Zentrum Mathematik

Lehrstuhl für Mathematische Statistik der
Technischen Universität München

Hierarchical Bayesian spatial regression models with applications to non-life insurance

Susanne Gschlößl

Zentrum Mathematik
Lehrstuhl für Mathematische Statistik
der Technischen Universität München

Hierarchical Bayesian spatial regression models with applications to non-life insurance

Susanne Gschlößl

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Herbert Spohn
Prüfer der Dissertation:	1.	Univ.-Prof. Claudia Czado, Ph.D.
	2.	Prof. Arnoldo Frigessi, University of Oslo / Norwegen

Die Dissertation wurde am 30.11.2005 bei der Technischen Universität eingereicht und durch die Fakultät für Mathematik am 27.01.2006 angenommen.

Abstract

In this thesis the modelling of overdispersed spatial count regression data is addressed. Particular emphasis is given to the Generalized Poisson distribution and zero inflated models. Further, the incorporation of spatial random effects which allows for the modelling of an underlying spatial dependency pattern, forms a central part. For the considered models efficient Markov Chain Monte Carlo (MCMC) algorithms are developed and implemented. In particular a novel Gibbs sampler for spatial Poisson regression models is developed using data augmentation techniques and compared to existing methods. An application to a comprehensive data set from a German car insurance company is given. Spatial regression models for the number of claims and claim size are developed. In contrast to the classical compound Poisson model we allow for dependencies between claim frequency and claim size. Based on these models the total claim sizes are simulated which are fundamental for premium calculation in insurance.

Zusammenfassung

In dieser Arbeit werden räumliche Regressionsmodelle für Zähldaten mit Überdispersion betrachtet. Das Hauptaugenmerk liegt hier bei der Generalisierten Poissonverteilung und Modellen, die einen Nullenüberschuss erlauben. Die Einführung von zufälligen räumlichen Effekten, durch welche überdies räumliche Abhängigkeiten in den Daten modelliert werden, ist von zentraler Bedeutung. Für die betrachteten Modelle werden effiziente "Markov Chain Monte Carlo" (MCMC) Algorithmen entwickelt und implementiert. Insbesondere wird ein neuartiger Gibbs Sampler für räumliche Poisson Regressionsmodelle vorgestellt und mit existierenden Methoden verglichen. Die erarbeiteten Modelle und Algorithmen werden zur Analyse eines umfassenden Datensatzes einer deutschen KFZ-Versicherung verwendet. Hier werden zum ersten Mal räumliche Regressionsmodelle entwickelt, welche Abhängigkeiten zwischen der Anzahl der Schäden und der Schadenshöhe zulassen. Basierend auf diesen Modellen werden die Gesamtschäden der Versicherungsnehmer simuliert, welche die Grundlage zur Prämienkalkulation in der Versicherung bilden.

Acknowledgements

First of all, I would like to express my gratitude to Prof. Claudia Czado for the excellent, intensive supervision during the last three years. This thesis has gained a lot from many fruitful discussions and valuables ideas proposed by her. I am also very grateful for her encouragement and support of research stays abroard and the participation on scientific conferences.

Further I would like to thank Dr. Carmen Fernández for the very good supervision on my thesis during my six month stay at the Statistics Department at Lancaster University. I have learned a lot from her wide experience in MCMC methods and gained many ideas which influenced this thesis.

I would also like to thank Prof. Arnoldo Frigessi for inviting me to a research stay in Oslo and his helpful and encouraging suggestions on parts of my thesis during this time.

Further my thanks go to Prof. Sylvia Frühwirth-Schnatter and Prof. Tilman Gneiting for fruitful discussions and many helpful comments and suggestions on my work.

Last but not least I would like to thank my colleagues at the Munich University of Technology, as well as the statistical department at Lancaster University for a pleasant time during the last years. Financial support by the Deutsche Forschungsgemeinschaft through the graduate program "Angewandte Algorithmische Mathematik" is gratefully acknowledged.

Contents

Chapter 1

Introduction

Count data are data taking only categorical values and arise in many statistical applications. In insurance the number of claims caused by a policyholder is an important quantity to be modelled, in epidemiology the number of cases with a certain disease is of interest. The most popular model for count data is the Poisson distribution where the probability for k counts is given by

$$P(Y = k) = \exp(-\mu)\frac{\mu^k}{k!}$$

with μ denoting the intensity of the Poisson distribution. When the data contain covariate information – for example age and gender of the policyholders, the type of car etc. in car insurance – a regression can be performed on the mean $E(Y_i) = \mu_i$ of the i-th observation $(i = 1, .., n)$, i.e. the mean can be modelled in terms of covariates $\mathbf{x}_i$ and unknown regression parameters $\boldsymbol{\beta}$. Usually a logarithmic link function is assumed resulting in

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}).$$

However, the Poisson distribution is rather restrictive in the sense that equality of the variance and the mean is assumed. Frequently, count data display a variance which is considerably higher than the mean, a phenomenon which is called *overdispersion*. In this case the Poisson assumption is violated and the use of the Poisson model will not be appropriate for analysing the data.

Therefore more flexible models which relax the Poisson assumption have been widely discussed in the literature, see Winkelmann (2003) for an overview. In many cases overdispersion is caused due to unobserved heterogeneity in the data. A frequently used model for overdispersed data is the negative binomial distribution which arises as a mixture distribution from a Poisson distribution where the parameter μ is assumed to be random and

to follow a Gamma distribution. By this the negative binomial distribution allows for unobserved heterogeneity among subjects. The generalized Poisson distribution introduced by Consul and Jain (1973) provides another possibility for modelling overdispersion. Here a second parameter is introduced which allows an independent modelling of the variance and the mean. However, in contrast to the negative binomial distribution, the generalized Poisson distribution seems to be much less known in the statistical community. For this reason, one aim of this thesis is to investigate the generalized Poisson distribution in more detail, to apply it on real data and to give a comparison to the negative binomial distribution.

Overdispersion might also be caused by an extraordinary large amount of zero counts in the data. Consider for example the number of traffic accidents in car insurance. Here, typically very few claims are observed, for most of the policyholders we expect to observe no claim at all. For data of this type zero inflated models may be used, see again Winkelmann (2003). Additionally to the zeros arising from the count data model, in zero inflated models part of the zero observations are assumed to be extra, strategic zeros. In car insurance this might be the case for policyholders which do not report minor claims in order to keep their no-claims bonus. Zero inflated models can be used in combination with any count data distribution, in this thesis in particular the zero inflated Poisson and the zero inflated generalized Poisson model are considered.

As mentioned already above, overdispersed data display an extra variability which can be interpreted as unobserved heterogeneity which is not satisfactorily explained by the incorporated covariates. Therefore, another approach for taking unobserved heterogeneity into account consists in the introduction of observation specific random effects $\boldsymbol{\gamma} = (\gamma_1, .., \gamma_n)$. The mean is then specified as

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_i).$$

Further, when the data are spatially indexed - for example for policyholders of a German insurance company the district they are living in is known- *spatial random effects*, i.e. random effects associated with geographic areas rather than individuals, might be assumed. For the estimation of these spatial effects strength can be borrowed from neighbouring regions by assuming that the effect in regions lying close together is similar. In a Bayesian approach this is done by assuming a prior distribution for the spatial effects which takes the neighbourhood structure of the regions into account and allows for dependencies between regions. The most popular model for spatial effects is probably the intrinsic conditional autoregressive (CAR) model introduced by Besag and Kooperberg

(1995). However, the intrinsic CAR model is not proper and care must be taken in order to achieve propriety of the posterior distribution. Several modifications of the intrinsic CAR model leading to a proper joint distribution of the spatial effects have been proposed in the literature, in this thesis a proper model based on Pettitt, Weir, and Hart (2002) will be used.

As already indicated, models are considered in a Bayesian context in this thesis. This allows for parameter uncertainty by assuming the parameters to be random and in particular enables the modelling of a spatial dependency pattern by imposing a CAR prior on the spatial effects. The resulting posterior distributions will be high dimensional, complex functions of the unknown parameters which in general are not tractable analytically anymore. Therefore, Markov Chain Monte Carlo (MCMC) methods will be used for parameter estimation. While in general MCMC samplers are easily implemented using Gibbs and Metropolis Hastings algorithms, convergence and mixing of the samplers crucially depends on issues like model parameterisation, choice of the proposal densities and update schemes for the parameters, see for example Roberts and Sahu (1997) and Gelfand et al. (1995). Therefore, the development and implementation of efficient MCMC algorithms for all considered models forms an important part of this thesis as well. The MCMC samplers have been implemented in Matlab, various techniques for improving convergence and mixing have been applied.

In particular, a new methodology for spatial Poisson regression models is presented. It is shown, that a data augmentation scheme developed in Frühwirth-Schnatter and Wagner (2004a) can be extended to spatial Poisson regression models leading to a straightforward Gibbs sampler. Usually, Metropolis Hastings algorithms have to be used for this kind of models which require the choice of adequate, often computationally costly proposal distributions. Using data augmentation however, the model can be transformed into an approximate normal linear model allowing for a Gibbs sampler, i.e. direct samples from the full conditionals are possible.

The work in this thesis has been mainly motivated by a very large and comprehensive data set from a German car insurance company. The analysis of these data using the discussed models is another main issue of this thesis.

Here, not only the modelling of the number of claims of the policyholders is of interest. In order to obtain a basis for premium calculation predictions of the total claim sizes are needed. Since the total claim size is determined both by the number of claims and the average or individual claim sizes, respectively, typically a separate analysis of the number of claims and claim size is performed. In the classical compound Poisson model going

back to Lundberg (1903) independence of claim frequency and claim size is assumed. One important contribution of this thesis is that we relax this assumption and allow for dependencies between these two quantities. In particular, claim size is modelled conditionally on the observed number of claims which allows us to incorporate the number of claims as covariate in the model for claim size. Similar to the count data models, covariates and spatial effects are incorporated in the models for claim size as well. In particular, spatial Gamma regression models will be assumed.

