 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 mean 1.64 and variance 7.53. Roughly 46% of the observations are zeros, the smallest and largest observed values are zero and 40. While these summary statistics do not take into account the potential covariate effects, they already provide a rough indication that overdispersion and zero-inflation may be relevant to obtain a realistic model for the number of citations.

To investigate the relevance of overdispersion and zero-inflation we consider the four candidates Poisson, ZIP, negative binomial and ZINB as possible distributions for the response and use the predictor structure

$$\eta = f_1(year) + f_2(ncountry) + f_3(nclaims) + \boldsymbol{x}'\boldsymbol{\beta}$$

for all relevant model parameters. Here, *year* is the grant year, *ncountry* denotes the number of designated states, *nclaims* are the number of claims against the patent and $\boldsymbol{x}'\boldsymbol{\beta}$ contains linear effects of further binary covariates described in [Fahrmeir et al., 2013] and an intercept term. The nonlinear effects are modeled by cubic P-splines with 20 inner knots and second order random walk prior. Estimates are usually based on 12,000 iterations and a burn-in phase of 2,000 iterations to ensure convergence. Every 10-th iterate is stored to obtain a sample of close-to-independent samples. Convergence and mixing of the Markov chains were assessed graphically. While no severe problems were found for the mixing and convergence of Poisson, ZIP and NB model, the mixing behavior for the parameters in the probability for additional zeros π of the ZINB model was somewhat problematic. This problem originates from the fact that there is only relatively weak evidence for zero-inflation when accounting for overdispersion and therefore the effects and in particular the level of the probability for additional zeros are only weakly identified. Therefore we increased the number of iterations for the ZINB model to 202,000 and a thinning parameter of 200.

The results of all models were compared in terms of normalized (randomized) quantile residuals as a graphical device suggested by Stasinopoulos et al. [2008]: For an observation y_i, the residual is given by $\hat{r}_i = \Phi^{-1}(u_i)$ where Φ^{-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 $[F(y_i - 1|\hat{\boldsymbol{\theta}}), F(y_i|\hat{\boldsymbol{\theta}})]$, $\hat{\boldsymbol{\theta}}$ comprises all estimated model parameters and $F(\cdot|\hat{\boldsymbol{\theta}})$ is the cumulative distribution function obtained by plugging in these 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 7 shows one realization for the Poisson, ZIP, negative binomial and ZINB model. It clearly indicates a preference for the negative binomial or ZINB model that provide a considerably better fit for estimating the distribution of patent citations. Although the residuals of the Poisson model can be improved applying the ZIP model, the sample quantiles greater than 2 are too high compared to the true quantiles. Both, the negative binomial and the ZINB model seem to overcome this problem. In a second step, we applied proper scoring rules

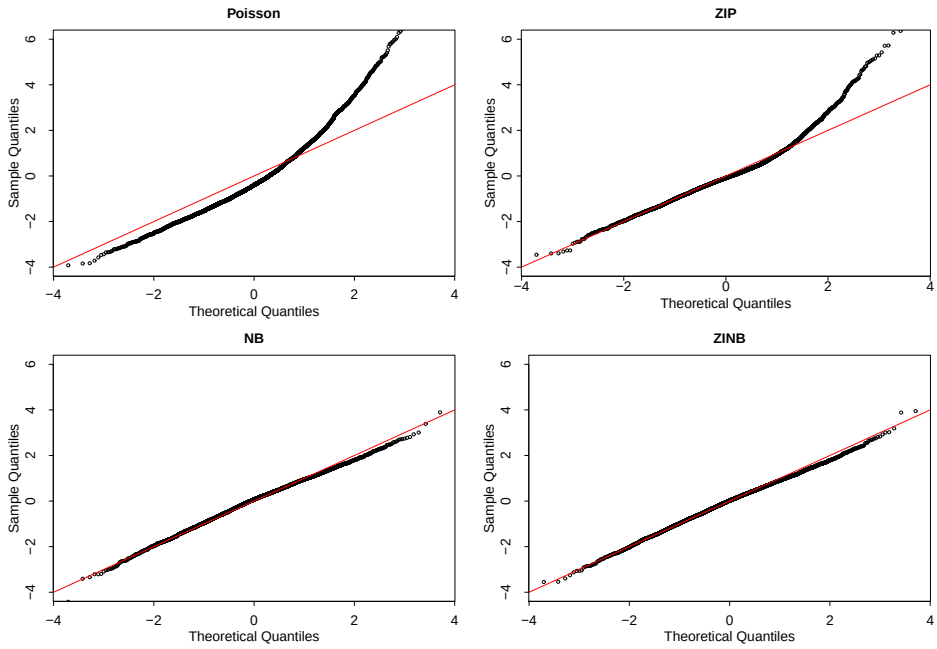

Figure 7: Patent citations. Comparison of quantile residuals

proposed by Gneiting and Raftery [2007] in order to confirm the findings assessed by the residuals: Let $y_1, \ldots, y_n$ be data in a hold out sample and $\hat{\boldsymbol{p}}_j$ the estimated probabilities of a predictive distribution, $\hat{p}_{jk} = p(y_j = k)$. Then a score is obtained by summing up individual score contributions, i.e. $S = \sum_{j=1}^{n} S(\hat{\boldsymbol{p}}_j, y_j)$. Let p_0 be the true distribution, then Gneiting and Raftery [2007] take the expected value of the score under p_0 in order 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 scores given in Gneiting and Raftery [2007]: the Brier score or quadratic score, $S(\hat{\boldsymbol{p}}_j, y_j) = -\sum_k (\mathbb{1}(y_j = k) - \hat{p}_{jk})^2$, the logarithmic score, $S(\hat{\boldsymbol{p}}_j, y_j) = \log(\hat{p}_{jy_j})$, and the spherical score $S(\hat{\boldsymbol{p}}_j, y_j) = \frac{\hat{p}_{jy_j}}{\sqrt{\sum_k \hat{p}_{jk}^2}}$. All these scoring rules are strictly proper but the logarithmic scoring rule has the drawback that it only takes into account one single probability of the predictive distribution and is therefore susceptible to extreme observations. In our application, the predictive distribution is assessed by a ten-fold cross validation. Table 1 summarizes the three scores for all four models. Similar to the residuals, the scores indicate that a Poisson distribution is the worst assumption. The scores of the ZIP are higher compared to Poisson but the best scores are obtained from NB and ZINB. In conclusion it can be said that overdispersion plays a major role in this data set and that there is some evidence for additional zero-inflation. Since the residuals look slightly better

Model	Brier Score	Logarithmic Score	Spherical Score
Poisson	-3,773.76	-10,530.62	32.41
ZIP	-3,456.48	-8,808.44	36.75
NB	-3,413.41	-8,120.43	37.31
ZINB	-3,388.40	-7,999.92	37.64

Table 1: Patent citations. Evaluated scores

for ZINB compared to NB and all three scores would prefer this model as well, we choose the ZINB model as our final model. Figure 8 displays mean sample results of the stored MCMC iterates for all three parameters and with respect to the three covariates *year*, *nclaims* and *ncountry* (row by row) together with pointwise 80% and 95% credible intervals. The vertical stripes indicate the relative amount of observations relating to the different covariate values (the darker the stripes, the more data). The following observations and interpretations on selected effects of Figure 8 can be made:

- The first row shows the estimated centered effects on the expectation μ of the underlying count process (which is not the same as the expectation of the response). For example, if we look at patents with grant year later than 1985, we estimate that patents are cited the less the newer they are.

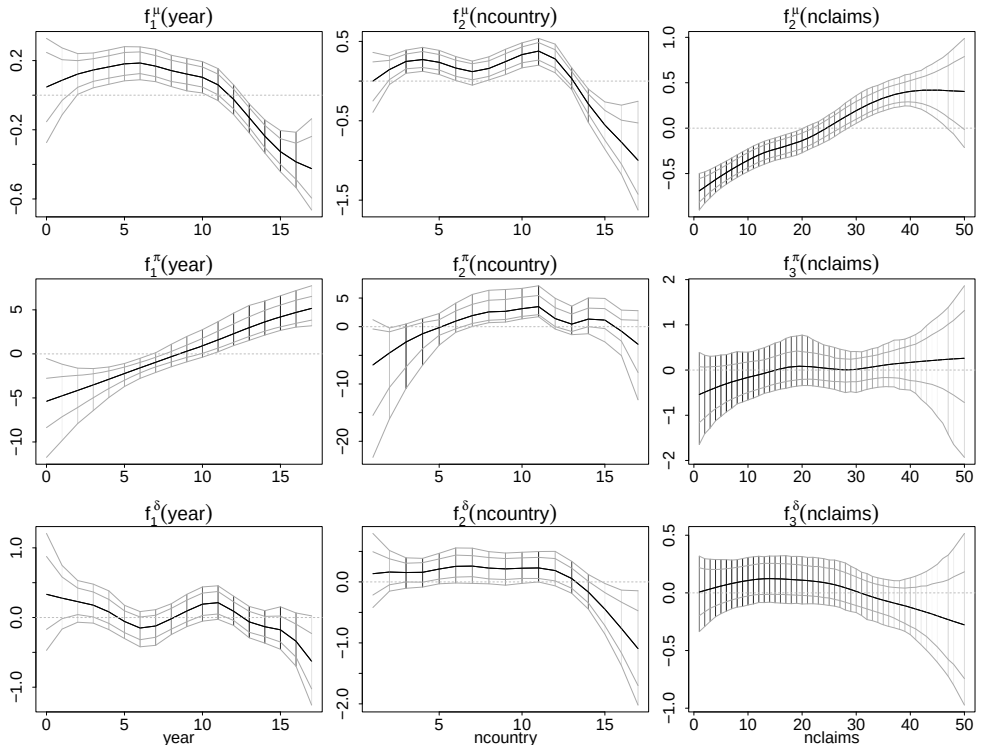

Figure 8: Patent data. Estimated centered nonlinear effects in the ZINB model

- In the second row, the corresponding estimates on the probability of structural zeros π indicate covariate values with a high probability of never being cited. With respect to the variable *year*, it is reasonable to have a decreasing chance of no citations for rising age of the patent. The effects of *ncountry* and *nclaims* are insignificant in the sense that the confidence bands cover the zero line.

- The expectation of y given the covariate information is given by $(1-\pi)\mu$: For an adequate interpretation it is important to see that an increase of the effects on μ and a decline of the function estimates on π result in a growing estimated expectation and vice versa. In general, the effect of a covariate on the expectation $(1-\pi)\mu$ is therefore hard to predict. For the patent data, we find that $(1-\pi)\mu$ behaves similar as μ in *year*, *ncountry* and *nclaims* when all other effects are kept constant.

- The variance of a zero-inflated negative binomial distributed variable can be derived from equation (1) as $\text{Var}(y_i) = (1-\pi_i)\mu_i\left(1+\mu_i\left(\delta_i^{-1}+\pi_i\right)\right)$. From this we find that δ is inversely proportional to the variance such that with respect to one effect in δ, and all others maintained fixed, an increasing function results in

a smaller variance. However, the estimated effects shown in Figure 8 are largely insignificant.

6 Application: Car Insurance

We also apply the developed methods to a data set of size $n = 162,548$ from car insurance in Belgium of the year 1997. The insurance premium in car insurances is based on detailed statistical analyses of the risk structure of the policyholder. One important step is to model the loss frequency which usually depends on the characteristics of the policyholder as well as the vehicle. Typical covariates are the age of the policyholder (*ageph*), age of the vehicle (*agec*), the engine power (*power*) and the previous claim experience. In Belgium, the claim experience is measured by a 22-step bonus-malus-score (*bm*). The higher the score, the better the history of the policyholder. The data also provides the geographical information in which of the 589 districts (*distr*) in Belgium the policyholder's car is registered.

The data set 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 introdued in Heller et al., 2006] for claim sizes in the framework of GAMLSS is provided in Klein et al. [2013]. 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

$$\eta = f_1(ageph) + sex\, f_2(ageph) + f_3(agec) + f_4(bm) + f_5(power) + f_{\text{spat}}(distr) + (\boldsymbol{x})'\boldsymbol{\beta}$$

for the mean parameter in the count process, i.e. λ in case of ZIP and μ in case of NB or ZINB. The spatial effect has been modeled by a Markov random field and the term $(\boldsymbol{x})'\boldsymbol{\beta}$ contains additional linear effects of dummy variables [Denuit and Lang, 2004] that will not be discussed here. 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 behavior 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 deviance information criterion (DIC), see Spiegelhalter et al. [2002]. 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 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 visualized in Figure 9, we refrained from searching for (even) more complex ZINB models.

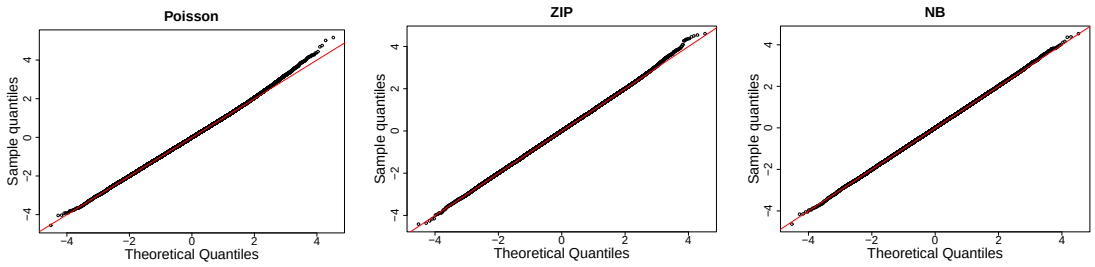

Figure 9: Insurance claims. Comparison of quantile residuals

Model	Brier Score	Logarithmic Score	Spherical Score
Poisson	-32,261.64	-62,131.83	360.9523
ZIP	-32,247.24	-61,997.6	360.9736
NB	-32,252.93	-61,981.25	360.9660

Table 2: Insurance claims. Evaluated scores

Table 2 shows the calculated scores for the Poisson, ZIP and NB distribution which have been introduced in Section 5 and which are again obtained by a ten-fold cross validation. In general, differences are smaller than for the patent application but still there is 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 favor of the ZIP model while the logarithmic score would prefer the NB model. The quantile residuals depicted in Figure 9 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 favors the NB model since this score reacts particularly sensitive to predictive problems of

extreme (in our case large) observations. In summary, there is no clear evidence in favor of ZIP or NB and both models seem to provide a reasonable fit. In the following, we present results for the ZIP model to illustrate the interpretation of estimated effects. The selected model for π provides the predictor

$$\eta^{\pi} = f_1^{\lambda}(ageph) + f_2^{\lambda}(agec) + f_{\text{spat}}^{\pi}(distr) + (\boldsymbol{x}^{\pi})' \boldsymbol{\beta}^{\pi}.$$

The spatial effect contains only a Markov random field since as in the predictor for λ an additional i.i.d random effect was neither significant nor selected by the DIC. In Figure 10 the estimated nonlinear effects on λ and π are plotted together with 80% and 95% pointwise credible intervals. Again, vertical stripes indicate relative amount of data of the corresponding covariate values. Figure 11 depicts the estimated spatial

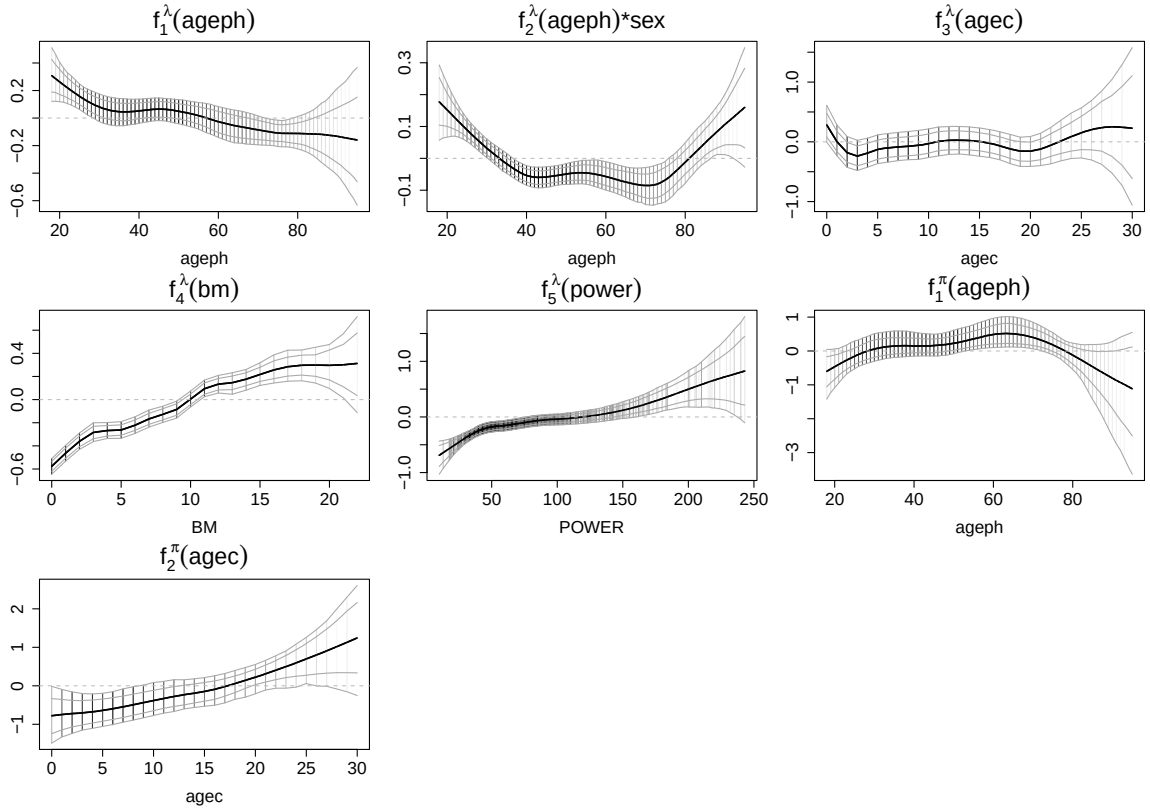

Figure 10: Insurance claims. Estimated centered nonlinear effects in the ZIP model

effects on λ and π. The estimated effects for λ in Figure 11 are generally close to those in Denuit and Lang [2004]. We discover for example that the age-sex interaction is significant in the way that males younger than 35 and males older than 80 report more accidents than females of the same ages. The peak of the effect of age at around 45 can be explained by the fact that asking older relatives to pay the policy is very common

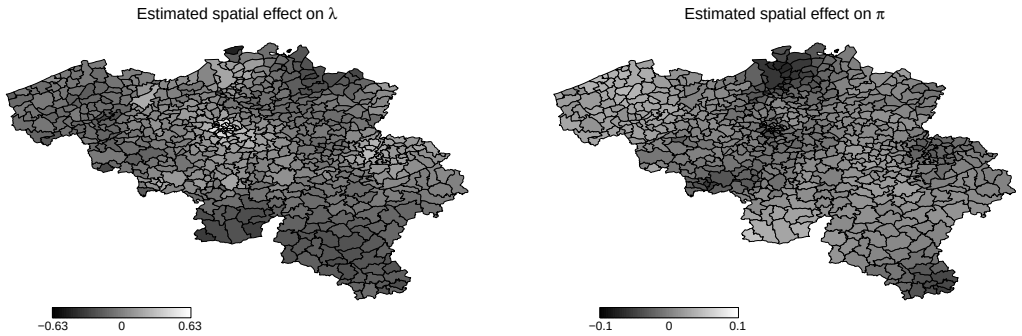

Figure 11: Insurance claims. Estimated spatial effects in the ZIP model

in Belgium because of the high premiums for young policyholders. The spatial effect in Figure 11 clearly indicates a large number of expected claims in urban areas like Brussels, Antwerp or Liège.

For π the monotonically increasing effect of *agec* can be seen as an indication for an excess of zero claims for older cars. The estimated spatial effect for π is pronounced as well but generally weaker than for λ.

7 Summary and Conclusions

In this paper, 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 `gamlss` package of R, [Stasinopoulos and Rigby, 2007], as our predictors may include complex, hierarchical spatial effects and may in general cope with hierarchical data situations as described in Lang et al. [2013]. Moreover, simulation studies revealed that the Bayesian approach yields reliable confidence 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 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 the multiple predictors involved. For 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 et al. [2012] in a Bayesian approach via spike and slab priors.

The Bayesian formulation of GAMLSS also provides the possibility to include modified / extended prior structured 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 non-normal 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. [2013] in the context of zero-adjusted models as introduced in a frequentist setting by Heller et al. [2006].

References

C. Belitz and S. Lang. Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis, 53:61–81, 2008.

C. Belitz, A. Brezger, T. Kneib, S. Lang, and N. Umlauf. Bayesx, 2012. - Software for Bayesian inference in structured additive regression models. Version 2.1. Available from http://www.bayesx.org.

A. Brezger and S. Lang. Generalized structured additive regression based on bayesian p-splines. Computational Statistics & Data Analysis, 50:967–991, 2006.

C. Czado, V. Erhardt, A. Min, and S. Wagner. 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, 2007.

M. Denuit and S. Lang. Non-life rate-making with bayesian gams. Insurance: Mathematics and Economics, 35:627–647, 2004.

P.K. Dunn and G.K. Smyth. Randomized quantile residuals. Computational and Graphical Statistics, 5:236–245, 1996.

L. Fahrmeir and T. Kneib. Propriety of posteriors in structured additive regression models: Theory and empirical evidence. Journal of Statistical Planning and Inference, 39:843–859, 2009.

L. Fahrmeir and L. Osuna Echavarría. Structured additive regression for overdispersed and zero-inflated count data. Applied Stochastic Models in Business and Industry, 22:351–369, 2006.

L. Fahrmeir and G. Tutz. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, 2001.

L. Fahrmeir, T. Kneib, and S. Lang. Penalized structured additive regression for space-time data: a Bayesian perspective. Statistica Sinica, 14:731–761, 2004.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. Regression - Models, Methods and Applications. Springer, 2013.

D. Gamerman. Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing, 7:57–68, 1997.

T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.

S. Graham, B. Hall, D. Harhoff, and D. Mowery. Post-issue patent "quality control": a comparative study of us patent reexaminations and european patent oppositions. Technical report, NBER, 2002. Working Paper 8807.

T.J. Hastie and R.J. Tibshirani. Generalized Additive Models. Chapman & Hall, 1990.

G. Heller, Stasinopoulos D. M., and Rigby R. A. The zero-adjusted inverse gaussian distribution as a model for insurance data. In J.Newell J. Hinde, J.Einbeck, editor, Proceedings of the 21th International Workshop on Statistical Modelling, 2006.

J.M. Hilbe. Negative binomial regression. Cambridge University Press, 2007.

A. Jerak and S. Wagner. Modeling probabilities of patent oppositions in a bayesian semiparametric regression framework. Empirical Economics, 31:513–533, 2006.

A. Jullion and P. Lambert. Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. Computational Statistics & Data Analysis, 51: 2542–2558, 2007.

N. Klein, M. Denuit, T. Kneib, and S. Lang. Nonlife ratemaking and risk management with bayesian additive model for location scale and shape. Technical report, 2013.

T. Kneib, T. Hothorn, and G. Tutz. Variable selection and model choice in geoadditive regression models. Biometrics, 65:626–634, 2009.

T. Krivobokova, T. Kneib, and G. Claeskens. Simultaneous confidence bands for penalized spline estimators. Journal of the American Statistical Association, 105:852–863, 2010.

S. Lang and A. Brezger. Bayesian p-splines. Journal of Computational and Graphical Statistics, 13: 183–212, 2004.

S. Lang and L. Fahrmeir. Bayesian generalized additive mixed models. a simulation study. discussion paper 230, sfb 386. supplement paper to Fahrmeir, L. and Lang, S. (2001): Bayesian semiparametric regression analysis of multicategorical time-space data. annals of the institute of statistical mathematics, 53, 10-30. Technical report, 2001. URL http://www.uibk.ac.at/statistics/personal/lang/publications/.

Stefan Lang, Nikolaus Umlauf, Peter Wechselberger, Kenneth Harttgen, and Thomas Kneib. Multilevel structured additive regression. Statistics and Computing, 23, 2013.

R.A. Rigby and D.M. Stasinopoulos. Generalized additive models for location, scale and shape,(with discussion). Applied Statistics, 54:507–554, 2005.

H. Rue and L. Held. Gaussian Markov Random Fields. Chapman & Hall / CRC, 2005.

D. Ruppert, M. P. Wand, and R. J. Carroll. Semiparametric Regression. Cambridge University Press, 2003.

F. Scheipl, L. Fahrmeir, and T. Kneib. Spike-and-slab priors for function selection in structured additive regression models. Journal of the American Statistical Association, 107:1518–1532, 2012.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, 65(B):583–639, 2002.

D.M. Stasinopoulos and R.A. Rigby. Generalized additive models for location scale and shape (gamlss) in r. Journal of Statistical Software, 23(7):1–46, 2007.

D.M. Stasinopoulos, B. Rigby, and C Akantziliotou. Instructions on how to use the gamlss package in R, Second Edition, 2008.

D. Sun, R.K. Tsutakawa, and H. Zhuoqiong. Propriety of posteriors with improper priors in hierarchical linear mixed models. Statistica Sinica, 11:77–95, 2001.

R. Winkelmann. Econometric Analysis of Count Data. Springer, 2008.

S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association, 99:673–686, 2004.

S. N. Wood. Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society, Series B, 70:495–518, 2008.

S.N. Wood. Generalized Additive Models : An Introduction with R. Chapmann & Hall, 2006.

Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in R. Journal of Statistical Software, 27(8), 2008. URL http://www.jstatsoft.org/v27/i08/.

Bayesian Generalized 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 summarize a backfitting algorithm, see [Hastie and Tibshirani, 1990], for obtaining the starting values for the MCMC sampler utilized in the paper. We basically approximate the maximum of the log-likelihood, this is the mode, by maximizing numerically its quadratic approximation:

1. Initialization of values: Set $\boldsymbol{\beta}_1^{(0)} = \ldots = \boldsymbol{\beta}_p^{(0)} = 0$ as well as $\beta_0^{(0)} = g(\hat{\rho})$ where g is the link function between the generic model parameter ρ and the predictor $\boldsymbol{\eta}$. $\hat{\rho}$ is an simple estimator for ρ, just depending on the responses. If for example, ρ stands for the average rate $\bar{\lambda} = \sum_{i=1}^{n} \lambda_i$ in the ZIP model, $\hat{\rho}$ could be the mean of the observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Let K be the maximum number of iterations in the algorithm and set $k = 0$.