An outline of the thesis is given in the following. In Chapter 2 the basics in Bayesian inference and MCMC methods are briefly summarized. The two main MCMC algorithms, the Gibbs sampler and the Metropolis Hastings sampler are introduced and several choices for proposal distributions are discussed.
Bayesian Model Choice including assessment of the model fit and model comparison is addressed in Chapter 3. Next to the well known deviance information criterion (DIC), the predictive model choice criterion and several proper scoring rules for categorical and continuous variables are presented.
In Chapter 4 the modelling of spatial effects is addressed. We introduce a proper Gaussian conditional autoregressive prior based on Pettitt et al. (2002) which allows the modelling of a spatial dependency structure and can be efficiently handled in a MCMC setting.
In Chapter 5, which is closely based on Gschlößl and Czado (2005b), various models for count data are presented. Besides the Poisson model, we additionally consider models allowing for overdispersion and an excessive number of zero observations, in particular the negative binomial, the generalized Poisson and zero inflated models. For these models a regression setup including spatial random effects is developed, prior specifications and details on the MCMC algorithms for parameter estimation are provided. The models are investigated in an application in epidemiology: the number of invasive meningococcal disease cases in Germany, reported in the year 2004, is analysed. Models are compared using the criteria presented in Chapter 3. We observe a rather high degree of overdispersion in the data which is captured best by the generalized Poisson model when spatial effects are neglected. However, when spatial effects are added to the models, a spatial Poisson model provides the best fit to the data.
A new MCMC methodology for spatial Poisson regression models is presented and evaluated in Chapter 6, following closely Gschlößl and Czado (2005a). Using data augmentation we show, that by the introduction of two sequences of latent variables a straightforward Gibbs sampler is available for a Poisson regression model including spatial effects. In par-

ticular, the influence of model parameterisation and different update strategies on the mixing of the MCMC chains is discussed. The developed Gibbs samplers are analysed in two simulation studies and applied to model the expected number of claims for policyholders of a German car insurance company. The mixing of the Gibbs samplers depends crucially on the model parameterisation and the update schemes. The best mixing is achieved when collapsed algorithms are used, reasonable low autocorrelations for the spatial effects are obtained in this case. For the regression effects however, autocorrelations are rather high, especially for data with very low heterogeneity. For comparison a single component Metropolis Hastings algorithms is applied which displays very good mixing for all components. Although the Metropolis Hastings sampler requires a higher computational effort, it outperforms the Gibbs samplers which would have to be run considerably longer in order to obtain the same precision of the parameters.

Chapters 7 and 8 are based on Gschlößl and Czado (2005c). In Chapter 7 spatial regression models for claim size are presented. Here, two approaches are considered. Both models for the individual claim sizes as well as models for the average claim size per policyholder are assumed. We consider models conditionally on the number of claims, which allows us to model claim size in dependence of the observed number of claims. Based on the models for claim frequency and claim size, finally the posterior predictive distribution of the total claim sizes can be approximated. For this independence of the number of claims and claim size is not required.

An application to car insurance data for policyholders in Germany within one year is presented in Chapter 8. For more than 350000 policyholders the data contain the number of claims, the corresponding claim sizes as well as several covariates like age, gender, type of car, kilometers driven per year and no-claims bonus. The number of claims is modelled using the regression models presented in Chapter 5. Besides a number of covariates, spatial random effects associated with the 440 districts in Germany are taken into account, allowing for spatial dependencies. For these data a spatial Poisson model turns out to be sufficient, an extension to models allowing for overdispersion like the generalized Poisson distribution or zero inflated models gives no improvement. For claim size the two approaches discussed in Chapter 7 are taken. Again covariates and spatial random effects are incorporated. Both the models for the number of claims and claim size are improved by the inclusion of spatial effects, in particular a smooth spatial pattern is observed. While the expected number of claims decreases from the south western parts of Germany to the east, a contrary trend is recognized for the average and individual claim sizes. Further, we quantify significant number of claims effects on claim size. With an increasing number of

claims, the average and individual claim sizes tend to decrease. Based on the MCMC output of the models for claim frequency and claim size the posterior predictive distribution of the total claim sizes is approximated. Also posterior prediction of the total claim size is improved by allowing for spatial effects, the impact of number of claims effects however diminishes.

Chapter 2

Basics on MCMC

In this section the basics of Bayesian inference and Markov Chain Monte Carlo (MCMC) simulation are briefly summarized. For more information see Gilks et al. (1996b) and Gelman et al. (2004). A good overview about MCMC methods is given by Dellaportas and Roberts (2003). Note, that the terms 'density' and 'distribution' are used interchangeably throughout this thesis.
Assume we have a statistical model $p(\mathbf{y}|\boldsymbol{\theta})$ for the vector of observed data $\mathbf{y}$ depending on a vector of unknown parameters $\boldsymbol{\theta}$. In a Bayesian context the parameter $\boldsymbol{\theta}$ is assumed to be random with prior distribution $\pi(\boldsymbol{\theta})$. Using Bayes theorem the posterior distribution of $\boldsymbol{\theta}$ is given by

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) &= \frac{p(\mathbf{y}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \\ &\propto p(\mathbf{y}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}). \end{aligned}$$

However, in general the posterior distribution for complex statistical models is a high dimensional, not analytically and numerically tractable function. MCMC provides a method to generate approximate samples from the posterior distribution which can be used to approximate quantities like the posterior mean, mode etc. by its empirical counterpart. The two basic algorithms used in MCMC are the Gibbs sampler first introduced by Geman and Geman (1984) and discussed by Gelfand and Smith (1990) and the Metropolis Hastings sampler developed by Metropolis et al. (1953) and Hastings (1970). These algorithms are described in the following sections.

2.1 Gibbs Sampler

Suppose we want to sample from a posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ where the parameter $\boldsymbol{\theta}$ is divided into d components $\boldsymbol{\theta} = (\theta_1, .., \theta_d)'$. The Gibbs sampler is based on the full conditional distributions denoted by $p(\theta_i|\boldsymbol{\theta}_{-i}) := p(\theta_i|\boldsymbol{\theta}_i, \mathbf{y})$, where $\boldsymbol{\theta}_i = (\theta_1, .., \theta_{i-1}, \theta_{i+1}, .., \theta_d)'$ denotes the vector of $\boldsymbol{\theta}$ without the i-th component. The algorithm of the Gibbs sampler proceeds as follows:

1. choose a starting value $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, .., \theta_d^{(0)})'$ and set $t = 1$
2. sample $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, .., \theta_d^{(t)})'$ by

$$\begin{aligned}
\theta_1^{(t)} &\sim p\Big(\theta_1|\theta_2^{(t-1)}, .., \theta_d^{(t-1)}\Big) \\
\theta_2^{(t)} &\sim p\Big(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, .., \theta_d^{(t-1)}\Big) \\
&\vdots \\
\theta_d^{(t)} &\sim p\Big(\theta_d|\theta_1^{(t)}, .., \theta_{d-1}^{(t)}\Big)
\end{aligned}$$

3. increase t by 1 and return to step 2.

2.2 Metropolis Hastings (MH) Sampler

The Gibbs sampler successively samples from the full conditional distributions. However, if the full conditionals do not belong to any standard distribution, samples from them can no longer be obtained directly and sampling might be very tedious. In this case, the Metropolis Hastings (MH) algorithm can be used. Instead of sampling directly from the full conditional of θ, a candidate value $\tilde{\theta}$ from an arbitrary proposal distribution $q(\tilde{\theta}, \cdot)$ is drawn and accepted with a certain probability. Like the Gibbs sampler, the MH sampler updates the parameters component by component. However, for notational simplicity the MH algorithm is described for one single component in this section. The algorithm works as follows:

1. choose a starting value $\theta^{(0)}$ and set $t = 1$
2. – propose a candidate value $\tilde{\theta} \sim q(\tilde{\theta}, \cdot)$
   – accept $\tilde{\theta}$ with probability $\min\Big\{1, \dfrac{p(\tilde{\theta}|\mathbf{y})}{p(\theta^{(t-1)}|\mathbf{y})} \dfrac{q(\theta^{(t-1)}, \tilde{\theta})}{q(\tilde{\theta}, \theta^{(t-1)})}\Big\}$ and set $\theta^{(t)} = \tilde{\theta}$, otherwise set $\theta^{(t)} = \theta^{(t-1)}$

3. increase t by 1 and return to step 2.

Efficiency of the MH sampler depends crucially on the choice of proposal distribution. In the following sections some proposal distributions are described, in particular we consider a symmetric random walk proposal, an independence proposal and Gamerman's proposal distribution.

2.2.1 Symmetric random walk proposal distribution

A symmetric random walk proposal density is a symmetric density of the form

$$q(\tilde{\theta}, \theta) = q(|\tilde{\theta} - \theta|).$$

In this case, the acceptance probability simplifies to $\min\left\{1, \frac{p(\tilde{\theta}|\mathbf{y})}{p(\theta|\mathbf{y})}\right\}$ with θ denoting the current value of the parameter. A common choice for a random walk proposal distribution is a normal distribution centered around the current value θ, i.e. $q(\tilde{\theta}, \theta) = \frac{1}{\sqrt{2\pi\tau^2}} \exp[-\frac{1}{2\tau^2}(\tilde{\theta} - \theta)^2]$. When only small moves are generated by the proposal distribution, the acceptance rates will be very high, but the chain will move very slowly and take a long time to explore the whole parameter space. In contrast, when large moves are proposed, the moves will often fall in the tail of the posterior distribution, resulting in very low acceptance rates. Therefore, we tune the variance τ^2 of the normal proposal using pilot runs in order to achieve acceptance rates between 30 and 60 % as recommended by Besag et al. (1995).

2.2.2 Independence proposal distribution

An independence proposal distribution is a proposal which is independent of the current value of the sampled parameters, i.e. $q(\tilde{\theta}, \cdot) = q(\tilde{\theta})$. In order to achieve high acceptance rates, the proposal distribution should be a good approximation to the target distribution $p(\theta|\mathbf{y})$ with slightly heavier tails. If a light tailed proposal distribution is used, the chain might get stuck in the tails of the target distribution resulting in low acceptance rates. A common independence proposal is a normal distribution with the same mode and inverse curvature at the mode as the target distribution $p(\theta|\mathbf{y})$, see for example Gilks et al. (1996a). Here, the mode has to be determined in every iteration using a numerical optimization routine, for example the Newton-Raphson algorithm or the bisection method. The algorithm for this independence MH sampler is given by

1. choose a starting value $\theta^{(0)}$ and set $t = 1$

2. – calculate the mode θ_{mode} and the inverse curvature at the mode $-H(\theta_{mode})^{-1}$ of the target distribution, where $H(\theta) := \frac{\delta^2 p(\theta|\mathbf{y})}{\delta\theta^2}$
   – propose a candidate value $\tilde{\theta} \sim N(\theta_{mode}, -H(\theta_{mode})^{-1})$
   – accept $\tilde{\theta}$ with probability $\min\left\{1, \frac{p(\tilde{\theta}|\mathbf{y})}{p(\theta^{(t-1)}|\mathbf{y})}\frac{q(\theta^{(t-1)})}{q(\tilde{\theta})}\right\}$ and set $\theta^{(t)} = \tilde{\theta}$, otherwise set $\theta^{(t)} = \theta^{(t-1)}$

3. increase t by 1 and return to step 2.

Alternatively to the normal distribution a t-distribution with v degrees of freedom can be used. For large v this is approximately a normal distribution, whereas for small values of v a distribution with thicker tails is obtained. In particular, $v = 1$ corresponds to the Cauchy distribution.
The density of the t-distribution with v degrees of freedom is given by

$$f_v(t) = \frac{\Gamma((v+1)/2)}{\sqrt{v\pi}\Gamma(v/2)(1+t^2v^{-1})^{(v+1)/2}} \propto \frac{1}{(1+t^2v^{-1})^{(v+1)/2}}.$$

The density for the transformation $y = \mu + \sigma \cdot z$ where $z \sim t(v)$, i.e. y is t-distributed with mean μ and variance $\frac{v}{v-2}\sigma^2$, is then proportional to

$$f_v(y) \propto \frac{1}{\sigma}\frac{1}{(1+\left(\frac{y-\mu}{\sigma}\right)^2 v^{-1})^{(v+1)/2}}. \tag{2.1}$$

In order to determine the mode and the inverse curvature at the mode of (2.1), we calculate the first and the second derivative of the logarithm of $f_v(y)$

$$\log f_v(y) \propto -\log\sigma - \frac{v+1}{2}\log(1+\left(\frac{y-\mu}{\sigma}\right)^2 v^{-1}).$$

$$(\log f_v(y))' = -\frac{v+1}{v\sigma}\frac{\left(\frac{y-\mu}{\sigma}\right)}{1+\left(\frac{y-\mu}{\sigma}\right)^2 v^{-1}}$$

$$(\log f_v(y))'' = -\frac{v+1}{v}\frac{1}{\sigma^2}\frac{1-\left(\frac{y-\mu}{\sigma}\right)^2\frac{1}{v}}{[1+\left(\frac{y-\mu}{\sigma}\right)^2\frac{1}{v}]^2}$$

Since $(\log f_v(y))' = 0$ for $y = \mu$, the mode is given by $y = \mu$ and since

$$(\log f_v(\mu))'' = -\frac{v+1}{v}\frac{1}{\sigma^2}$$

the inverse curvature at the mode is given by $\frac{v}{v+1}\sigma^2$. Equating the mode and the inverse curvature at the mode of both the target and the t-distribution, we finally obtain

$$\theta_{mode} = \mu \quad \text{and} \quad -H(\theta_{mode})^{-1} = \frac{v}{v+1}\sigma^2$$

and therefore $\sigma = \left(-\frac{v+1}{v}H(\theta_{mode})^{-1}\right)^{\frac{1}{2}}$. For the MH independence sampler using the t-distribution the sampling procedure is the following:

1. choose a starting value $\theta^{(0)}$ and set $t = 1$
2. – calculate the mode θ_{mode} and the inverse curvature at the mode $-H(\theta_{mode})^{-1}$ of the target distribution
   – propose a candidate value $\tilde{\theta} \sim q(\tilde{\theta}) = \frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sqrt{v\pi}}\frac{1}{\sigma}\left(1+\left(\frac{\tilde{\theta}-\theta_{mode}}{\sigma}\right)^2 v^{-1}\right)^{-(v+1)/2}$ where $\sigma = \left(-\frac{v+1}{v}H(\theta_{mode})^{-1}\right)^{\frac{1}{2}}$
   – accept $\tilde{\theta}$ with probability $\min\left\{1, \frac{p(\tilde{\theta}|\mathbf{y})}{p(\theta^{(t-1)}|\mathbf{y})}\frac{q(\theta^{(t-1)})}{q(\tilde{\theta})}\right\}$ and set $\theta^{(t)} = \tilde{\theta}$, otherwise set $\theta^{(t)} = \theta^{(t-1)}$
3. increase t by 1 and return to step 2.

Throughout this thesis, independence proposals based on the t-distribution with $v = 20$ degrees of freedom will be used.

2.2.3 Gamerman's proposal distribution

In a GLM setting, Gamerman (1997) proposes a MH algorithm based on the iterative weighted least squares (IWLS) algorithm. Like the independence proposal described in the previous section, Gamerman's proposal distribution takes the structure of the model into account by both incorporating likelihood and prior of the model.
Assume, that for independent observations $y_i, i = 1, .., n$ we have a model from the exponential family, i.e. the likelihood can be written in the form

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right)c(y_i, \phi_i), i = 1, .., n,$$

where $b(\cdot)$ and $c(\cdot)$ are known functions depending on the model under consideration and the scale parameters ϕ_i are assumed to be known. The canonical parameter θ_i is related to the mean via $\mu_i = b'(\theta_i)$ and to the regression parameters $\boldsymbol{\alpha} = (\alpha_1, .., \alpha_p)'$ via the link function $g(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\alpha}$. The vector $\mathbf{x}_i = (x_{i1}, .., x_{ip})'$ denotes the vector of covariates for observation i, the design matrix containing the covariate vectors for all observations is denoted by $\mathbf{x} = (\mathbf{x}_1, .., \mathbf{x}_n)$. By this setting, a Generalized Linear Model (GLM) is defined. The maximum likelihood estimate $\hat{\boldsymbol{\alpha}}$ in a GLM is obtained using the IWLS algorithm to the vector of transformed variables $\tilde{\mathbf{y}}(\boldsymbol{\alpha})$ with components $\tilde{y_i}(\boldsymbol{\alpha}) := \mathbf{x}_i'\boldsymbol{\alpha} + (y_i - \mu_i)g'(\mu_i), i = 1, .., n$, see McCullagh and Nelder (1989). The IWLS algorithm starts from an arbitrary value $\boldsymbol{\alpha} = \mathbf{m^{(0)}}$ and then iteratively obtains $\mathbf{m}^{(t)}, t = 1, ..$ as the least squares estimator of the weighted linear model $\tilde{\mathbf{y}}(\mathbf{m}^{(t-1)}) \sim N_p(\mathbf{x}'\boldsymbol{\alpha}, W^{-1}(\mathbf{m}^{(t-1)}))$, where W is a diagonal matrix of weights $W(\boldsymbol{\alpha}) = diag(w_{ii})_{i=1,..,n}$ with $w_{ii}^{-1} = b''(\theta_i)(g'(\mu_i))^2$. That is, $\hat{\boldsymbol{\alpha}}$ is obtained after convergence of the iteration $\mathbf{m}^{(t)} = (\mathbf{x}'W(\mathbf{m}^{(t-1)})\mathbf{x})^{-1}\mathbf{x}'W(\mathbf{m}^{(t-1)})\tilde{\mathbf{y}}(\mathbf{m}^{(t-1)})$ with associated asymptotic covariance matrix $C^{(t)} = (\mathbf{x}'W(\mathbf{m}^{(t-1)})\mathbf{x})^{-1}$.
In a Bayesian context we additionally assume a prior for $\boldsymbol{\alpha}$, in particular we consider the normal prior $\boldsymbol{\alpha} \sim N_p(\boldsymbol{\mu_0}, V_0)$. According to the results from the Bayesian version of the IWLS algorithm for a canonical link function, see West (1985), again the posterior mode $\hat{\boldsymbol{\alpha}}$ and the approximate posterior covariance matrix C of $\boldsymbol{\alpha}$ can be obtained from iterative schemes which now additionally incorporate the prior information, in particular $\mathbf{m}^{(t)} = (V_0^{-1} + \mathbf{x}'W(\mathbf{m}^{(t-1)})\mathbf{x})^{-1}(V_0^{-1}\boldsymbol{\mu_0} + \mathbf{x}'W(\mathbf{m}^{(t-1)})\tilde{\mathbf{y}}(\mathbf{m}^{(t-1)}))$ and $C^{(t)} = (V_0^{-1} + \mathbf{x}'W(\mathbf{m}^{(t-1)})\mathbf{x})^{-1}$. The posterior distribution of $\boldsymbol{\alpha}$ is then asymptotically normal distributed with

$$\boldsymbol{\alpha}|\mathbf{y} \approx N_p(\hat{\boldsymbol{\alpha}}, C),$$

where $C = (V_0^{-1} + \mathbf{x}'W(\hat{\boldsymbol{\alpha}})\mathbf{x})^{-1}$. Gamerman (1997) proposes a MH algorithm with a proposal density based on only one step of this iteration, starting at the current value $\boldsymbol{\alpha}^{(t-1)}$. This algorithm reads as follows:

1. choose a starting value $\boldsymbol{\alpha}^{(0)}$ and set $t = 1$

2. – propose a candidate value $\tilde{\boldsymbol{\alpha}}$ with
   $\tilde{\boldsymbol{\alpha}} \sim N\Big((V_0^{-1} + \mathbf{x}'W(\boldsymbol{\alpha}^{(t-1)})\mathbf{x})^{-1}(V_0^{-1}\boldsymbol{\mu_0} + \mathbf{x}'W(\boldsymbol{\alpha}^{(t-1)})\tilde{\mathbf{y}}(\boldsymbol{\alpha}^{(t-1)}),$
   $(V_0^{-1} + \mathbf{x}'W(\boldsymbol{\alpha}^{(t-1)})\mathbf{x})^{-1}\Big)$
   – accept $\tilde{\boldsymbol{\alpha}}$ with probability $\min\Big\{1, \frac{p(\tilde{\boldsymbol{\alpha}}|\mathbf{y})}{p(\boldsymbol{\alpha}^{(t-1)}|\mathbf{y})}\frac{q(\boldsymbol{\alpha}^{(t-1)}|\tilde{\boldsymbol{\alpha}})}{q(\tilde{\boldsymbol{\alpha}}|\boldsymbol{\alpha}^{(t-1)})}\Big\}$ and set $\boldsymbol{\alpha}^{(t)} = \tilde{\boldsymbol{\alpha}}$,

otherwise set $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)}$

3. increase t by 1 and return to step 2.

Note, that the above algorithm is given in a very general notation. Often it is more convenient not to update the whole vector $\boldsymbol{\alpha}$ in one block, but to divide it in several components, say for example $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2)'$ and corresponding block structure for the design matrix $\mathbf{x}' = (\mathbf{x}^1, \mathbf{x}^2)$. In this case, for the update of $\boldsymbol{\alpha}^1$ the transformed variables change to $\tilde{y}_i(\boldsymbol{\alpha}^1) = \mathbf{x}_i'\boldsymbol{\alpha} + (y_i - \mu_i)g'(\mu_i) - \mathbf{x}_i^{2\prime}\boldsymbol{\alpha}^2$, i.e. $\mathbf{x}_i^{2\prime}\boldsymbol{\alpha}^2$ is treated as an offset. Similarly, for the update of $\boldsymbol{\alpha}^2$, $\mathbf{x}_i^{1\prime}\boldsymbol{\alpha}^1$ appears as an offset. For details see Gamerman (1997).
Note, that in comparison to the independence proposal discussed in the previous section where the mode of the target is computed using several iterations of a numerical optimization routine, for Gamerman's proposal only one step of the iteration is performed and therefore less computation time is required. However, Gamerman's proposal can only be applied in a straightforward manner to models from the exponential family.