2. Estimation of $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_p$ and β_0 :

   (a) Set $r = 0$ and for $j = 1, \ldots, p$

   $$\boldsymbol{f}_j^{(r)} = \boldsymbol{f}_j^{(k)} = \boldsymbol{Z}_j \boldsymbol{\beta}_j^{(k)} \quad \text{as well as} \quad \beta_0^{(r)} = \beta_0^{(k)} = g(\hat{\rho})$$

   (b) Outer backfitting slope: Compute

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

and define $\boldsymbol{S}_j^{(k)} := \boldsymbol{Z}_j \left(\boldsymbol{Z}_j' \boldsymbol{W}^{(k)} \boldsymbol{Z}_j + \frac{1}{\tau_j^2} \boldsymbol{K}_j \right)^{-1} \boldsymbol{Z}_j' \boldsymbol{W}^{(k)}$, $j = 1, \ldots, p$

(c) Inner backfitting slope: Calculate for $j = 1, \ldots, p$

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

(d) Centering of the estimations

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

$$\frac{\left| \beta_0^{(r+1)} - \beta_0^{(r)} \right| + \sum_{j=1}^{p} \left\| \boldsymbol{f}_j^{(r+1)} - \boldsymbol{f}_j^{(r)} \right\|}{\left| \beta_0^{(r+1)} \right| + \sum_{j=1}^{p} \left\| \boldsymbol{f}_j^{(r+1)} \right\|} < \epsilon$$

end the inner backfitting slope, set for $j = 1, \ldots, p$

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

and go to (f). Otherwise set $r = r + 1$ and go to (c).

(f) If $k < K$ go to (b). Otherwise stop the algorithm.

B Working Weights

B.1 Computation of the Working Weights

The working weights given in Section 3 might not be that obvious at the first sight. For some of them several steps of calculations and simplifications had to be done. In principle, the approach is simple: For the score vectors $\boldsymbol{v}$ the first derivatives of the log-likelihood with respect to each predictor have to be computed. The working weights are achieved by taking the expectation of the 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:

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

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

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

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

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

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

For the ZINB model calculations can be written as follows:

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

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

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

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

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

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

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

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

B.2 Positive Definiteness of the Working Weights

Lemma B.1. *The working weights $\boldsymbol{W}^{\lambda}$ and $\boldsymbol{W}^{\pi}$ in the ZIP model are positive definite.*

Proof. As both matrices are diagonal it is only to show that all entries on the diagonal are greater than zero. Let us start with $\boldsymbol{W}^{\lambda}$: We need to proof that

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

remains true because the denominator in (7) is obviously greater than zero. 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 eigenvalues of $\boldsymbol{W}^{\lambda}$ are greater than zero. For $\boldsymbol{W}^{\pi}$ in (8) we need only to show

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

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

Lemma B.2. *The working weights $\boldsymbol{W}^{\mu}$ and $\boldsymbol{W}^{\pi}$ in the ZINB model are positive definite.*

Proof. As both matrices are diagonal it is only to show that all entries on the diagonal are greater than zero. Let us start with $\boldsymbol{W}^{\mu}$ in (9) 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

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

The first term is greater than zero as we assume $\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}<1$. The second an 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 differ between the two cases $\delta_i\leq\mu_i$ and $\delta_i>\mu_i$.

(i) $\delta_i\leq\mu_i$: It is sufficient $\left(\frac{\delta_i+\mu_i}{\delta_i}\right)^{\delta_i}\geq\delta_i$ or equivalently $\delta_i\log\left(1+\frac{\mu_i}{\delta_i}\right)\geq\log(\delta_i)$ to show. 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 true. If $0\leq\delta_i\leq 1$ it is nothing to do. For $\delta_i>1$:

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

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:

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

It is sufficient to prove that

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

If $\delta_i \leq 1$, it 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 $\lfloor\delta_i\rfloor = \max\{j \in \mathbb{Z} | j \leq \delta_i\}$. Remains the case $1 < \delta_i < 2$. But then it is

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

All in all we get that the entries on the diagonal of $\boldsymbol{W}^{\mu}$ are greater than zero and therefore that the working weight $\boldsymbol{W}^{\mu}$ is positive definite.

For $\boldsymbol{W}^{\pi}$ we need to look at (10). 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 (10) are greater than zero. In consequence all entries on the diagonal of $\boldsymbol{W}^{\pi}$ are greater than zero as desired. □

C Propriety of the Posterior Distribution

Proof of Theorem 3.1. Following the steps in Fahrmeir and Kneib [2009], we finally represent (2) in terms of $\tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{X}}\boldsymbol{\xi} + \tilde{\boldsymbol{V}}\boldsymbol{b} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \tau_0\boldsymbol{I})$ with reduced dimension $\dim(\boldsymbol{\epsilon})=\tilde{k}_0$ and for a subsample of observations. The advantage here is that this model contains random effects with proper priors and individual-specific $\boldsymbol{\epsilon}$, such that the findings of Sun et al. [2001] can be applied.

For each of the model terms $\boldsymbol{Z}_j\boldsymbol{\beta}_j$ with improper prior precision matrix $\boldsymbol{K}_j$ we have $\text{rk}(\boldsymbol{K}_j) = k_j < d_j = \dim(\boldsymbol{\beta}_j)$. 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 $\boldsymbol{b}_j$ of random effects with proper prior such that $\boldsymbol{Z}_j\boldsymbol{\beta}_j = \boldsymbol{X}_j\boldsymbol{\xi}_j + \boldsymbol{V}_j\boldsymbol{b}_j$ holds. In order to clearly separate effects with respect to their priors, we therefore first of all rewrite (2) as

$$\boldsymbol{\eta} = \boldsymbol{U}\boldsymbol{\gamma} + \sum_{j=1}^{p} \boldsymbol{Z}_j\boldsymbol{\beta}_j + \boldsymbol{Z}_0\boldsymbol{\beta}_0 \tag{C.1}$$

where $\boldsymbol{U}$ consists of all linear effects and has full rank r and $p(\boldsymbol{\gamma}) \propto \text{const}$. The additional term $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ is a random effect with full rank $n \times d_0$ design matrix $\boldsymbol{Z}_0$, $\text{rk}(\boldsymbol{Z}_0) = d_0 = \dim(\boldsymbol{\beta}_0)$ such that $d_0 \geq d_j$ and $k_0 \geq k_j$ for $j = 1, \ldots, p$ with $k_0 = \text{rk}(\boldsymbol{K}_0)$. The distribution of $\boldsymbol{\beta}_0$ is then given by a possibly improper prior of the form (4), i.e.

$$p(\boldsymbol{\beta}_0) \propto \left(\frac{1}{\tau_0^2}\right)^{\frac{k_0}{2}} \exp\left(-\frac{1}{2\tau_0^2}\boldsymbol{\beta}_0'\boldsymbol{K}_0\boldsymbol{\beta}_0\right).$$

The smoothing variance τ_0^2 is assumed to have an inverse gamma prior with parameters a_0, b_0. As said in Fahrmeir and Kneib [2009] setting $\boldsymbol{Z}_0 = \boldsymbol{I}, \boldsymbol{\beta}_0 = \boldsymbol{\epsilon} \sim N(0, \tau_0\boldsymbol{I})$, (C.1) also covers individual specific random effects as a special case. In geoadditive models, $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ may represent a structured or unstructured spatial effect and usually simply corresponds to the term with the largest number of parameters.

The so far presented considerations and Fahrmeir and Kneib [2009, Section 4] allow now to write (C.1) in a mixed model representation, including the additional term $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{V}\boldsymbol{b} + \boldsymbol{V}_0\boldsymbol{b}_0, \quad \boldsymbol{b}_0 \sim N(0, \tau_0\boldsymbol{I}), \tag{C.2}$$

with $\dim(b_0) = k_0$ where $\boldsymbol{X}$ captures terms with partially improper priors and is a full rank augmentation of $\boldsymbol{U}$, for details see Fahrmeir and Kneib [2009, Remark 1], and $\boldsymbol{V}$ represents the model part with proper priors and random effects. The augmented design matrix $\boldsymbol{X}$ may possibly contain additional columns constructed from the unpenalized part of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1 \ldots, \boldsymbol{\beta}_p$. Let q be the number of additional columns, such that $\mathrm{rk}(\boldsymbol{X}) = r + q$.

We make the following assumptions for eventually reordered observations y_i:

1. $\int f_i(y_i|\eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m})\mathrm{d}\eta_i^{\theta_1} \ldots \mathrm{d}\eta_i^{\theta_m} < \infty$ holds for observations $i = 1, \ldots, n^*$
2. $f_i(y_i|\eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m}) \leq M$ holds for the remaining observations $i = n^* + 1, \ldots, n$.

As in Fahrmeir and Kneib [2009] we denote the submatrices of $\boldsymbol{U}^{\theta_l}, \boldsymbol{X}^{\theta_l}, \boldsymbol{Z}^{\theta_l} = (\boldsymbol{Z}_1^{\theta_l}, \ldots, \boldsymbol{Z}_p^{\theta_l})$ and $\boldsymbol{Z}_0^{\theta_l}$ corresponding to $i = 1, \ldots, n^*$, by $(\boldsymbol{U}^{\theta_l})^*, (\boldsymbol{X}^{\theta_l})^*, (\boldsymbol{Z}^{\theta_l})^*$ and $(\boldsymbol{Z}_0^{\theta_l})^*$, $l = 1, \ldots, m$, and assume

3. $\mathrm{rk}\left(\boldsymbol{U}^{\theta_l}\right) = \mathrm{rk}\left((\boldsymbol{U}^{\theta_l})^*\right) = r^{\theta_l}$, $\mathrm{rk}(\boldsymbol{X}^{\theta_l}) = \mathrm{rk}\left((\boldsymbol{X}^{\theta_l})^*\right) = r^{\theta_l} + q^{\theta_l}$, $\mathrm{rk}(\boldsymbol{U}^{\theta_l}, \boldsymbol{Z}^{\theta_l}) = \mathrm{rk}\left((\boldsymbol{U}^{\theta_l})^*, (\boldsymbol{Z}^{\theta_l})^*\right) = r^{\theta_l} + t^{\theta_l}$ for $t^{\theta_l} \geq 0$ and $\mathrm{rk}\left((\boldsymbol{Z}_0^{\theta_l})^*\right) \geq \tilde{k}_0$ for $l = 1, \ldots, m$.

Note that the rank assumptions for $\boldsymbol{Z}_0^*$ allow to select $\tilde{k}_0$ linear independent rows of $(\boldsymbol{Z}_0^{\theta_1})^*, \ldots, (\boldsymbol{Z}_0^{\theta_m})^*$ corresponding to a subset $\{i_1, \ldots, i_{\tilde{k}_0}\} \subset \{1, \ldots, n^*\}$ of observations. The submodel to those observations is denoted by

$$\boldsymbol{\eta}_s^{\theta_l} = \boldsymbol{U}_s^{\theta_l}\boldsymbol{\gamma}^{\theta_l} + \boldsymbol{Z}_s^{\theta_l}\boldsymbol{\beta}^{\theta_l} + \boldsymbol{Z}_{0s}^{\theta_l}\boldsymbol{\beta}_0^{\theta_l}, \tag{C.3}$$

with accordingly submatrices $\boldsymbol{U}_s^{\theta_l}, \boldsymbol{Z}_s^{\theta_l}, \boldsymbol{Z}_{0s}^{\theta_l}$, $l = 1, \ldots, m$. This enables to use the arguments of Fahrmeir and Kneib [2009] who themselves refer to Theorem 3 in Sun et al. [2001]. Then we claim

4. $\mathrm{rk}(\boldsymbol{X}_s^{\theta_l}) = \mathrm{rk}\left((\boldsymbol{X}^{\theta_l})^*\right) = r^{\theta_l} + q^{\theta_l}$, $\mathrm{rk}(\boldsymbol{U}_s^{\theta_l}, \boldsymbol{Z}_s^{\theta_l}) = r^{\theta_l} + \tilde{t}^{\theta_l}$ for $\tilde{t}^{\theta_l} < t^{\theta_l}$.

Conditions 1. and 2. correspond to conditions (i) and (ii) in Fahrmeir and Kneib [2009] and they themselves to (B1) and (B2) of Sun et al. [2001]. There, the case of individual-specific effects is assumed. Fahrmeir and Kneib [2009] note that then condition 4. is not needed. Last, we assume

5. $k_j^{\theta_l} + 2a_j^{\theta_l} > \sum_{j=1}^{p^{\theta_l}} k_j^{\theta_l} - \tilde{t}^{\theta_l} + q^{\theta_l},\ j = 1, \ldots, p^{\theta_l},\ l = 1, \ldots, m,$

6. $\tilde{k}_0 - r^{\theta_l} - q^{\theta_l} + 2a_0^- + 2\sum_{j=1}^{p^{\theta_l}} (a_j^-)^{\theta_l} > 0,\ (a_j^-)^{\theta_l} = \min(0, a_j^{\theta_l}),\ l = 1, \ldots, m,$

where $a_j^{\theta_l}, b_j^{\theta_l}$ are the parameters of the inverse gamma prior for $(\tau^2)^{\theta_l}$.