2.3 Further algorithms and improving MCMC

In case that not all full conditionals can be easily sampled from directly, a popular mixture of the Gibbs and the MH sampler is the Metropolis-within-Gibbs sampler, see for example Dellaportas and Roberts (2003). Like in the Gibbs sampler, the parameters are updated component by component, however, components with a non standard full conditional are updated using a MH-step instead of a Gibbs step.
For univariate log-concave distributions adaptive rejection sampling (ARS) introduced by Gilks and Wild (1992) enables efficient sampling. ARS is an adaptive method where the distribution function to be sampled from is approximated by an envelope and a squeezing function which are both piecewise exponential, for details see Gilks and Wild (1992).
Many techniques for improving MCMC are discussed in the literature, here we will only briefly mention some techniques used in this thesis later on.
For highly correlated components convergence and mixing of MCMC algorithms can be improved by blocking these components and update them jointly in one step.
Another technique for improving MCMC when high correlations between components are present is using collapsed algorithms, see Liu, Wong, and Kong (1994). Collapsing means, that components are updated according to their marginal distribution, i.e. with the highly correlated components integrated out.

Of particular importance is the parameterisation of the model which can crucially determine mixing and convergence of the MCMC samplers. See Gelfand et al. (1995), Roberts and Sahu (1997) and Papaspiliopoulos et al. (2003) for further information on this topic. Reparameterisation issues will be discussed in more detail in Chapter 6. A very good review and summary of the existing parameterisation techniques is given in Frühwirth-Schnatter (2004), with special emphasis on time series models.

Chapter 3

Model choice

MCMC enables parameter estimation for a wide range of models. Once a model has been fitted, an important issue is model choice. Model choice includes both the assessment of the model fit as well as model comparison. We will discuss posterior predictive p-values for checking the model, while the deviance information criterion (DIC), the predictive model choice criterion (PMCC) and several scoring rules will be considered for comparing models.

3.1 Assessment of model fit using posterior predictive p-values

Gelman et al. (1996) propose a method for model checking using the posterior predictive distribution for a discrepancy measure. Stern and Cressie (2000) use this method for model checking in disease mapping models. Assume a model having likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$ and posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. The posterior predictive density for a replication $\mathbf{y}_{rep} = (y_{rep,1}, .., y_{rep,n})$ of the observed data $\mathbf{y} = (y_1, .., y_n)$ is defined as

$$\begin{aligned} p(\mathbf{y}_{rep}|\mathbf{y}) &:= \int p(\mathbf{y}_{rep}, \boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int p(\mathbf{y}_{rep}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \end{aligned} \tag{3.1}$$

Here independence of $\mathbf{y}_{rep}$ and $\mathbf{y}$ given $\boldsymbol{\theta}$ is assumed. The posterior predictive distribution therefore averages the conditional density of the data over the posterior distribution of the parameters $\boldsymbol{\theta}$. It can be estimated by $\hat{p}(\mathbf{y}_{rep}|\mathbf{y}) := \frac{1}{R}\sum_{j=1}^{R} p(\mathbf{y}|\hat{\boldsymbol{\theta}}^j)$ where $\hat{\boldsymbol{\theta}}^j, j = 1, .., R$ denotes the j-th MCMC iterate of $\boldsymbol{\theta}$ after burnin.

For model checking choose a discrepancy measure $D(\mathbf{y}, \boldsymbol{\theta})$ which may depend on $\boldsymbol{\theta}$, like for example the deviance or the χ^2-statistic given later in this section. Then, the posterior predictive p-value for this discrepancy measure is given by

$$p_b(\mathbf{y}) = P(D(\mathbf{y}_{rep}, \boldsymbol{\theta}) \geq D(\mathbf{y}, \boldsymbol{\theta})|\mathbf{y}).$$

This posterior predictive p-value can be computed as follows using a set of MCMC draws $\hat{\boldsymbol{\theta}}^j, j = 1, .., R$ from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$.

- For each draw $\hat{\boldsymbol{\theta}}^j, j = 1, .., R$ simulate a replicated data set $\mathbf{y}^j_{rep}$ according to its sampling distribution $p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}^j)$
- Compute the discrepancy measures $D(\mathbf{y}^j_{rep}, \hat{\boldsymbol{\theta}}^j)$ and $D(\mathbf{y}, \hat{\boldsymbol{\theta}}^j)$

If the discrepancy measure $D(\mathbf{y}, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$, Gelman et al. (1996) propose to produce a scatterplot of the pairs $(D(\mathbf{y}, \hat{\boldsymbol{\theta}}^j), D(\mathbf{y}^j_{rep}, \hat{\boldsymbol{\theta}}^j)), j = 1, .., R$. If the model fit is good, the observed data should be similar to the simulated data, i.e. the discrepancy measures evaluated for the observed data should not be outliers compared to the discrepancy measures evaluated for the simulated data. Therefore, for a good model fit about half of the points should fall below and half above the 45 degree line. The p-value can then be estimated by the proportion of pairs for which $D(\mathbf{y}^j_{rep}, \hat{\boldsymbol{\theta}}^j) \geq D(\mathbf{y}, \hat{\boldsymbol{\theta}}^j)$, i.e.

$$\hat{p}_b = \frac{1}{R} \sum_{j=1}^{R} I(D(\mathbf{y}^j_{rep}, \hat{\boldsymbol{\theta}}^j) \geq D(\mathbf{y}, \hat{\boldsymbol{\theta}}^j)),$$

where $I(\cdot)$ denotes the indicator function. An extreme p-value close to 0 or 1 indicates a lack of fit of the model according to the chosen discrepancy measure. If $D(\mathbf{y}, \boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$ a histogram of $D(\mathbf{y}^j_{rep})$ can be displayed and compared to $D(\mathbf{y})$.
Possible discrepancy measures include the

- non standardized deviance: $D(\mathbf{y}, \boldsymbol{\theta}) = Dev(\mathbf{y}, \boldsymbol{\theta}) = -2 \log p(\mathbf{y}|\boldsymbol{\theta})$
- χ^2-discrepancy: $D(\mathbf{y}, \boldsymbol{\theta}) = \chi^2(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - E(Y_i|\boldsymbol{\theta}))^2}{Var(Y_i|\boldsymbol{\theta})}$
- $D(\mathbf{y}) = DIC(\mathbf{y})$ which will be introduced in Section 3.2.1.

Only recently Hjort et al. (2005) have pointed out limitations of posterior predictive p-values. They state, that nothing can be learned from medium p-values and propose to base model evaluation on calibrated posterior predictive p-values instead. However, this would require an immense computational effort for the models considered in this thesis.

3.2 Model comparison

Bayes factors based on marginal likelihood provide a method for model comparison, see Kass and Raftery (1995). Further, Bayesian Model Averaging (BMA), see for example Hoeting et al. (1999), which is based on Bayes factors presents a method for model selection taking model uncertainty into account. However, for complex hierarchical models like those which will be considered in this thesis, the computation of Bayes factors requires substantial efforts , see Han and Carlin (2001). Therefore, we use model choice criterions and scoring rules which can be easily computed using the available MCMC output in this thesis.

3.2.1 Deviance Information Criterion (DIC)

Spiegelhalter et al. (2002) suggest to use the following criterion for model comparison in Bayesian inference. Assume a probability model $p(\mathbf{y}|\boldsymbol{\theta})$. The Bayesian deviance $D(\boldsymbol{\theta})$, which is used as a measure for goodness of fit, is defined as

$$D(\boldsymbol{\theta}) = -2\log p(\mathbf{y}|\boldsymbol{\theta}) + 2\log f(\mathbf{y})$$

where $f(\mathbf{y})$ is some fully specified standardizing term. To measure the model complexity Spiegelhalter et al. (2002) introduce the effective number of parameters p_D defined by

$$\begin{aligned} p_D &:= E[D(\boldsymbol{\theta}|\mathbf{y})] - D(E[\boldsymbol{\theta}|\mathbf{y}]) \\ &= \text{posterior mean of the deviance} - \text{deviance of the posterior means.} \end{aligned}$$

Finally they define the deviance information criterion (DIC) as the sum of the posterior mean of the deviance and the effective number of parameters

$$DIC := E[D(\boldsymbol{\theta}|\mathbf{y})] + p_D.$$

According to this criterion the model with the smallest DIC is to be preferred. p_D and DIC are easily computed using the available MCMC output by taking the posterior mean of the deviance to obtain $E[D(\boldsymbol{\theta}|\mathbf{y})]$ and the plug-in estimate of the deviance $D(E[\boldsymbol{\theta}|\mathbf{y}])$ using the posterior means $E[\boldsymbol{\theta}|\mathbf{y}]$ of the parameter $\boldsymbol{\theta}$.
The DIC may be also used for comparing non nested models. However, in this case, the full normalizing constants of $p(\mathbf{y}|\boldsymbol{\theta})$ have to be taken into account when computing the deviance $D(\boldsymbol{\theta})$. An information theoretic discussion of the DIC as criterion for posterior predictive model comparison is given in van der Linde (2005).

3.2.2 Predictive model choice criterion (PMCC)

A related approach for model comparison is given by the predictive model choice criterion (PMCC) considered by Laud and Ibrahim (1995) and Gelfand and Ghosh (1998). It is based on the posterior predictive distribution $p(\mathbf{y}_{rep}|\mathbf{y})$ (see (3.1)). The PMCC is defined by

$$PMCC := \sum_{i=1}^{n}(\mu_i - y_i)^2 + \sum_{i=1}^{n}\sigma_i^2, \tag{3.2}$$

where $\mu_i := E(y_{rep,i}|\mathbf{y})$ and $\sigma_i^2 := Var(y_{rep,i}|\mathbf{y})$ denote the expected value and the variance of a replicate $y_{rep,i}$ of the posterior predictive distribution. Similar to the DIC, models with a smaller value of the PMCC are preferred. While the first term $\sum_{i=1}^{n}(\mu_i - y_i)^2$ gives a goodness-of-fit measure which will decrease with increasing model complexity, the second term $\sum_{i=1}^{n}\sigma_i^2$ can be considered as penalty term which will tend to be large both for poor and overfitted models, see Gelfand and Ghosh (1998). The quantities μ_i and σ_i^2 can be estimated based on the MCMC output $\hat{\boldsymbol{\theta}}^j, j = 1, .., R$ by $\hat{\mu}_i := \frac{1}{R}\sum_{j=1}^{R}\mu_i(\hat{\boldsymbol{\theta}}^j)$ and $\hat{\sigma}_i^2 := \frac{1}{R}\sum_{j=1}^{R}\sigma_i^2(\hat{\boldsymbol{\theta}}^j)$, where $\mu_i(\boldsymbol{\theta})$ and $\sigma_i^2(\boldsymbol{\theta})$ denote the mean and the variance of the underlying model $p(\mathbf{y}|\boldsymbol{\theta})$ depending on the parameters $\boldsymbol{\theta}$. When the mean $\mu_i(\boldsymbol{\theta})$ and the variance $\sigma_i^2(\boldsymbol{\theta})$ of the model are not explicitly available, the PMCC can be alternatively evaluated using simulation. For every MCMC iteration $j = 1, .., R$ after burnin, a replicated data set $\mathbf{y}_{rep}^j = (y_{rep,1}^j, .., y_{rep,n}^j)$ can be simulated from $p(\mathbf{y}|\hat{\boldsymbol{\theta}}^j)$. The mean μ_i and the variance σ_i^2 can then be estimated by the empirical counterparts $\hat{\mu}_i := \frac{1}{R}\sum_{j=1}^{R} y_{rep,i}^j$ and $\hat{\sigma}_i^S := \frac{1}{R-1}\sum_{j=1}^{R}(y_{rep}^j - \hat{\mu}_i)^2$.
The PMCC will be used in the application given in Chapter 8 for comparing models for the number of claims and for individual, average and total claim sizes. Since the mean $\mu_i(\boldsymbol{\theta})$ and the variance $\sigma_i^2(\boldsymbol{\theta})$ are explicitly given in the models for the number of claims, the individual and the average claim sizes, we will compute the PMCC directly using the MCMC output of these models. The distribution of the total claim sizes however, is not available in an analytically closed form, therefore here, the PMCC will be evaluated using simulation as described above.

3.2.3 Proper scoring rules

Gneiting and Raftery (2004) consider scoring rules for assessing the quality of probabilistic forecasts. A scoring rule assigns a numerical score based on the forecast of the predictive distribution of a specific model and the value that was observed and can be used for

comparing the predictive distribution of several models. Ideally, both calibration and sharpness of the predictive distribution are taken into account. Calibration refers to the statistical consistency between the observed and the predicted data, whereas sharpness is determined by the concentration of the predictive distribution and is independent of the observed value. Gneiting and Raftery (2004) also use scoring rules in estimation problems for assessing the optimal score estimator for the unknown model parameters. Assume a parametric model $P_\theta := p(\mathbf{y}|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$ based on the sample $\mathbf{y} = (y_1, .., y_n)$. Then, the mean score

$$S_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} S(P_\theta, y_i)$$

can be taken as a goodness-of-fit measure, where S is a strictly proper scoring rule, i.e. the highest score is obtained for the true model. For details on strictly proper and proper scoring rules see Gneiting and Raftery (2004). Since for the true parameter vector $\boldsymbol{\theta}_0$, see Gneiting and Raftery (2004),

$$argmax_\theta S_n(\boldsymbol{\theta}) \rightarrow \boldsymbol{\theta}_0, n \rightarrow \infty,$$

the optimum score estimator based on scoring rule S is given by $\hat{\boldsymbol{\theta}}_n = argmax_\theta S_n(\boldsymbol{\theta})$. We will use scoring rules in a Bayesian context as measures for comparing models based on their posterior predictive distribution. Gneiting and Raftery (2004) provide and discuss several scoring rules, we will present some of the proper scoring rules for categorical and continuous variables here. All considered scores are positively oriented, i.e. the model with the highest mean score $S_n(\boldsymbol{\theta})$ is favoured.

Scoring rules for categorical variables

For categorical variables scores are based on the posterior predictive probability vector $\mathbf{p}_i = (p_{i1}, p_{i2}, .., p_{im})$. Here the component $p_{ij} := P(y_i = j|\mathbf{y})$ denotes the posterior predictive probability that the i-th observation takes the value j which can be estimated by $\hat{p}_{ij} := \frac{1}{R}\sum_{k=1}^{R} p(y_i = j|\boldsymbol{\theta}^k)$ where $\boldsymbol{\theta}^k, k = 1, .., R$, denotes the k-th MCMC iterate of $\boldsymbol{\theta}$ after burnin. For computational reasons we set $p_{im} := 1 - \sum_{k=1}^{m-1} p_{ik}$ where $m - 1$ gives the highest response value observed in the data. This ensures that the probability vector $\mathbf{p}_i$ sums up to 1. We consider the logarithmic score and the Brier score for categorical variables.

The logarithmic score is defined by

$$S(\mathbf{p}_i, y_i) = \log p_{iy_i}$$

where $p_{iy_i} = P(y = y_i|\mathbf{y})$ denotes the posterior predictive probability for the true value y_i under the considered model. Therefore, the logarithmic score chooses the model which gives the highest probability for observing the true value.
The Brier score first proposed by Brier (1950) is defined by

$$S(\mathbf{p}_i, y_i) = 2p_{iy_i} - 1 - \sum_{k=1}^{m} p_{ik}^2$$

The mean Brier score corresponds to the expression

$$-\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}(p_{ij} - \hat{p}_{ij}^{emp})^2$$

where $\hat{p}_{ij}^{emp} = \begin{cases} 1 & y_i = j \\ 0 & \text{otherwise} \end{cases}$ denotes the empirical probability that the i-th observation takes the value j. Hence, according to the Brier score the model which minimizes the squared difference between the observed and the estimated probabilities is considered best.
Both scoring rules are strictly proper, see Gneiting and Raftery (2004). Further, when parameter estimation is done using MCMC both scores are computed easily based on the available MCMC output as indicated above.

Scoring rules for continuous variables

For continuous variables we consider the logarithmic score (LS), the continuous ranked probability score (CRPS), the interval score (IS) and a score for quantiles which we denote as quantile score (QS).
The logarithmic score LS for the i-th observation is given by

$$LS(p(\mathbf{y}_{rep}|\mathbf{y}), y_i) := \log p(y_{rep} = y_i|\mathbf{y}),$$

where $p(y_{rep} = y_i|\mathbf{y})$ denotes the posterior predictive density at $y_{rep} = y_i$ of the model under consideration. When a sample of MCMC iterates $\hat{\boldsymbol{\theta}}^j, j = 1, .., R$ after burnin is available, an approximation of $\log p(y_{rep} = y_i|\mathbf{y})$ for the i-th observation is straightforward, i.e.

$$\hat{\log} p(y_{rep} = y_i|\mathbf{y}) := \log\Big(\frac{1}{R}\sum_{j=1}^{R} p(y_i|\hat{\boldsymbol{\theta}}^j)\Big),$$

where $p(y|\hat{\boldsymbol{\theta}}^j)$ denotes the density at the observed value y based on the j-th MCMC iterates. In contrast to the logarithmic score which only considers the posterior predictive

distribution evaluated at the observed value, the following scoring rules take both calibration and sharpness into account.

The continuous ranked probability score CRPS for a parametric model P_θ with posterior predictive cumulative density function (cdf) $F(x) := \int_{-\infty}^{x} p(\tilde{y}|\mathbf{y})d\tilde{y}$ is defined by

$$CRPS(F, y_i) \;=\; -\int_{-\infty}^{\infty} (F(x) - 1\{x \geq y_i\})^2 dx,$$

where $1\{x \geq y\}$ takes the value 1 if $x \geq y$ and 0 otherwise. Hence, the CRPS can be interpreted as the integrated squared difference between the predictive and the empirical cdf based on the single observation y_i. The CRPS can be seen as the analogue to the Brier score for continuous variables. A graphical illustration of the CRPS is presented in Figure 3.1 when P_θ is a normal distribution with mean θ_1 and variance θ_2. Here the

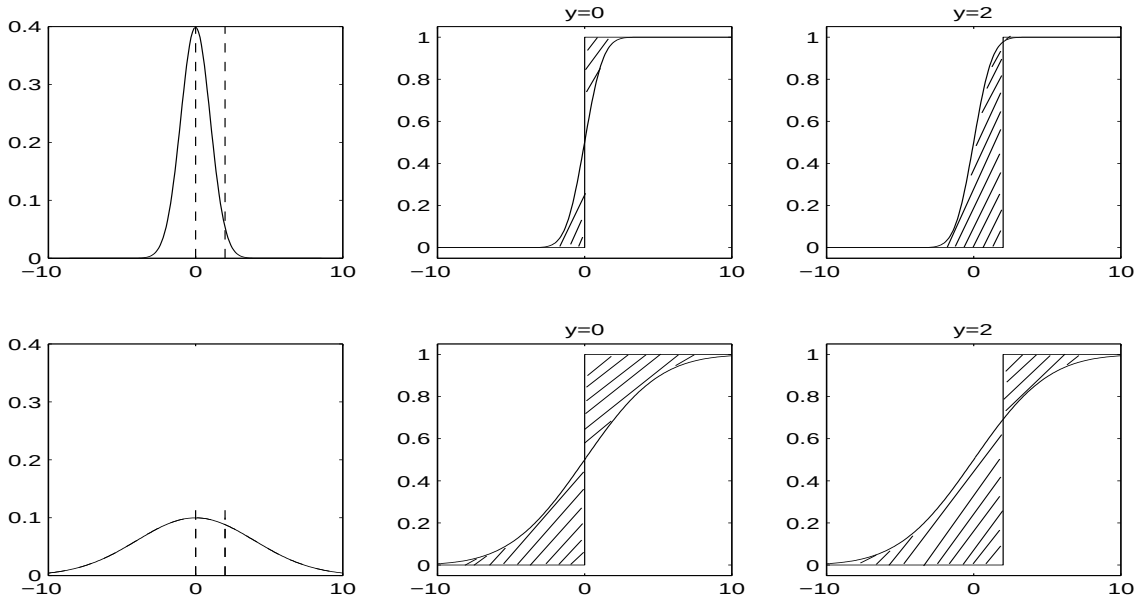

Figure 3.1: Pdf (left column with $y = 0, 2$ indicated as dashed lines) and cdf of a normal distribution with mean 0 and standard deviation 1 (first row) and 4 (second row), respectively. The differences between the cdf and the empirical cdf for two observations $y = 0$ (middle) and $y = 2$ (right) are indicated as dashed regions.

pdf of a normal distribution with mean 0 and standard deviation 1 (left panel in first row) and 4 (left panel in second row), respectively, is plotted. The difference between the corresponding cdf and the empirical cdf for two observations $y = 0$ and $y = 2$ is indicated in the middle and right plot of each row as dashed regions. These plots show that the CRPS rewards sharp distributions, but also takes into account if the observation y is close

to the center or rather in the tails of the distribution. According to Székely (2003) the CRPS can be expressed as

$$CRPS(F, y_i) \quad = \quad \frac{1}{2} E|y_{rep,i} - \tilde{y}_{rep,i}| - E|y_{rep,i} - y_i|. \tag{3.3}$$

Here $y_{rep,i}, \tilde{y}_{rep,i}$ are independent replicates from the posterior predictive distribution $p(\cdot|\mathbf{y})$ and the expectation is taken with respect to $p(\cdot|\mathbf{y})$. Estimation of the CRPS is again straightforward using the available MCMC output: for $j = 1, .., R$ simulate two replicated data sets $\mathbf{y}^j_{rep} = (y^j_{rep,1}, .., y^j_{rep,n}), \tilde{\mathbf{y}}^j_{rep} = (\tilde{y}^j_{rep,1}, .., \tilde{y}^j_{rep,n})$ based on the distribution $p(\mathbf{y}|\hat{\boldsymbol{\theta}}^j)$ and estimate the mean formula in (3.3) by $\hat{E}|y_{rep,i} - \tilde{y}_{,rep,i}| := \frac{1}{R}\sum_{j=1}^{R} |y^j_{rep,i} - \tilde{y}^j_{rep,i}|$ and $\hat{E}|y_{rep,i} - y_i| := \frac{1}{R}\sum_{j=1}^{R} |y^j_{rep,i} - y_i|$.