Submodel (C.3) can for $l = 1, \ldots, m$, be rewritten in mixed model representation as

$$\boldsymbol{\eta}_s^{\theta_l} = \boldsymbol{X}_s^{\theta_l}\boldsymbol{\xi}^{\theta_l} + \boldsymbol{V}_s^{\theta_l}\boldsymbol{b}^{\theta_l} + \boldsymbol{V}_{0s}^{\theta_l}\boldsymbol{b}_0^{\theta_l}, \quad \boldsymbol{V}_{0s}^{\theta_l}\boldsymbol{b}_0^{\theta_l} \sim N\left(0, \tau_0^{\theta_l}\boldsymbol{V}_{0s}^{\theta_l}(\boldsymbol{V}_{0s}^{\theta_l})'\right) \tag{C.4}$$

and $\boldsymbol{V}_{0s}^{\theta_l}(\boldsymbol{V}_{0s}^{\theta_l})'$ has full rank $\tilde{k}_0$. This leads to a normalized model,

$$\boldsymbol{\nu}^{\theta_l} = \tilde{\boldsymbol{X}}^{\theta_l}\boldsymbol{\xi}^{\theta_l} + \tilde{\boldsymbol{V}}^{\theta_l}\boldsymbol{b}^{\theta_l} + \boldsymbol{\epsilon}^{\theta_l}, \quad \boldsymbol{\epsilon}^{\theta_l} \sim N\left(0, \tau_0^{\theta_l}\boldsymbol{I}\right) \tag{C.5}$$

via multiplication with $\left(\boldsymbol{V}_{0s}^{\theta_l}(\boldsymbol{V}_{0s}^{\theta_l})'\right)^{-1/2}$, compare Fahrmeir and Kneib [2009]. As $\boldsymbol{V}_{0s}^{\theta_l}(\boldsymbol{V}_{0s}^{\theta_l})'$ has full rank, we also have that $\left(\boldsymbol{V}_{0s}^{\theta_l}(\boldsymbol{V}_{0s}^{\theta_l})'\right)^{-1/2}$ is regular for $l = 1, \ldots, m$, and hence, condition 4. also holds for the normalized submodel (C.5). Showing that the posterior $p(\boldsymbol{\xi}^{\theta_1}, \ldots, \boldsymbol{\xi}^{\theta_m}, \boldsymbol{b}^{\theta_1}, \ldots, \boldsymbol{b}^{\theta_m}, \boldsymbol{b}_0^{\theta_1}, \ldots, \boldsymbol{b}_0^{\theta_m}, (\tau^2)^{\theta_1}, (\tau^2)^{\theta_m}, (\tau_0^2)^{\theta_1}, (\tau_0^2)^{\theta_m}|\boldsymbol{y})$ is proper is obviously equivalent to show the propriety of $p(\boldsymbol{\xi}^{\theta_1}, \ldots, \boldsymbol{\xi}^{\theta_m}, \boldsymbol{b}^{\theta_1}, \ldots, \boldsymbol{b}^{\theta_m}, \boldsymbol{\nu}^{\theta_1}, \ldots, \boldsymbol{\nu}^{\theta_m}, (\tau^2)^{\theta_1}, (\tau^2)^{\theta_m}, (\tau_0^2)^{\theta_1}, (\tau_0^2)^{\theta_m}|\boldsymbol{y})$. First of all it is

$$\begin{aligned}
& p(\boldsymbol{\xi}^{\theta_1}, \ldots, \boldsymbol{\xi}^{\theta_m}, \boldsymbol{b}^{\theta_1}, \ldots, \boldsymbol{b}^{\theta_m}, \boldsymbol{\nu}^{\theta_1}, \ldots, \boldsymbol{\nu}^{\theta_m}, (\tau^2)^{\theta_1}, \ldots, (\tau^2)^{\theta_m}, (\tau_0^2)^{\theta_1}, \ldots, (\tau_0^2)^{\theta_m}|\boldsymbol{y}) \\
\propto\ & \prod_{i=1}^{n} f(y_i|\eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m}) \prod_{l=1}^{m} p(\boldsymbol{\nu}^{\theta_l}|\boldsymbol{\xi}^{\theta_l}, \boldsymbol{b}^{\theta_l}, (\tau_0^2)^{\theta_l}) \prod_{l=1}^{m} p(\boldsymbol{b}^{\theta_l}|(\tau^2)^{\theta_l}) \\
& \times \prod_{l=1}^{k} p((\tau^2)^{\theta_l}) \prod_{l=1}^{m} p((\tau_0^2)^{\theta_l}) \\
\propto\ & \prod_{i=1}^{n} f(y_i|\eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m}) G,
\end{aligned}$$

where

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

and $\boldsymbol{Q}^{\theta_l} = \mathrm{Cov}(\boldsymbol{b}^{\theta_l})$. From assumption 2. we get

$$p(\boldsymbol{\xi}^{\theta_1}, \ldots, \boldsymbol{\xi}^{\theta_m}, \boldsymbol{b}^{\theta_1}, \ldots, \boldsymbol{b}^{\theta_m}, \boldsymbol{\nu}^{\theta_1}, \ldots, \boldsymbol{\nu}^{\theta_m}, (\tau^2)^{\theta_1}, (\tau^2)^{\theta_m}, (\tau_0^2)^{\theta_1}, (\tau_0^2)^{\theta_m} | \boldsymbol{y})$$
$$\leq M^* \prod_{i=1}^{n^*} f(y_i | \eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m}) G,$$

with $M^* = M^{n-n^*}$. Integrating over $\boldsymbol{\xi}^{\theta_1}, \boldsymbol{\xi}^{\theta_m}, \boldsymbol{b}^{\theta_1}, \boldsymbol{b}^{\theta_m}, (\boldsymbol{\tau}^2)^{\theta_1}, \ldots, (\boldsymbol{\tau}^2)^{\theta_m}$ implies

$$p(\boldsymbol{\nu}^{\theta_1}, \ldots, \boldsymbol{\nu}^{\theta_m}, (\tau_0^2)^{\theta_1}, (\tau_0^2)^{\theta_m} | \boldsymbol{y})$$
$$\leq M^* \prod_{i=1}^{n} f(y_i | \eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m}) \int G \mathrm{d}\boldsymbol{\xi}^{\theta_1}, \ldots, \mathrm{d}\boldsymbol{\xi}^{\theta_m}, \mathrm{d}\boldsymbol{b}^{\theta_1}, \ldots, \mathrm{d}\boldsymbol{b}^{\theta_m}, \mathrm{d}(\boldsymbol{\tau}^2)^{\theta_1}, \ldots, \mathrm{d}(\boldsymbol{\tau}^2)^{\theta_m}.$$

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 $\int G < \infty$. It can be bounded by their expressions (A.25), if $\tilde{t}^{\theta_l} = q^{\theta_l}$ for $l = 1, \ldots, m$, and by (A.27), if $\tilde{t}^{\theta_l} < q^{\theta_l}$ for any $l \in \{1, \ldots, m\}$. For some constant $\tilde{M}$ we therefore get the inequality

$$p(\boldsymbol{\nu}^{\theta_1}, \ldots, \boldsymbol{\nu}^{\theta_m}, (\tau^2)^{\theta_1}, (\tau^2)^{\theta_m}, (\tau_0^2)^{\theta_1}, (\tau_0^2)^{\theta_m} | \boldsymbol{y}) \quad \leq \quad \tilde{M} \prod_{i=1}^{n^*} f(y_i | \eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m}) g\left((\tau_0^2)^{\theta_1}, \ldots, (\tau_0^2)^{\theta_m}\right),$$

with

$$g\left((\tau_0^2)^{\theta_1}, \ldots, (\tau_0^2)^{\theta_m}\right) \quad = \quad \prod_{l=1}^{k} \left(\frac{1}{(\tau_0^2)^{\theta_l}}\right)^{-(\tilde{k}_0 - r^{\theta_l} - q^{\theta_l})/2 - a_0^- - \sum_{j=1}^{p^{\theta_l}} (a_j^-)^{\theta_l}} \exp\left(-\frac{\mathrm{SSE}_s^{\theta_l} + 2b_0^{\theta_l}}{2(\tau_0^2)^{\theta_l}}\right)$$

and

$$\mathrm{SSE}_s^{\theta_l} := (\boldsymbol{\nu} - \tilde{\boldsymbol{X}}^{\theta_l} \boldsymbol{\xi}^{\theta_l} - \tilde{\boldsymbol{V}}^{\theta_l} \boldsymbol{b}^{\theta_l})'(\boldsymbol{\nu} - \tilde{\boldsymbol{X}}^{\theta_l} \boldsymbol{\xi}^{\theta_l} - \tilde{\boldsymbol{V}}^{\theta_l} \boldsymbol{b}^{\theta_l}). \tag{C.6}$$

Assumption 6., and $\mathrm{SSE}_s^{\theta_l} + 2b_0^{\theta_l} > 0$ for $l = 1, \ldots, m$, imply $\int g\left((\tau_0^2)^{\theta_1}, \ldots, (\tau_0^2)^{\theta_m}\right) \mathrm{d}(\boldsymbol{\tau}^2)^{\theta_1}, \ldots, \mathrm{d}(\boldsymbol{\tau}^2)^{\theta_m} < \infty$ and therefore for some constant C

$$p(\boldsymbol{\nu}^{\theta_1}, \ldots, \boldsymbol{\nu}^{\theta_m} | \boldsymbol{y}) \quad \leq \quad C \prod_{i=1}^{n^*} f(y_i | \eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m}).$$

Finally, it remains $\int p(\boldsymbol{\nu}^{\theta_1}, \ldots, \boldsymbol{\nu}^{\theta_m} | \boldsymbol{y}) \mathrm{d}\boldsymbol{\nu}^{\theta_1}, \ldots, \mathrm{d}\boldsymbol{\nu}^{\theta_m} < \infty$ to show. This is received using the relation $\boldsymbol{\nu}^{\theta_l} = \left(\boldsymbol{V}_{0s}^{\theta_l} (\boldsymbol{V}_{0s}^{\theta_l})'\right)^{-1/2} \boldsymbol{\eta}_s^{\theta_l}$ for $l = 1, \ldots, m$:

$$\int p(\boldsymbol{\nu}^{\theta_1}, \ldots, \boldsymbol{\nu}^{\theta_m} | \boldsymbol{y}) \mathrm{d}\boldsymbol{\nu}^{\theta_1}, \ldots, \mathrm{d}\boldsymbol{\nu}^{\theta_m}$$
$$= \int p(\boldsymbol{\eta}_s^{\theta_1}, \ldots, \boldsymbol{\eta}_s^{\theta_m} | \boldsymbol{y}) \det\left(\left(\boldsymbol{V}_{0s}^{\theta_1} (\boldsymbol{V}_{0s}^{\theta_1})'\right)^{-1/2}\right) \mathrm{d}\boldsymbol{\eta}_s^{\theta_1}, \ldots, \det\left(\left(\boldsymbol{V}_{0s}^{\theta_m} (\boldsymbol{V}_{0s}^{\theta_m})'\right)^{-1/2}\right) \mathrm{d}\boldsymbol{\eta}_s^{\theta_m}$$
$$\leq K \int \prod_{i=1}^{n^*} f_i(y_i | \eta_i^{\theta_1}, \ldots, \eta_i^{\theta_m}) \mathrm{d}\eta_i^{\theta_1} \ldots \mathrm{d}\eta_i^{\theta_m} < \infty$$

for some constant K. □

D Additional Graphics to Simulation Studies of the Main Paper

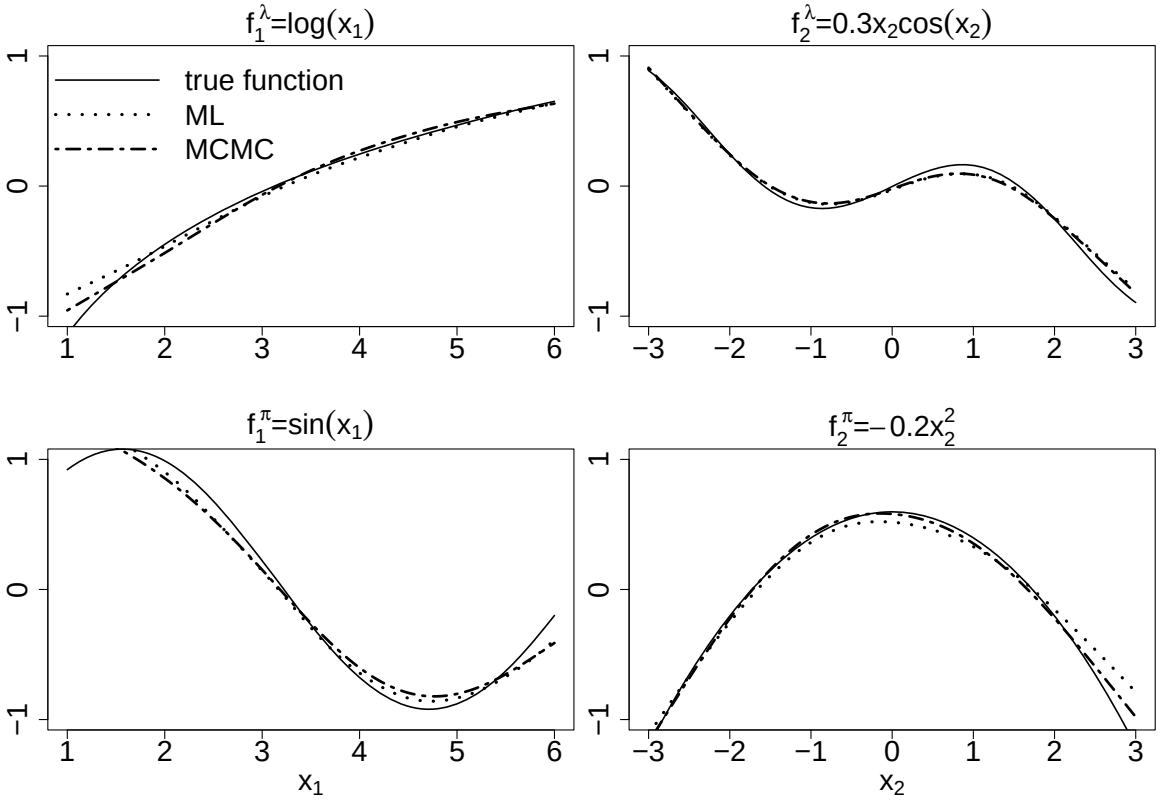

Figure D1: ZIP additive model. True curves of nonlinear effects together with overall mean ML and MCMC estimates

E Further Simulation Studies

E.1 Negative Binomial Regression

E.1.1 Additive Models

For reasons of simplicity we keep the study design 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 parameters μ and δ of a negative binomial distribution, is written as the sum of two nonlinear functions f_1 and f_2,

$$f_1^\mu(\boldsymbol{x}_1) = f_1^\mu(\boldsymbol{x}_1) = \log(\boldsymbol{x}_1), \quad f_2^\mu(\boldsymbol{x}_2) = f_2^\mu(\boldsymbol{x}_2) = 0.3\boldsymbol{x}_2 \cos(\boldsymbol{x}_2)$$

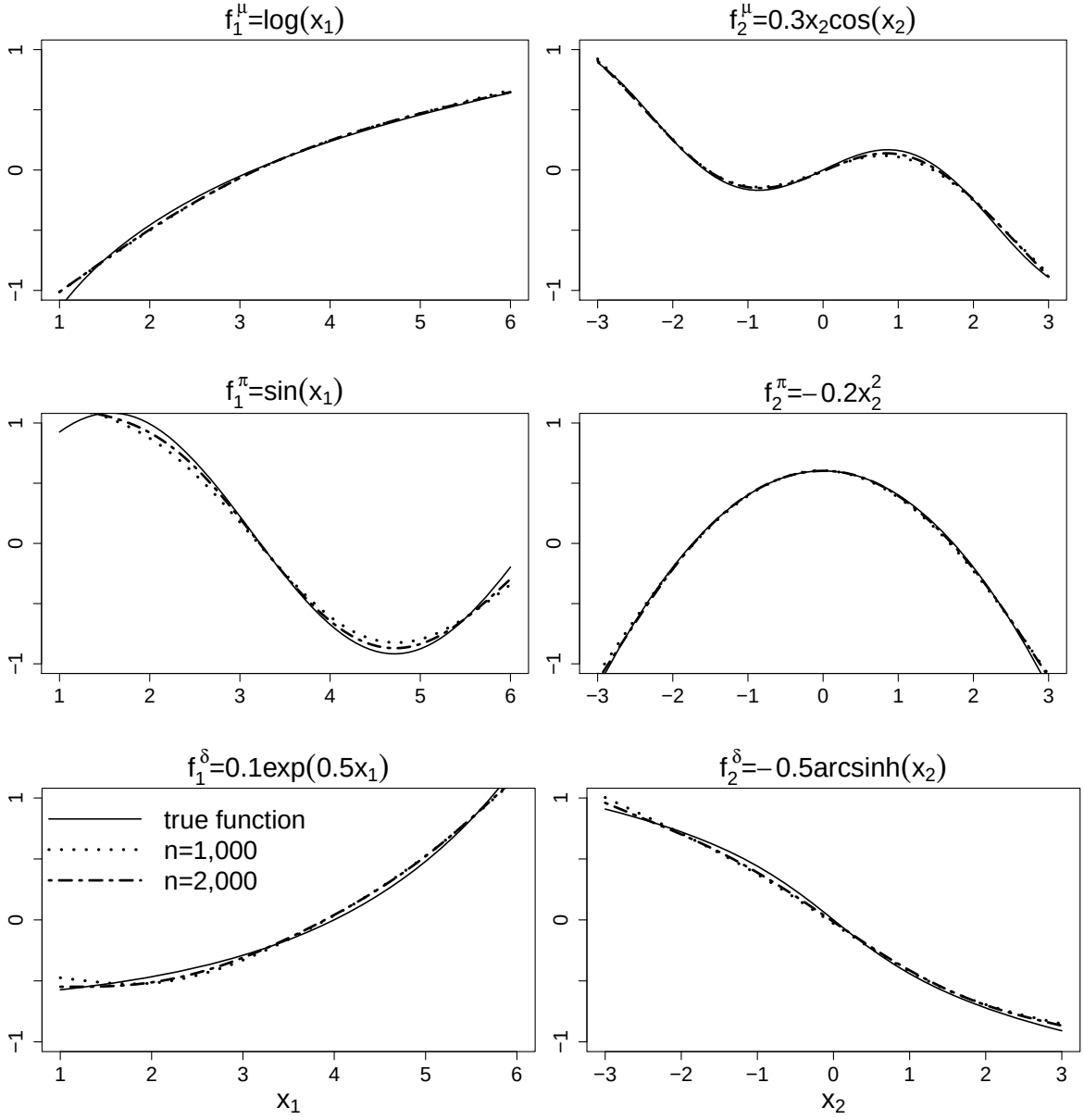

Figure D2: ZINB additive model. True curves of nonlinear effects together with overall mean MCMC estimates

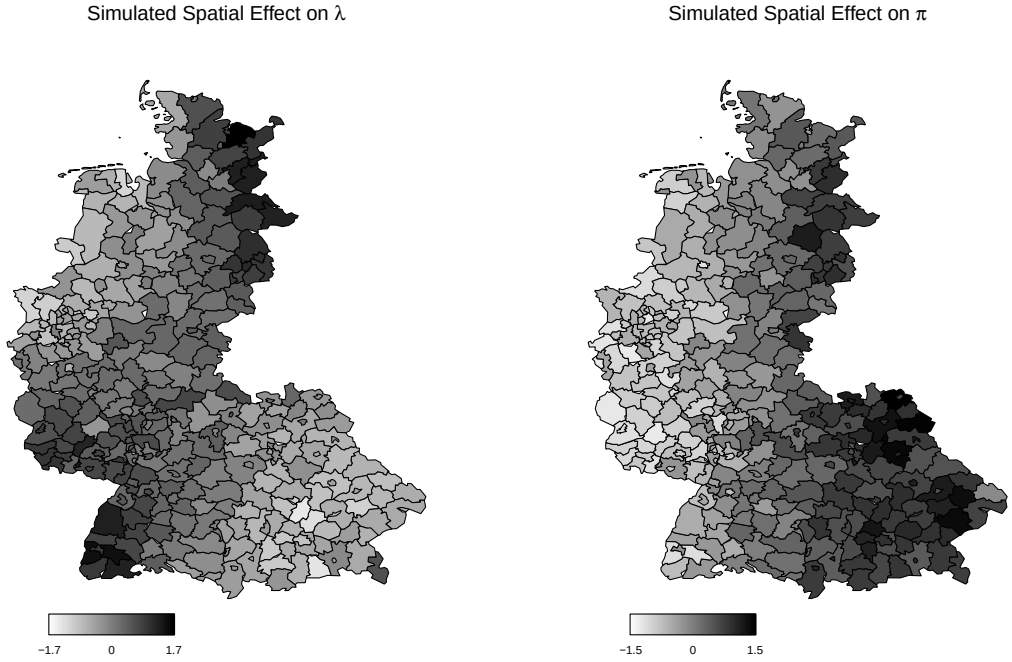

Figure D3: ZIP geoadditive model. Simulated complete spatial effects

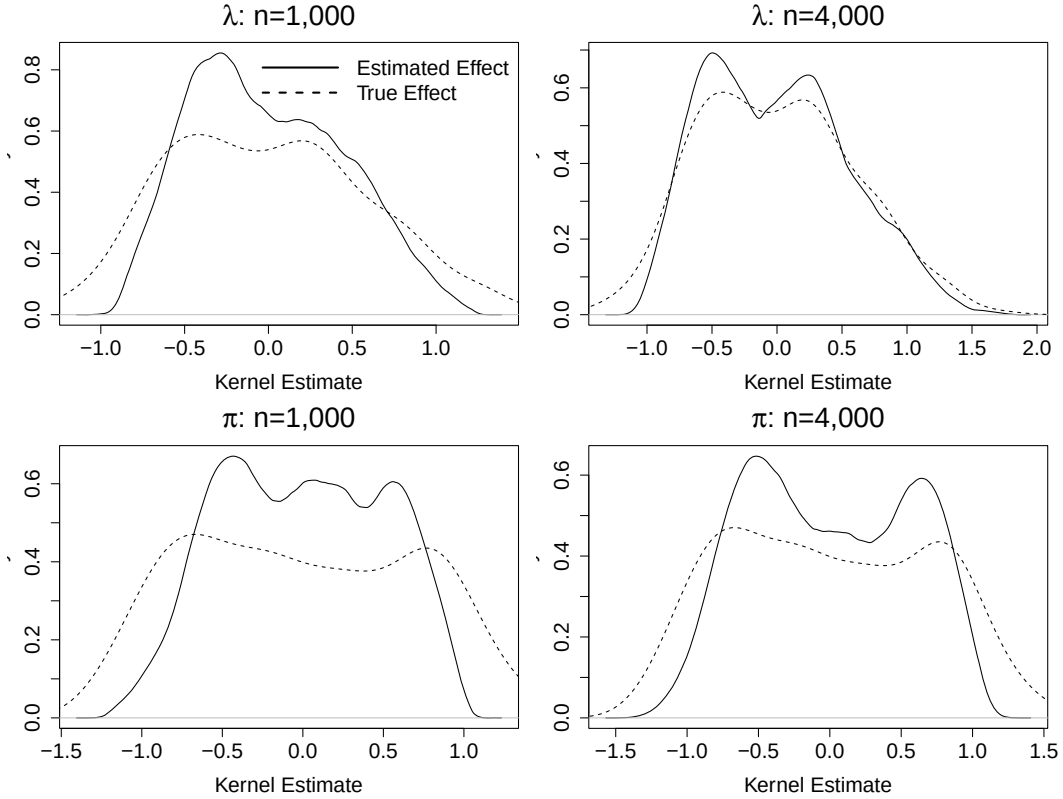

Figure D4: ZIP geoadditive model. Kernel density estimates of complete spatial effects

$$f_1^\delta(\boldsymbol{x}_1) = 0.1\exp(0.5\boldsymbol{x}_1), \quad f_2^\delta(\boldsymbol{x}_2) = -0.5\,\text{arcsinh}(\boldsymbol{x}_2),$$

with covariates x_1, x_2 as defined in Section 4.1. All other settings are identical to the ones described in the paper. Similar to the ZINB study, we simulated 250 replications for sample sizes $n = 1,000$ and $n = 2,000$. In Figure E5, the mean squared errors for both samples sizes are summarized in terms of boxplots. Pointwise 95% credible

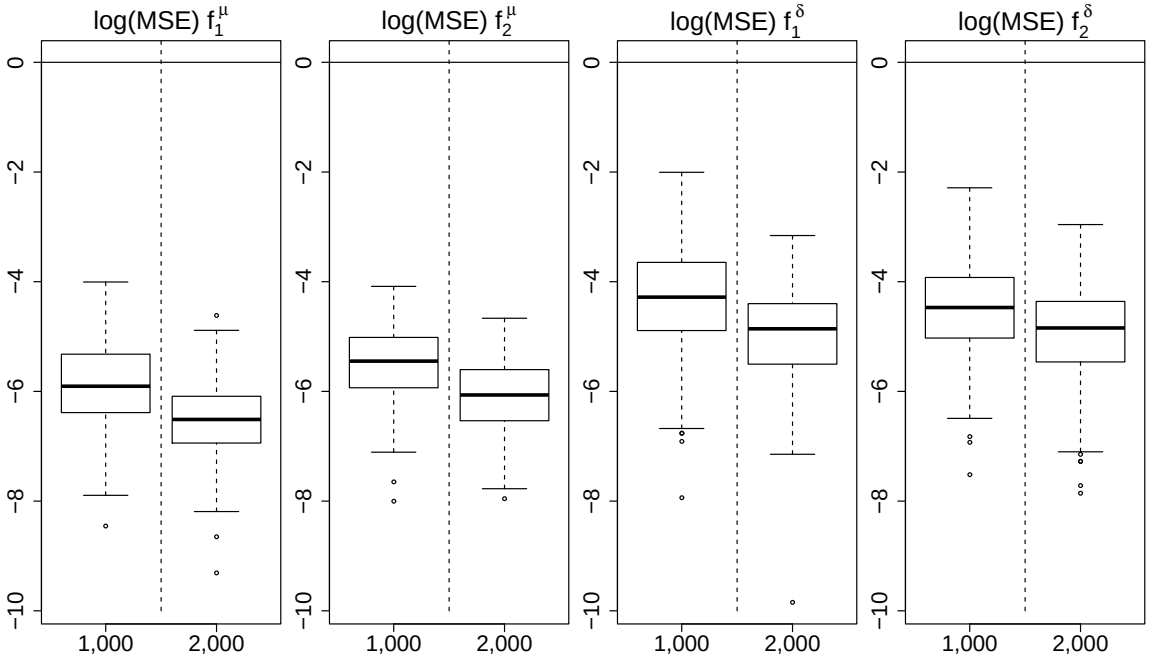

Figure E5: NB additive model. log(MSE) of MCMC

intervals can be compared in Figure E6. Figure E7 shows the average of mean estimates of all 250 simulated replications. Similar to the corresponding studies of ZIP and ZINB results for the NB model can briefly be summed up as follows:

- Bias: Figure E7 indicates that in average we obtain satisfactory mean estimates for all nonlinear effects. The effects on δ are more difficult to estimate compared to μ.
- MSE: The logarithmic mean squared errors in Figure E5 of the effects on δ are higher compared to the ones of μ. Increasing the sample size can reduce the MSE in all effects.
- Pointwise coverage rates: 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 wide confidence intervals.