The interval score IS_α is based on the $(1-\alpha)$ 100 % posterior prediction interval defined by $I = [l_i, u_i]$ where l_i and u_i denote the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantile of the posterior predictive distribution for the i-th observation. It rewards narrow prediction intervals and assigns a penalty for observations which are not covered by the interval. The interval score is defined by

$$IS_\alpha(l_i, u_i, y_i) = \begin{cases} -2\alpha(u_i - l_i) - 4(l_i - y_i) & \text{if} \quad y_i \leq l_i \\ -2\alpha(u_i - l_i) & \text{if} \quad l_i \leq y_i \leq u_i \\ -2\alpha(u_i - l_i) - 4(y_i - u_i) & \text{if} \quad y_i \geq u_i \end{cases} .$$

Using the available MCMC output, replicated data sets $\mathbf{y}^j_{rep} = (y^j_{rep,1}, .., y^j_{rep,n}), j = 1, .., R$, can be simulated from which l_i and $u_i, i = 1, .., n$ can be determined. In order to compare models based on prediction intervals with both moderate and large coverage, we will use $\alpha = 0.1$ and $\alpha = 0.5$, respectively.

As will be seen in the application given in Chapter 8, the posterior predictive distribution of the total claim size in car insurance typically has most of its mass at zero. In particular, zero will in general be included in the posterior prediction intervals and the interval score will not be appropriate for model comparison. Here one sided scores might be more interesting to investigate. Gneiting and Raftery (2004) propose a proper scoring rule based on the quantiles $r_{\alpha,i}$ at level $\alpha \in (0, 1)$ of the predictive distribution for the i-th observation given by

$$S(r_{\alpha,i}, y_i) = \alpha s(r_{\alpha,i}) + (s(y_i) - s(r_{\alpha,i}))1\{y_i \leq r_{\alpha,i}\} + h(y_i)$$

for a nondecreasing function s and h arbitrary. We will use this score with the special choice $s(x) = x$ and $h(x) = -\alpha x$ and refer to the resulting scoring rule as quantile score

QS_α which is given by

$$QS_\alpha(r_{\alpha,i}, y_i) = (y_i - r_{\alpha,i})[1_{\{y_i \leq r_{\alpha,i}\}} - \alpha].$$

Analogous to the interval score, the α-quantile $r_{\alpha,i}$ of the posterior predictive distribution can be computed based on the MCMC output and evaluation of the quantile score is straightforward.

The predictive model choice criterion PMCC discussed in the previous section, can be expressed as a scoring rule as well. The corresponding positively oriented score function is defined by

$$S(P_\theta, y_i) = -(E(y_{rep,i}|\mathbf{y}) - y_i)^2 - Var(y_{rep,i}|\mathbf{y}).$$

However, this is not a proper scoring rule, see Gneiting and Raftery (2004), and should be used with care.

Chapter 4

Spatial modelling

In this thesis, models allowing for spatial dependencies for regional data are considered. This is carried out by introducing spatial random effects which incorporate a certain dependency structure. For regional data conditional autoregressive (CAR) models are a popular choice for modelling spatial patterns. CAR models are based on the assumption that the effects of adjacent sites are similar, leading to a spatially smoothed pattern. By this strength can be borrowed from neighbouring sites for the estimation of the spatial effects. Pettitt et al. (2002) propose a Gaussian conditional autoregressive (CAR) model for univariate data on irregularly spaced sites. In contrast to the widely used intrinsic CAR model, introduced by Besag and Kooperberg (1995), this model gives a proper joint distribution of the spatial effects. Further in a MCMC setting it allows for an efficient update of the spatial hyperparameter included in the model. We will define this model for regional data and use it for the modelling of spatial effects.

4.1 Gaussian conditional autoregressive models

Assume J sites $\{1, ..., J\}$ and let $\boldsymbol{\gamma} = (\gamma_1, .., \gamma_J)'$ denote the vector containing the spatial effects for each site. Then Pettitt et al. (2002) assume a multivariate normal distribution for $\boldsymbol{\gamma}$, in particular

$$\boldsymbol{\gamma} \sim N_J(\boldsymbol{\mu}, \sigma^2(I_J - C)^{-1}M)$$

where $\boldsymbol{\mu} = (\mu_1, .., \mu_J)'$, σ is a scale parameter and I_J the J-dimensional identity matrix. Here $C = (c_{ij})_{i,j=1,..,J}$ is a matrix with zeros along its main diagonal, i.e. $c_{ii} = 0$, whose off diagonal entries $c_{ij}, i \neq j$, are non-zero only when site i and j are neighbours, $M = diag(m_{ii}), i = 1, .., J$ is a diagonal matrix such that $(I_J - C)^{-1}M$ is symmetric and

positive-definite. The precision matrix Q is defined by $Q = M^{-1}(I_J - C)$. Pettitt et al. (2002) show that the full conditional of γ_i given all the remaining components $\boldsymbol{\gamma}_{-i} = (\gamma_1, .., \gamma_{i-1}, \gamma_{i+1}, .., \gamma_J)$ is given by

$$\gamma_i | \boldsymbol{\gamma}_{-i} \sim N(\mu_i + \sum_{j \sim i} c_{ij}(\gamma_j - \mu_j), \sigma^2 m_{ii}),$$

where we write $i \sim j$ if sites i and j are contiguous, i.e. if $c_{ij} \neq 0$. Pettitt et al. (2002) define sites to be neighbours if they lie within a certain distance δ. The scale of spatial dependence between site i and j is determined via the matrix $\boldsymbol{\varrho} = (\varrho_{ij})_{i,j=1,..,J}$ where

$$\varrho_{ij} = \begin{cases} \varrho(d_{ij}), & i \neq j \\ 0, & i = j \end{cases} \tag{4.1}$$

and d_{ij} denotes the Euclidean distance between sites i and j. They propose several choices for ϱ_{ij}, for example

$$\varrho_{ij} = \begin{cases} 1, & 0 < d_{ij} < \delta \\ 0, & d_{ij} \geq \delta \end{cases}$$

which gives equal weight to all neighbours of site i which are a distance δ away. The precision matrix $Q = (Q_{ij})_{i,j=1,..,J}$ is constructed using this matrix $\boldsymbol{\varrho}$ and the parameter ψ which determines the overall degree of spatial dependence by

$$Q_{ij} = \begin{cases} 1 + |\psi| \sum_{i \sim j} \varrho_{ij} & i = j \\ -\psi \varrho_{ij} & i \neq j \end{cases}.$$

This representation results from using matrices C and M defined as

$$c_{ij} := \begin{cases} \dfrac{\psi \varrho_{ij}}{1 + |\psi| \sum_{k \sim i} \varrho_{ik}}, & j \neq i \\ 0, & j = i \end{cases}$$

and

$$m_{ii} := \frac{1}{1 + |\psi| \sum_{k \sim i} \varrho_{ik}}, \quad i = 1, 2, .., J$$

which satisfies $c_{ij} m_{jj} = c_{ji} m_{ii} \quad \text{for} \quad i, j = 1, .., J$ and implies that Q is symmetric. In addition they show that Q is positive-definite.

In this thesis we are dealing with regional data and we will use a slightly altered version of the above model. We define regions to be neighbours if they share a common border and additionally take $\boldsymbol{\mu}$ to be zero, i.e.

$$\boldsymbol{\gamma} \sim N_J(0, \sigma^2 Q^{-1}).$$

Defining the matrix $\boldsymbol{\varrho}$ as

$$\varrho_{ij} = \begin{cases} 1 & i \sim j \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

we obtain with this special choice

$$Q_{ij} = \begin{cases} 1 + |\psi| \cdot N_i & i = j \\ -\psi & i \neq j, i \sim j \\ 0 & \text{otherwise} \end{cases}, \tag{4.3}$$

where N_i denotes the number of neighbours of region i. Thus the conditional distribution of γ_i, given all the remaining components $\boldsymbol{\gamma}_{-i}, i = 1, .., J$ is given by

$$\gamma_i | \boldsymbol{\gamma_{-i}} \sim N\Big(\frac{\psi}{1 + |\psi| \cdot N_i} \sum_{j \sim i} \gamma_j, \sigma^2 \frac{1}{1 + |\psi| \cdot N_i}\Big). \tag{4.4}$$

The partial correlation between γ_i and γ_j given all other regions is given by

$$corr(\gamma_i, \gamma_j | \boldsymbol{\gamma}_{-\{i,j\}}) = \begin{cases} \dfrac{\psi}{\sqrt{1 + |\psi| N_i}\sqrt{1 + |\psi| N_j}} & , i \sim j \\ 0 & , \text{otherwise} \end{cases},$$

see Pettitt et al. (2002). In the conditional mean the conditional autoregressive structure of the model is reflected: conditionally on the effects of all other regions, the effect of region i only depends on the neighbouring regions. The conditional variance of the effect of region i depends both on the spatial hyperparameters ψ, σ^2 and on the number of neighbours of this region. The more neighbours a region has, the stronger is the spatial smoothing leading to a decreased conditional variance. For $\psi = 0$ all regions are spatially independent, the precision matrix Q given in (4.3) reduces to the identity matrix in this case. For $\psi \to \infty$ the asymptotic behaviour of this model can be presented as follows:

$$\begin{cases} \gamma_J \sim N(0, \sigma^2 V_{22}) \quad \text{with} \quad \lim_{\psi \to \infty} V_{22} = \frac{1}{J} \\ \gamma_{-\mathbf{J}|\gamma_J} \sim N_{J-1}(\mu_J, \sigma^2 Q_{11}^{-1}) \\ \text{with} \lim_{\psi \to \infty} \mu_J = \gamma_J \cdot \mathbf{1}_{J-1} \quad \text{and} \quad \lim_{\psi \to \infty} Q_{11}^{-1} = 0 \end{cases}, \tag{4.5}$$

where $Q_{11} \in \mathbb{R}^{J-1,J-1}, V_{22} \in \mathbb{R}$ and $-\mathbf{J} = (1, 2, .., J-1)$ denotes all indices except J. The proof is shown in the Appendix A. Hence, with $\psi \to \infty$ the covariance matrix of the spatial effects tends to zero, resulting in a strong smoothed spatial pattern, i.e. the degree of dependence increases. Note, that $\psi > 0$ induces positive conditional partial correlations, while for $\psi < 0$ negative conditional partial correlations are obtained. In the

remainder of this thesis ψ will be restricted to take positive values only since we expect similar, positive correlated effects for neighbouring regions.
To get a better idea of the overall dependence parameter ψ, we consider the correlations between neighbours of first to fifth order for several values of ψ. Since σ is only a scale parameter, it has no influence on the correlations. Note, that we look at the unconditional correlations here. The order of a neighbour will be explained below. Assume 100 regions on a 10×10 grid. We consider two different neighbourhood structures:

Structure 1: regions with a common border are neighbours

Structure 2: regions with at least a common corner are neighbours

These structures are illustrated in Figure 4.1 for a region in the corner, on the middle edge and in the middle of the grid. Neighbours of the same order are plotted in the same colour. Direct neighbours are referred to as neighbours of order 1, the next neighbours of those to neighbours of order 2 and so on.
The correlations between the region in the corner, middle edge and middle of the grid and its neighbours are plotted in Figure 4.2 for Structure 1 and in Figure 4.3 for Structure 2 for $\psi = 1, 3$ and 6. On the x-axis the order of the neighbours is given. First of all, the correlations decrease polynomial with increasing order of the neighbours. Correlations to neighbours of the same order are not necessarily the same for all neighbours as the distances to the region of interest are different. The highest correlations are obtained for the region in the corner of the grid since the spatial dependency is distributed on fewer neighbours there. Consequently the correlations between the middle region and its neighbours are the smallest. As supposed by the model the correlations increase with increasing values of ψ. The correlations behave similarly for both neighbourhood structures with the difference that there is a higher variation of the correlations to neighbours of the same order for structure 2. This is due to the fact that this structure considers more regions as neighbours.