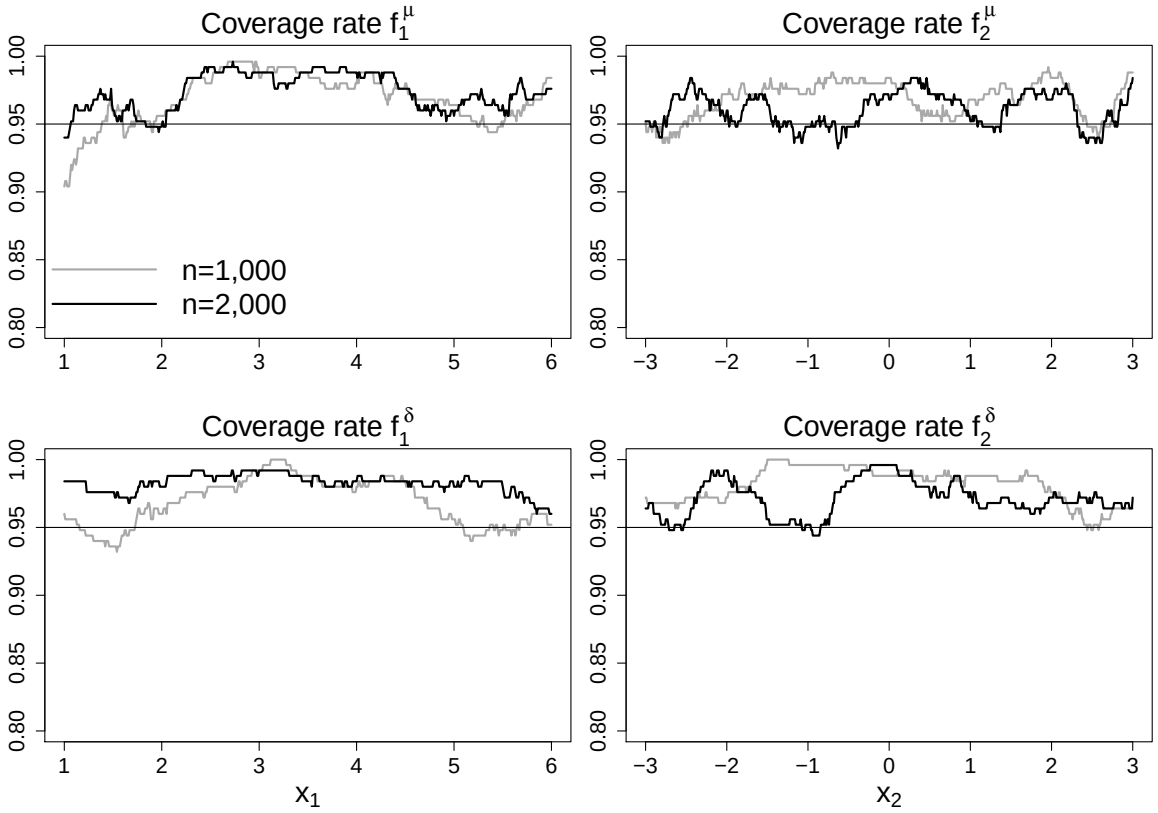

Figure E6: NB additive model. Pointwise 95% coverage rates of MCMC

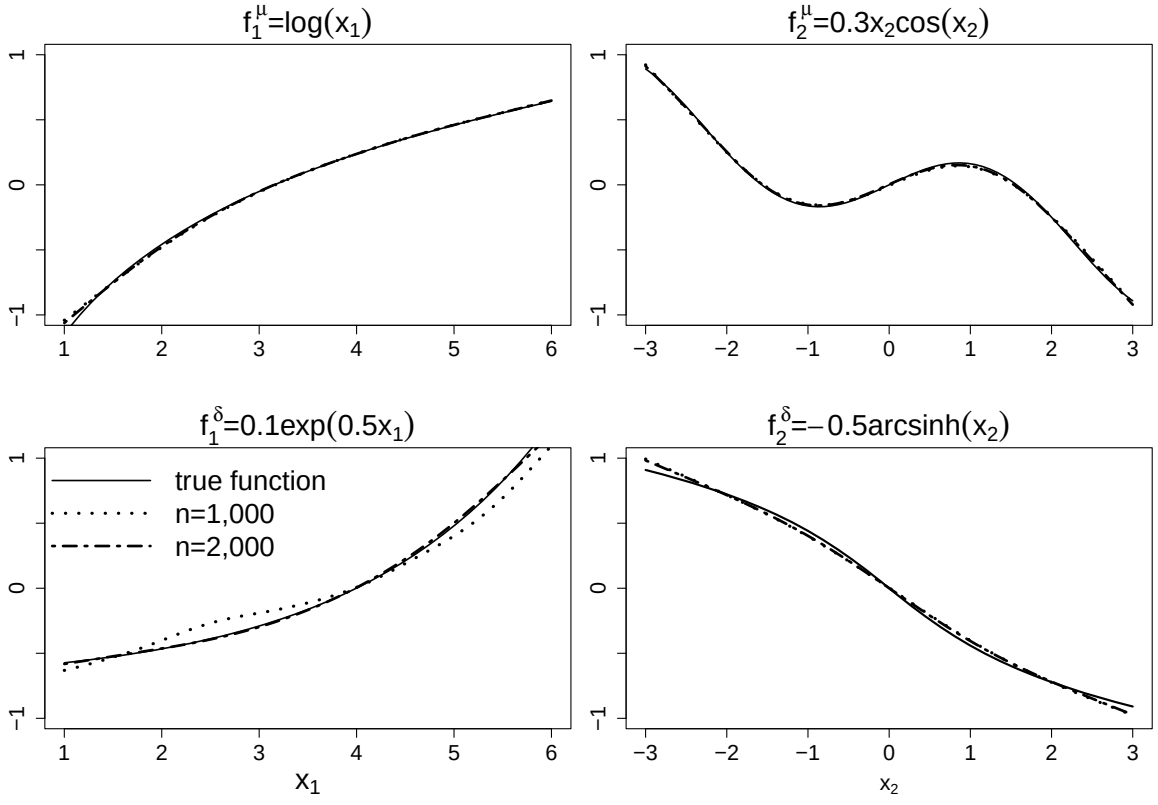

Figure E7: NB additive model. True curves of nonlinear effects together with overall mean MCMC estimates

In conclusion, it can be said that 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 levels is kept or even higher 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 uncertainties.

E.1.2 Geoadditive Models

In order to extend the additive model with a spatial effect, we proceed as in the ZIP and ZINB case. The nonlinear effects retain the ones of Section E.1.1. Both parameters μ and δ are simulated with an additional spatial effect, consisting of an unstructured part, generated by $\epsilon \sim \mathrm{N}(0, 0.125)$ and a structured part, modeled by a Markov random field and simulated as

$$\begin{aligned} f^{\mu}_{\text{spat}}(l) &= \sin(x^c_l y^c_l) + \epsilon^{\mu}_l \\ f^{\delta}_{\text{spat}}(l) &= 0.5 x^c_l y^c_l + \epsilon^{\delta}_l. \end{aligned}$$

Again, $l \in \{1, \ldots, S\}$ describes one of the $S = 327$ districts in Western Germany. In Figure E8, the results of the simulated complete spatial effects are visualized. As before, we performed studies for four different sample sizes $n = 1,000$, $n = 2,000$, $n = 4,000$ and $n = 16,000$, each consisting of 250 replications. Again, we restrict 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 E9. Beside this, the log(MSE) in Figure E10 and the kernel estimates of the complete spatial effects in Figure E11 give further information about the quality of the inference.

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 can improve our estimates. The complete spatial effect on δ is underestimated and too smooth but with a sample size of $n = 4,000$ it gets closer to the true effect.

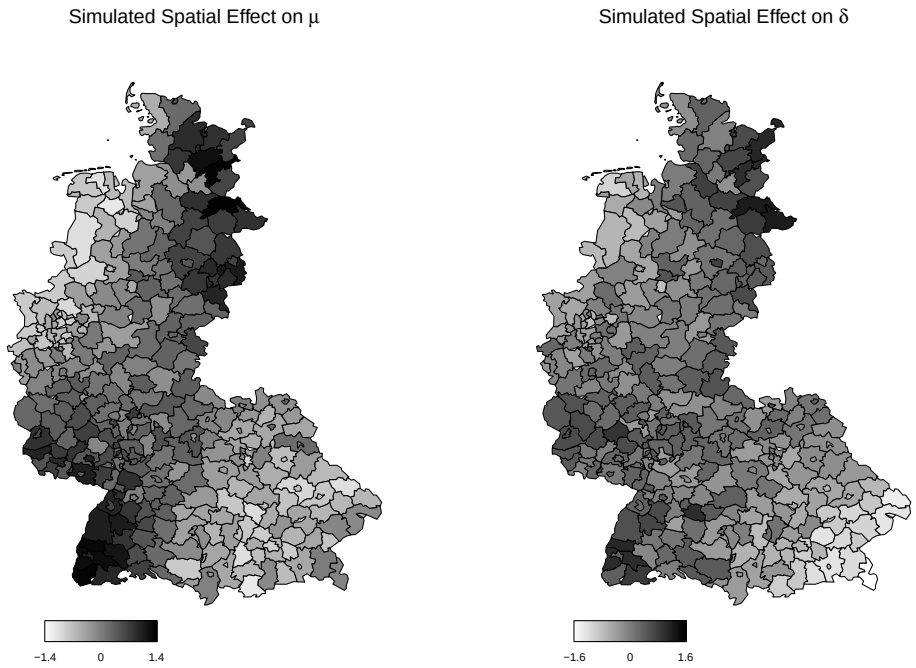

Figure E8: NB geoadditive model. Simulated complete spatial effects

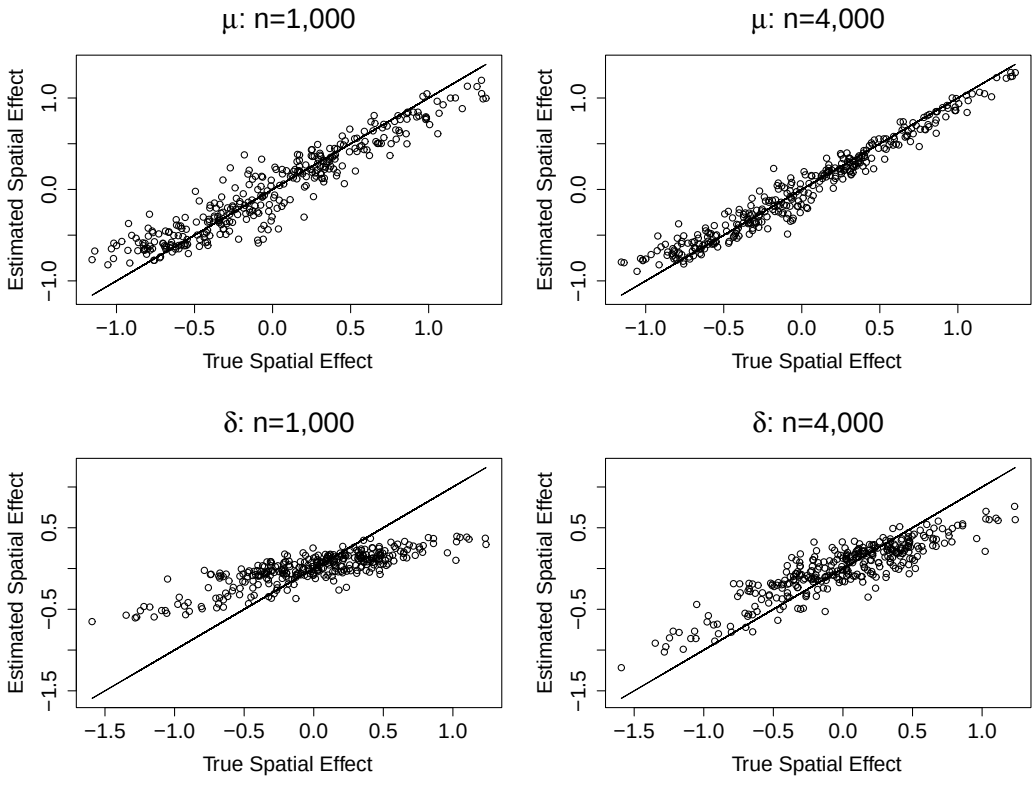

Figure E9: NB geoadditive model. Estimated complete spatial effects

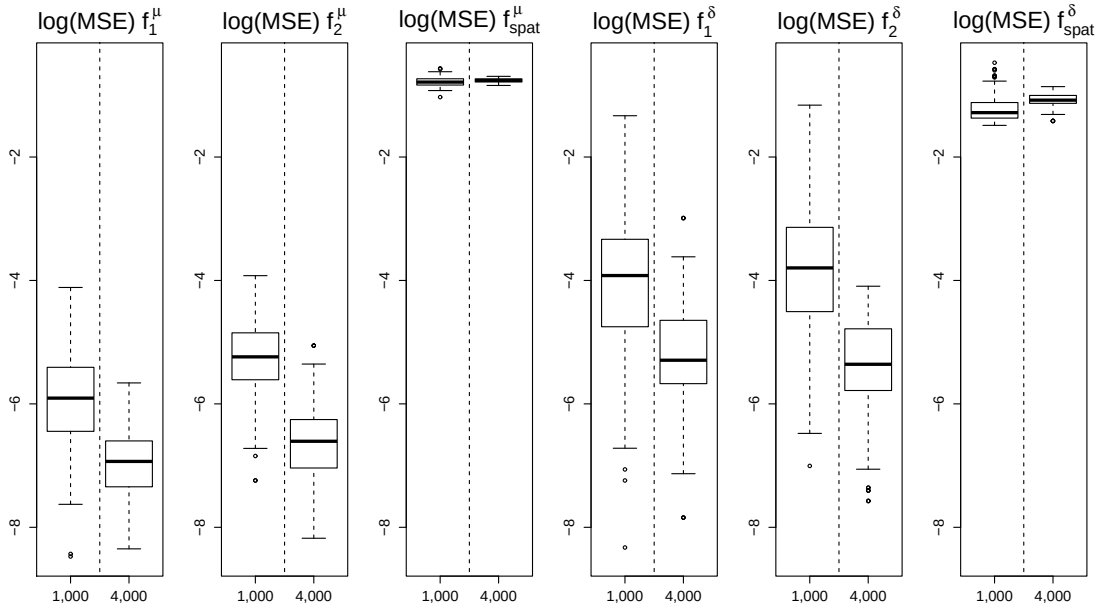

Figure E10: NB geoadditive model. log(MSE) of nonlinear and complete spatial effects

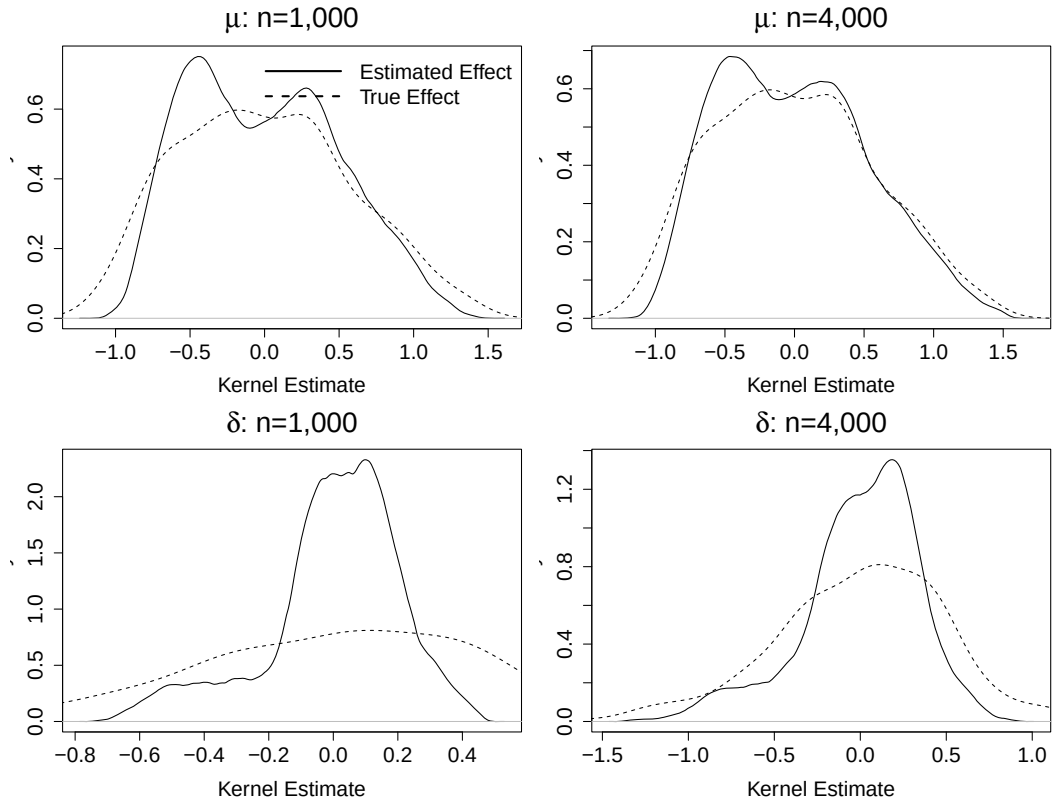

Figure E11: NB geoadditive model. Kernel density estimates of complete spatial effects

E.2 Zero-Inflated Negative Binomial Regression

E.2.1 Geoadditive Models

Similar to the structure of the corresponding study for the ZIP model, we performed simulations for sample sizes $n = 1,000,\ 2,000,\ 4,000$ and $16,000$. All settings are given in Section 4 of the main paper. The true model is defined as follows:

$$f_1^{\mu}(\boldsymbol{x}_1) = \log(\boldsymbol{x}_1), \quad f_2^{\mu}(\boldsymbol{x}_2) = 0.3\boldsymbol{x}_2 \cos(\boldsymbol{x}_2)$$

$$f_1^{\pi}(\boldsymbol{x}_1) = \sin(\boldsymbol{x}_1), \quad f_2^{\pi}(\boldsymbol{x}_2) = -0.2\boldsymbol{x}_2^2$$

$$f_1^{\delta}(\boldsymbol{x}_1) = 0.1 \exp(0.5\boldsymbol{x}_1), \quad f_2^{\delta}(\boldsymbol{x}_2) = 0.5 \operatorname{arcsinh}(\boldsymbol{x}_2),$$

where the covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained as i.i.d. samples from equidistant grids of steps 0.01, such that for $i = 1, \ldots, n$, we have $x_{i1} \in [1, 6]$ and $x_{i2} \in [-3, 3]$. The complete spatial effect is obtained as

$$\begin{aligned} f_{\text{spat}}^{\mu}(l) &= \sin(x_l^c y_l^c) + \epsilon_l^{\lambda} \\ f_{\text{spat}}^{\pi}(l) &= \sin(x_l^c) \cos(0.5 y_l^c) + \epsilon_l^{\pi} \\ f_{\text{spat}}^{\delta}(l) &= 0.5 x_l^c y_l^c + \epsilon_l^{\delta}. \end{aligned}$$

and $\boldsymbol{\epsilon}^{\mu}$, $\boldsymbol{\epsilon}^{\pi}$, $\boldsymbol{\epsilon}^{\delta}$ are Gaussian with mean 0 and variance 1/16. Since we make use of the multilevel structure which was mentioned in Section 3.3, the random effects are estimated on a second level equation. The resulting complete spatial effects for all three model parameters are plotted in Figure E12. Our results of the study are summarized in Figures E13 and E14 where the logarithmic mean-squared errors of all 250 replications and for both sample sizes are visualized in form of boxplots and the estimated complete spatial effects can be compared to the simulated effects.

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

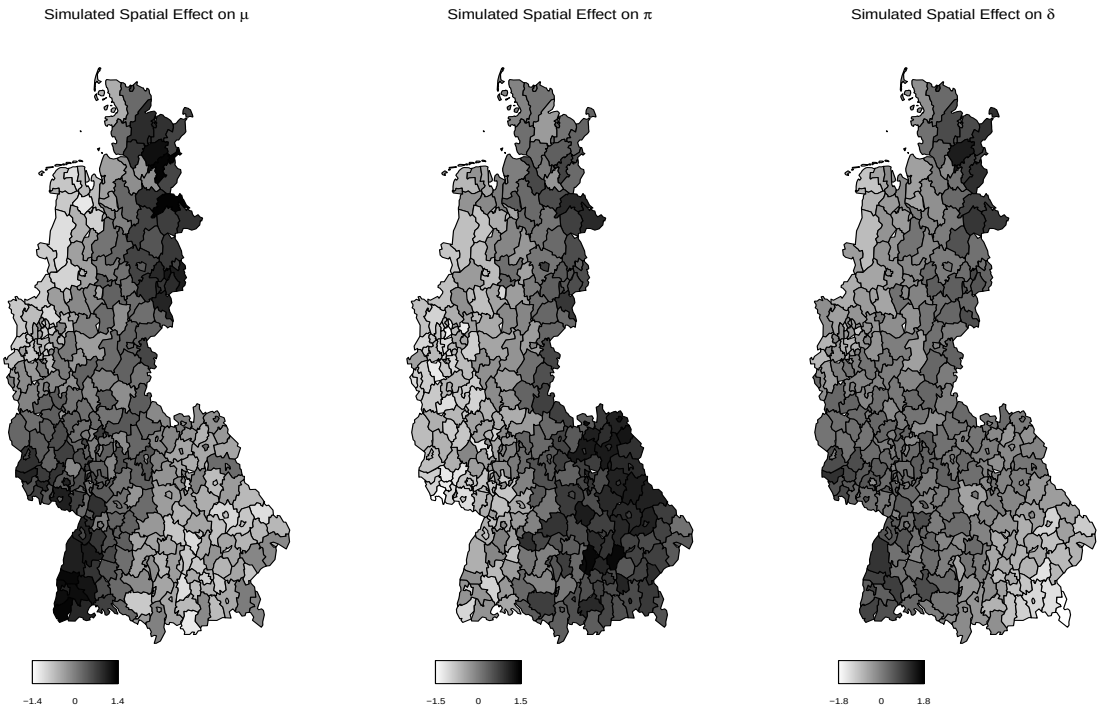

Figure E12: ZINB geoadditive model. Simulated complete spatial effects

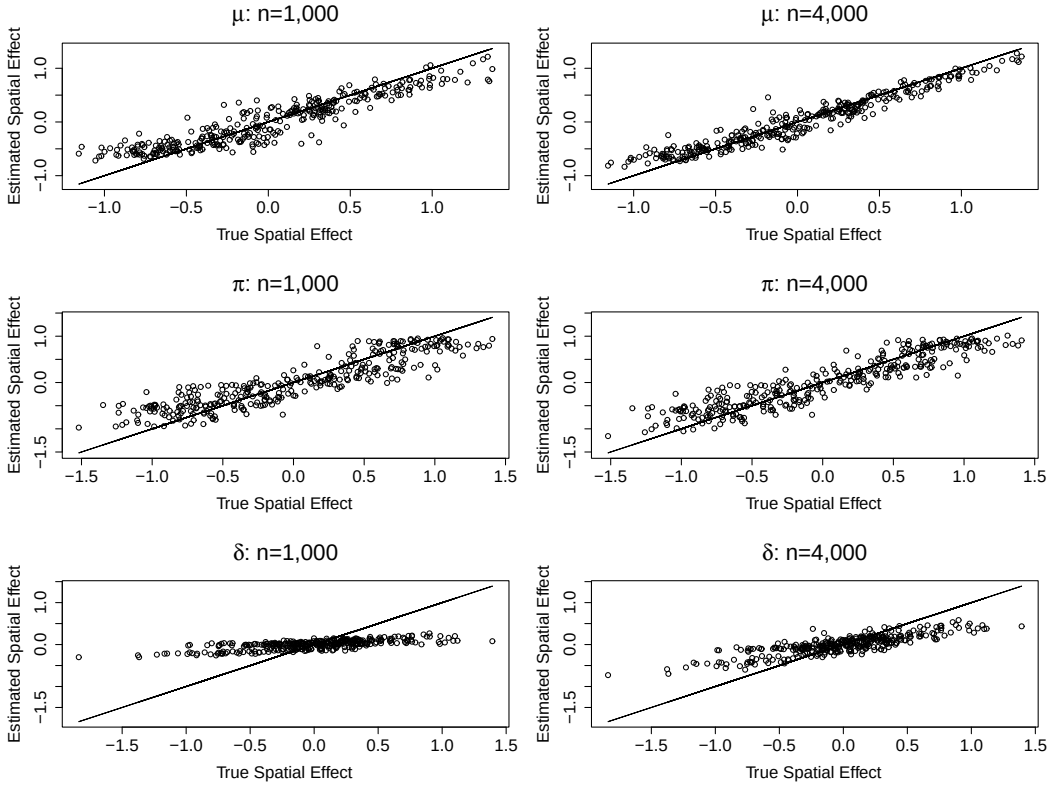

Figure E13: ZINB geoadditive model. Estimated complete spatial effects

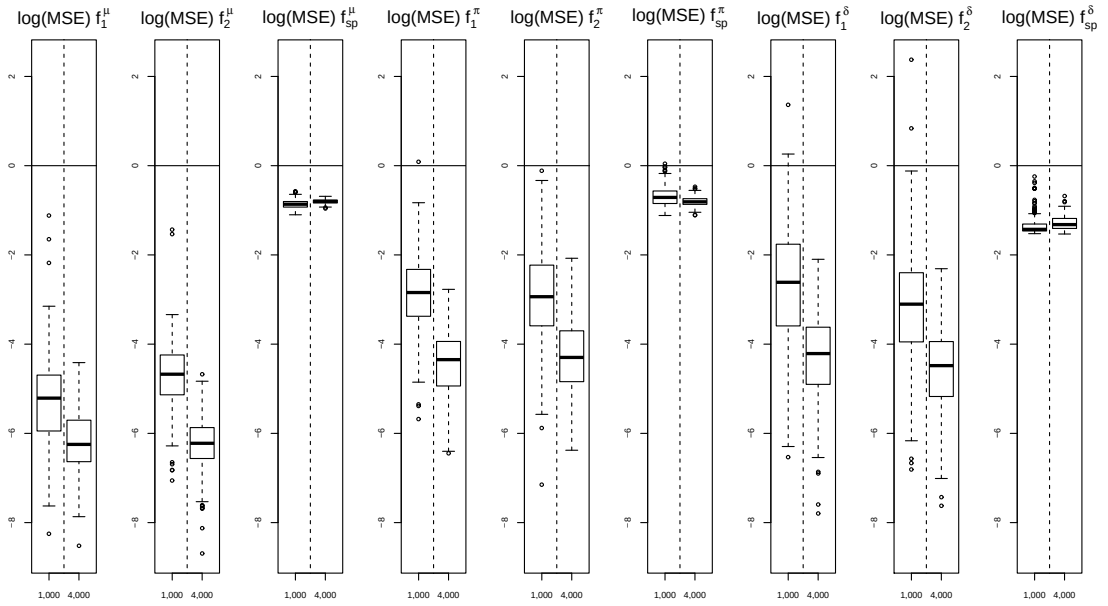

Figure E14: ZINB geoadditive model. log(MSE) of nonlinear and complete spatial effects