4.2 Computational issues

A convenient feature of this Gaussian CAR model is that the determinant of Q which is needed for the update of ψ in a Markov Chain Monte Carlo (MCMC) algorithm can be

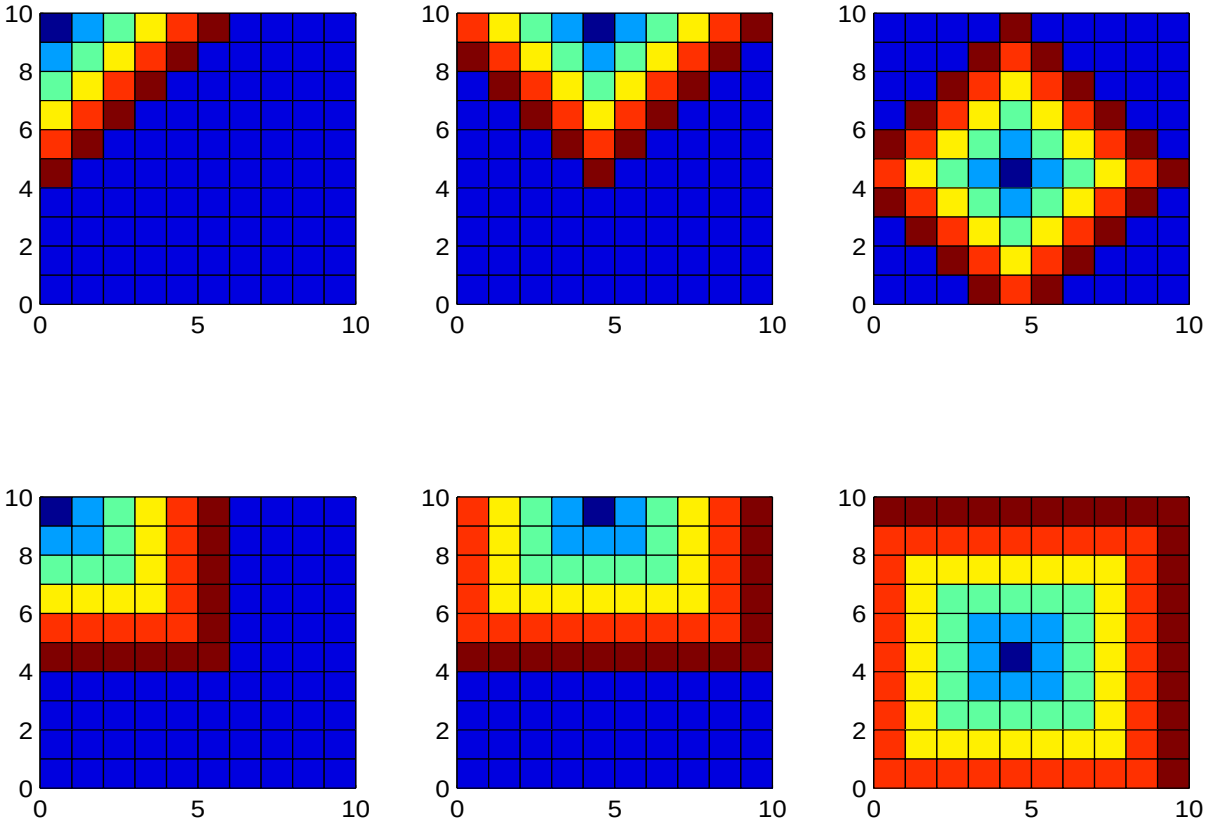

Figure 4.1: Neighbours of order 1 to 5 of a region on the corner, middle edge and middle of the grid, using two different neighbour structures (first row: Structure 1, second row: Structure 2).

computed efficiently using the following results by Pettitt et al. (2002) :

$$\begin{aligned} Q &= I + |\psi| D - \psi \boldsymbol{\rho} \\ &= \begin{cases} I - \psi(\boldsymbol{\rho} - D), & \psi > 0 \\ I, & \psi = 0 \\ I - \psi(\boldsymbol{\rho} + D), & \psi < 0 \end{cases}, \end{aligned} \tag{4.6}$$

where $D = diag(\sum_{k \sim i} \rho_{ik}, i = 1, .., J)$. With (4.2), D can be written as $D = diag(N_i, i = 1, .., J)$. Let $\lambda_i, i = 1, .., J$ denote the eigenvalues of $\boldsymbol{\rho} - D$ and $\nu_i, i = 1, .., J$ denote the eigenvalues of $\boldsymbol{\rho} + D$, then the eigenvalues $\xi_i, i = 1, .., J$ of Q are given by

$$\xi_i = \begin{cases} 1 - \psi\lambda_i, & \psi > 0 \\ 1, & \psi = 0 \\ 1 - \psi\nu_i, & \psi < 0 \end{cases}. \tag{4.7}$$

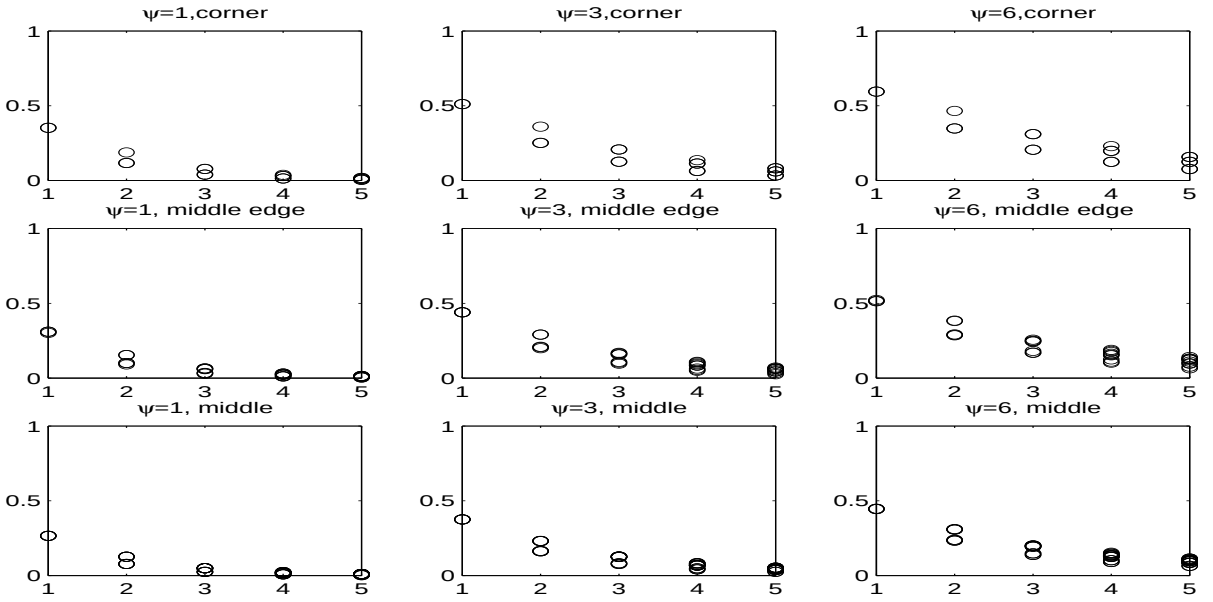

Figure 4.2: Correlations between a region on the corner, middle edge and middle of the grid and its neighbours of order 1 to 5, using neighbour structure 1.

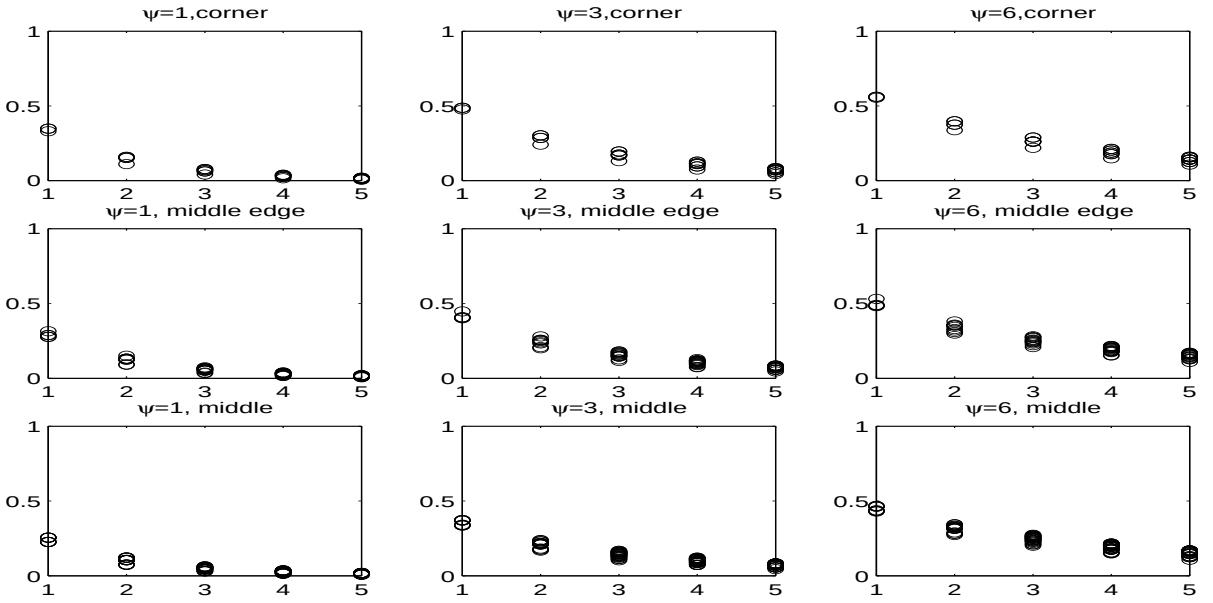

Figure 4.3: Correlations between a region on the corner, middle edge and middle of the grid and its neighbours of order 1 to 5, using neighbour structure 2.