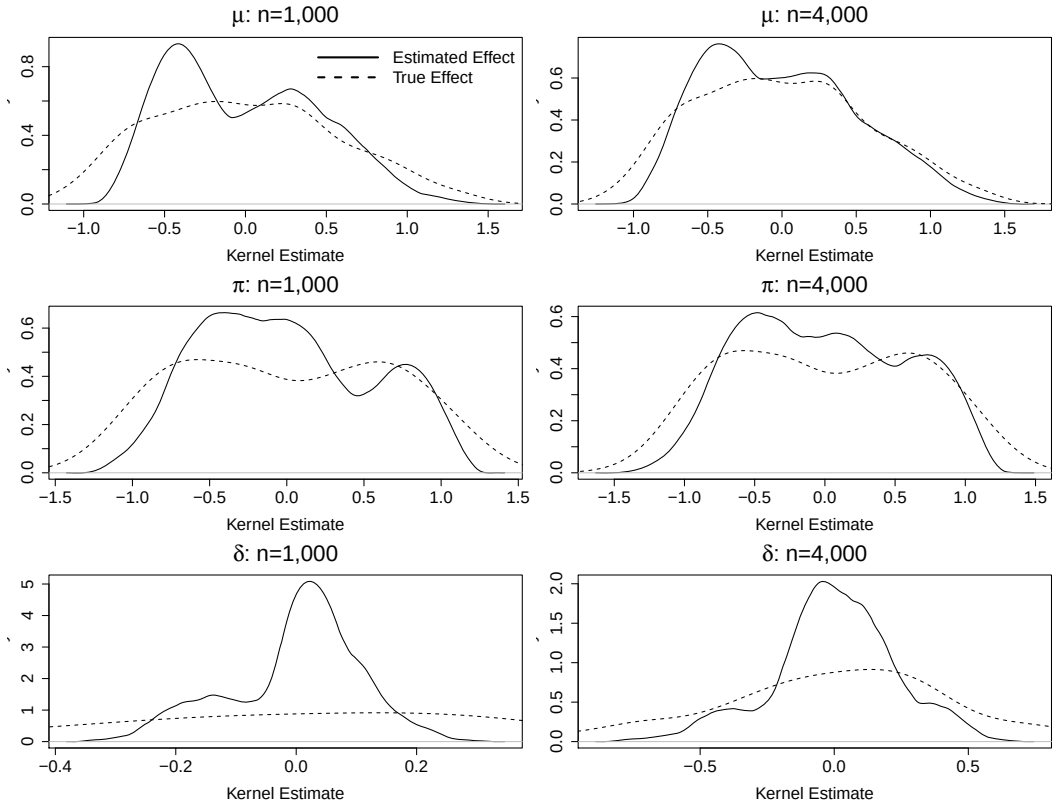

Figure E15: ZINB geoadditive model. Kernel density estimates of complete spatial effects

the rule of thumb where differences of more than 10 in DIC might "'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 selected by choosing the one with the lowest DIC.

F.1 Simulation Setup

The true models are defined as in Section 4.2. This means that one generic predictor η linked to a generic parameter of the ZIP or ZINB distribution is given by

$$\begin{aligned}\boldsymbol{\eta} &= f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) + \boldsymbol{f}_{\text{spat}} + \boldsymbol{\epsilon} \\ &= \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}_{\text{spat}}\boldsymbol{\beta}_{\text{spat}} + \boldsymbol{\epsilon}.\end{aligned}$$

Depending on the chosen model, we used the functions

$$f_1^\lambda(\boldsymbol{x}_1) = f_1^\mu(\boldsymbol{x}_1) = \log(\boldsymbol{x}_1), \quad f_2^\lambda(\boldsymbol{x}_2) = f_2^\mu(\boldsymbol{x}_2) = 0.3\boldsymbol{x}_2\cos(\boldsymbol{x}_2)$$

$$f_1^\pi(\boldsymbol{x}_1) = \sin(\boldsymbol{x}_1), \quad f_2^\pi(\boldsymbol{x}_2) = -0.2\boldsymbol{x}_2^2$$

$$f_1^\delta(\boldsymbol{x}_1) = 0.1\exp(0.5\boldsymbol{x}_1), \quad f_2^\delta(\boldsymbol{x}_2) = 0.5\,\text{arcsinh}(\boldsymbol{x}_2),$$

where the covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained as i.i.d. samples from equidistant grids of steps 0.01, such that for $i = 1, \ldots, n$, we have $x_{i1} \in [1, 6]$ and $x_{i2} \in [-3, 3]$. The complete spatial effect is obtained as

$$\begin{aligned}f_{\text{spat}}^\lambda(l) = f_{\text{spat}}^\mu(l) &= \sin(x_l^c y_l^c) + \epsilon_l \\ f_{\text{spat}}^\pi(l) &= \sin(x_l^c)\cos(0.5y_l^c) + \epsilon_l^\pi \\ f_{\text{spat}}^\delta(l) &= 0.5x_l^c y_l^c + \epsilon_l^\delta.\end{aligned}$$

and $\boldsymbol{\epsilon}^\lambda$, $\boldsymbol{\epsilon}^\mu$, $\boldsymbol{\epsilon}^\pi$, $\boldsymbol{\epsilon}^\delta$ are Gaussian with mean 0 and variance 1/16.

To validate the ability of the DIC to find the best model in geoadditive structured regression models for responses with NB, ZIP or ZINB distribution, we reestimated different misspecified versions of each of the three simulated models, compare Table F1 for a scheme of all tested candidates. All models are replicated 250 times for sample sizes $n = 1,000$ and $n = 4,000$ in each model.

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 F1: Description of estimated models

F.2 Results

Table F2 displays the percentage of times where the true/misspecified model was selected or non of the models would be preferred with the decision rule defined before. For all ZIP models, it can be stated that 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 where for sample size $n = 1,000$ the differences in DIC is in the range of 15 and in favor of the false model, applying the predefined rule, 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 is that the additional nonlinear effect was simulated in the interval $[-1, 1]$ which is similar to the

Model comparison	Select true model	indecisive	select misspecified model
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	11.2%/100%	88.8%/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 F2: Relative amount in percent of the decision made by DIC for $n = 1,000/4,000$ in the NB, ZIP and ZINB model

true covariate values. Furthermore it can stated the following:

- Included irrelevant effects: 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 into account another criterion. In particular, a covariate can be seen as irrelevant or not significant if the corresponding Bayesian credible interval contains zero. Applying this rule, the irrelevant effect would be excluded in at least 85% of all tested models and for all distributions where an irrelevant variable was included.

- Omitted relevant effects: 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 favor of the true model such that with our decision rule we would choose the correct model in at least 75.6% of the replications. 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 a effect is omitted. But in all these models in at least 11.2% we would prefer the true model whereas in the remaining cases the DIC is indecisive.

- Wrong specification: 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.

References

L. Fahrmeir and T. Kneib. Propriety of posteriors in structured additive regression models: Theory and empirical evidence. Journal of Statistical Planning and Inference, 39:843–859, 2009.

T.J. Hastie and R.J. Tibshirani. Generalized Additive Models. Chapman & Hall, 1990.

H. Rue and L. Held. Gaussian Markov Random Fields. Chapman & Hall / CRC, 2005.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, 65(B):583–639, 2002.

D. Sun, R.K. Tsutakawa, and H. Zhuoqiong. Propriety of posteriors with improper priors in hierarchical linear mixed models. Statistica Sinica, 11:77–95, 2001.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://eeecon.uibk.ac.at/wopec/

2013-12 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

2013-11 **Thomas Stöckl:** Price efficiency and trading behavior in limit order markets with competing insiders

2013-10 **Sebastian Prediger, Björn Vollan, Benedikt Herrmann:** Resource scarcity, spite and cooperation

2013-09 **Andreas Exenberger, Simon Hartmann:** How does institutional change coincide with changes in the quality of life? An exemplary case study

2013-08 **E. Glenn Dutcher, Loukas Balafoutas, Florian Lindner, Dmitry Ryvkin, Matthias Sutter:** Strive to be first or avoid being last: An experiment on relative performance incentives.

2013-07 **Daniela Glätzle-Rützler, Matthias Sutter, Achim Zeileis:** No myopic loss aversion in adolescents? An experimental note

2013-06 **Conrad Kobel, Engelbert Theurl:** Hospital specialisation within a DRG-Framework: The Austrian case

2013-05 **Martin Halla, Mario Lackner, Johann Scharler:** Does the welfare state destroy the family? Evidence from OECD member countries

2013-04 **Thomas Stöckl, Jürgen Huber, Michael Kirchler, Florian Lindner:** Hot hand belief and gambler's fallacy in teams: Evidence from investment experiments

2013-03 **Wolfgang Luhan, Johann Scharler:** Monetary policy, inflation illusion and the Taylor principle: An experimental study

2013-02 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Tensions between the resource damage and the private benefits of appropriation in the commons

2013-01 **Jakob W. Messner, Achim Zeileis, Jochen Broecker, Georg J. Mayr:** Improved probabilistic wind power forecasts with an inverse power curve transformation and censored regression

2012-27 **Achim Zeileis, Nikolaus Umlauf, Friedrich Leisch:** Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond

2012-26 **Francisco Campos-Ortiz, Louis Putterman, T.K. Ahn, Loukas Balafoutas, Mongoljin Batsaikhan, Matthias Sutter:** Security of property as a public good: Institutions, socio-political environment and experimental behavior in five countries

2012-25 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Appropriation in the commons: variations in the opportunity costs of conservation

2012-24 **Edgar C. Merkle, Jinyan Fan, Achim Zeileis:** Testing for measurement invariance with respect to an ordinal variable *forthcoming in Psychometrika*

2012-23 **Lukas Schrott, Martin Gächter, Engelbert Theurl:** Regional development in advanced countries: A within-country application of the Human Development Index for Austria

2012-22 **Glenn Dutcher, Krista Jabs Saral:** Does team telecommuting affect productivity? An experiment

2012-21 **Thomas Windberger, Jesus Crespo Cuaresma, Janette Walde:** Dirty floating and monetary independence in Central and Eastern Europe - The role of structural breaks

2012-20 **Martin Wagner, Achim Zeileis:** Heterogeneity of regional growth in the European Union

2012-19 **Natalia Montinari, Antonio Nicolo, Regine Oexl:** Mediocrity and induced reciprocity

2012-18 **Esther Blanco, Javier Lozano:** Evolutionary success and failure of wildlife conservancy programs

2012-17 **Ronald Peeters, Marc Vorsatz, Markus Walzl:** Beliefs and truth-telling: A laboratory experiment

2012-16 **Alexander Sebald, Markus Walzl:** Optimal contracts based on subjective evaluations and reciprocity

2012-15 **Alexander Sebald, Markus Walzl:** Subjective performance evaluations and reciprocity in principal-agent relations

2012-14 **Elisabeth Christen:** Time zones matter: The impact of distance and time zones on services trade

2012-13 **Elisabeth Christen, Joseph Francois, Bernard Hoekman:** CGE modeling of market access in services

2012-12 **Loukas Balafoutas, Nikos Nikiforakis:** Norm enforcement in the city: A natural field experiment *forthcoming in European Economic Review*

2012-11 **Dominik Erharter:** Credence goods markets, distributional preferences and the role of institutions

2012-10 **Nikolaus Umlauf, Daniel Adler, Thomas Kneib, Stefan Lang, Achim Zeileis:** Structured additive regression models: An R interface to BayesX

2012-09 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** History repeating: Spain beats Germany in the EURO 2012 Final

2012-08 **Loukas Balafoutas, Glenn Dutcher, Florian Lindner, Dmitry Ryvkin:** The optimal allocation of prizes in tournaments of heterogeneous agents

2012-07 **Stefan Lang, Nikolaus Umlauf, Peter Wechselberger, Kenneth Harttgen, Thomas Kneib:** Multilevel structured additive regression

2012-06 **Elisabeth Waldmann, Thomas Kneib, Yu Ryan Yu, Stefan Lang:** Bayesian semiparametric additive quantile regression

2012-05 **Eric Mayer, Sebastian Rueth, Johann Scharler:** Government debt, inflation dynamics and the transmission of fiscal policy shocks

2012-04 **Markus Leibrecht, Johann Scharler:** Government size and business cycle volatility; How important are credit constraints?

2012-03 **Uwe Dulleck, David Johnston, Rudolf Kerschbamer, Matthias Sutter:** The good, the bad and the naive: Do fair prices signal good types or do they induce good behaviour?

2012-02 **Martin G. Kocher, Wolfgang J. Luhan, Matthias Sutter:** Testing a forgotten aspect of Akerlof's gift exchange hypothesis: Relational contracts with individual and uniform wages

2012-01 **Loukas Balafoutas, Florian Lindner, Matthias Sutter:** Sabotage in tournaments: Evidence from a natural experiment *published in Kyklos*

University of Innsbruck

Working Papers in Economics and Statistics

2013-12

Nadja Klein, Thomas Kneib, Stefan Lang

Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

Abstract
Frequent problems in applied research that prevent 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 special instances, we consider the zero-inflated Poisson, the negative binomial and the zero-inflated negative binomial distribution as standard options for applied work. 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 to quantify the predictive ability of the models. The deviance information criterion is used for further model specification.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)