The determinant of Q is therefore given by

$$|Q| = \begin{cases} \prod_{i=1}^{J}(1-\psi\lambda_i), & \psi > 0 \\ 1, & \psi = 0 \\ \prod_{i=1}^{J}(1-\psi\nu_i), & \psi < 0 \end{cases} . \tag{4.8}$$

Since the eigenvalues λ_i and ν_i only depend on the number of neighbours and the neighbourhood structure, they only have to be determined once at the beginning of the MCMC algorithm and can then be reused for the calculation of the determinant of Q in every iteration.

4.3 Related conditional autoregressive models

Many other authors have dealt with conditional autoregressive models. An overview about CAR models is given in the book by Banerjee et al. (2004) and in Jin et al. (2004) where also multivariate CAR models are discussed. The most popular model is probably the intrinsic CAR model introduced by Besag and Kooperberg (1995) where the full conditional of γ_i given $\boldsymbol{\gamma_{-i}}$ is given by

$$\gamma_i|\boldsymbol{\gamma_{-i}} \sim N\Big(\sum_{j\sim i}\frac{\gamma_j}{N_i}, \frac{\sigma^2}{N_i}\Big). \tag{4.9}$$

This model can be extended to the weighted version

$$\gamma_i|\boldsymbol{\gamma_{-i}} \sim N\Big(\sum_{j}\frac{\rho_{ij}}{\sum_j \rho_{ij}}\gamma_j, \frac{\sigma^2}{\sum_j \rho_{ij}}\Big), \tag{4.10}$$

where ρ_{ij} are the elements of a symmetric positive-definite matrix similar to (4.1). If ρ_{ij} is chosen as in (4.2) this model reduces to the unweighted intrinsic CAR model (4.9). The joint density for $\boldsymbol{\gamma}$ in the intrinsic CAR model is improper in contrast to model (4.4) described above which has a proper joint density. Therefore, using the intrinsic CAR model care must be taken in order to achieve propriety of the posterior distribution.
Czado and Prokopenko (2004) consider a modification of model (4.4) given by

$$\gamma_i|\boldsymbol{\gamma_{-i}} \sim N\Big(\frac{\psi}{1+|\psi|\cdot N_i}\sum_{j\sim i}\gamma_j, \sigma^2\frac{1+|\psi|}{1+|\psi|\cdot N_i}\Big) \tag{4.11}$$

where the conditional variance is multiplied by the additional term $1+|\psi|$. This is a proper model as well but in the limit $\psi \to \infty$ reduces to the intrinsic CAR model.
Another modification of the intrinsic CAR model leading to a proper joint distribution has been presented by Sun et al. (1999). They introduce the parameter $|\varrho| < 1$ such that

$$\gamma_i|\boldsymbol{\gamma_{-i}} \sim N\Big(\varrho\sum_{j\sim i}\frac{\gamma_j}{N_i}, \frac{\sigma^2}{N_i}\Big) \tag{4.12}$$

to get a proper multivariate normal distribution for $\boldsymbol{\gamma}$. Here, the intrinsic CAR model is the limiting case for $\varrho = 1$. A multivariate version of this model has also been used by Gelfand and Vounatsou (2003) to model spatial effects in hierarchical models.

Chapter 5

Spatial regression models for count data

In this chapter regression models for count data allowing for overdispersion and spatial dependence patterns are considered. Probably the most popular model for count data is the Poisson distribution. For a random variable Y the density of the Poisson distribution with parameter μ is given by

$$P(Y = y|\mu) = \exp(-\mu)\frac{\mu^y}{y!}.$$

We use the notation $Y \sim Poi(\mu)$. In the Poisson model equality of the mean and the variance is assumed, in particular

$$E(Y|\mu) = Var(Y|\mu) = \mu.$$

However, for count data often overdispersion is observed, i.e. the variance in the data is greater than the mean. In this case, the Poisson distribution is not appropriate any more and more flexible models, which allow the variance to be larger than the mean, should be used. We follow two approaches for dealing with the extra variability in overdispersed data. On the one hand, we consider a wider class of models allowing for overdispersion, on the other hand spatial random effects are introduced to capture unobserved spatial heterogeneity in the data.
Overdispersion with respect to the Poisson model can be modelled by introducing an additional parameter. In particular we consider the negative binomial (NB) distribution and the generalized Poisson (GP) distribution introduced by Consul and Jain (1973). Both models allow an independent modelling of the mean and the variance by the inclusion of an additional parameter.

When dealing with a data set with an excessive number of zeros, zero-inflated models might be used, see for example Winkelmann (2003). In contrast to the GP and the NB model, overdispersion in zero inflated models is caused by the occurrence of more zero observations than expected. Zero inflated models can be used in combination with any model for count data. Additionally to the zero observations arising from the count data model an extra proportion of zeros is incorporated. Lambert (1992) introduced the zero inflated Poisson regression model, a Bayesian analysis of the zero inflated Poisson model is given in Rodrigues (2003). Zero inflated regression models in combination with the generalized Poisson distribution have been addressed in Famoye and Singh (2003b) and Famoye and Singh (2003a) using maximum likelihood estimation, a Bayesian analysis without the inclusion of covariates is given in Angers and Biswas (2003). Agarwal et al. (2002) use a zero inflated Poisson regression model for spatial count data in a Bayesian framework.

The second approach for modelling unobserved data heterogeneity is the introduction of random effects. For spatially indexed data which are the focus of this thesis, spatial random effects associated with each region or site may be used, allowing for the modelling of an underlying spatial dependency structure.

In the following we consider Poisson, NB, GP and zero inflated (ZI) regression models both including and without spatial random effects in a Bayesian context. A spatial correlation structure is incorporated by assuming the Gaussian conditional autoregressive spatial prior presented in Chapter 4 for the spatial effects. Since this results in a high dimensional, complex posterior distribution, Markov Chain Monte Carlo (MCMC) is used for parameter estimation.

The considered models will then be applied for analysing the number of invasive meningococcal disease cases reported in Germany in the year 2004. An application of some of the models on a large data set from a German car insurance company will be given in Chapter 8. Here the number of claims of a policyholder within one year will be modelled. This chapter is organized as follows. In Section 5.1 the negative binomial, the generalized Poisson and zero-inflated regression models are presented. The regression set up and prior specifications for the regression and model dependent overdispersion parameters as well as for the spatial effects are given in Section 5.2. Finally, in Section 5.3 the application on invasive meningococcal disease data in Germany is given. Models are compared using the deviance information criterion (DIC) and proper scoring rules, see Chapter 3. We observe a substantial degree of overdispersion in the data which is modelled best by the GP distribution when spatial effects are neglected. While the addition of spatial random

effects gives no or little improvement to the models allowing for overdispersion, spatial effects turn out to be significant for the Poisson model. In particular, according to the DIC and the scoring rules a spatial Poisson model gives the best fit for these data. However, no smooth spatial pattern is modelled. Instead some isolated regions with high risk are detected by the spatial effects, indicating that the risk is not sufficiently explained by the incorporated covariates in these regions.
Conclusions are drawn in Section 5.4. Details about the MCMC algorithms can be found in the Appendix B, a brief summary of the used update schemes is given in Section 5.5.

5.1 Models for overdispersed count data

5.1.1 Negative Binomial (NB) distribution and Regression

The density of the negative binomial distribution with parameters $r > 0$ and $\mu > 0$ denoted by NB(r,μ) is defined by

$$P(Y = y|r, \mu) = \frac{\Gamma(y + r)}{\Gamma(r)y!} \cdot \Big(\frac{r}{\mu + r}\Big)^r \cdot \Big(\frac{\mu}{\mu + r}\Big)^y, \qquad y = 0, 1, 2, ... \tag{5.1}$$

with mean and variance specified by

$$E(Y|r, \mu) = \mu \quad \text{and} \quad Var(Y|r, \mu) = \mu\Big(1 + \frac{\mu}{r}\Big).$$

The variance is the mean multiplied by the positive factor $\varphi := 1 + \frac{\mu}{r}$ and therefore greater than the mean, i.e. overdispersion can be modelled in the negative binomial distribution. We call the factor φ dispersion factor. In the limit $r \to \infty$ the NB distribution converges to the Poisson distribution with parameter μ, see Winkelmann (2003). The negative binomial distribution also arises from a Poisson distribution where the parameter θ is assumed to be random and to follow a Gamma distribution with mean $E(\theta) = \mu$ and variance $Var(\theta) = \frac{\mu^2}{r}$. Therefore, overdispersion in the NB model can be interpreted by unobserved heterogeneity among observations. From $\dfrac{P(Y = y + 1)}{P(Y = y)} = \dfrac{y + r}{y + 1}\Big(\dfrac{\mu}{\mu + r}\Big)$, we can derive that $P(Y = y + 1) > P(Y = y)$ if $y < \frac{\mu r - \mu - r}{r} := k$. Therefore, if k is not an integer, the NB distribution is unimodal with mode at $y = \lfloor k \rfloor$, i.e. the integer part of k. If k is an integer there are two modes at $y = k$ and $y = k + 1$.
In a regression model with $Y_i \sim \text{NB}(r, \mu_i)$ independent for $i = 1, .., n$, the mean of Y_i is specified in terms of covariates $\mathbf{x}_i$ and unknown regression parameters $\boldsymbol{\beta}$ by

$$E(Y_i|\mathbf{x_i}, \boldsymbol{\beta}) = \mu_i > 0.$$

Note, that in the NB regression model the dispersion factor $\varphi_i := 1 + \frac{\mu_i}{r}$ takes observation specific values.

5.1.2 Generalized Poisson (GP) distribution and Regression

The generalized Poisson distribution has been introduced by Consul and Jain (1973) and is investigated in detail in Consul (1989). A random variable Y is called generalized Poisson distributed with parameters θ and λ, denoted by GP(θ, λ), if

$$P(Y = y|\theta, \lambda) = \begin{cases} \theta(\theta + y\lambda)^{y-1}\frac{1}{y!}\exp(-\theta - y\lambda), & y = 0, 1, 2, ... \\ 0 & \text{for } y > m \text{ when } \lambda < 0 \end{cases} \tag{5.2}$$

where $\theta > 0$, $\max(-1, -\frac{\theta}{m}) < \lambda \leq 1$ and $m \geq 4$ is the largest positive integer for which $\theta + m\lambda > 0$ for negative λ. For $\lambda < 0$ the model gets truncated. In this case, the lower limit for λ ensures, that there are at least five classes with positive probability.
Mean and variance are given by

$$E(Y|\theta, \lambda) = \frac{\theta}{1-\lambda} \quad \text{and} \quad Var(Y|\theta, \lambda) = \frac{\theta}{(1-\lambda)^3} = E(Y|\theta, \lambda) \cdot \frac{1}{(1-\lambda)^2}, \tag{5.3}$$

hence $\varphi := \frac{1}{(1-\lambda)^2}$ can be interpreted as an dispersion factor for the GP distribution. For $\lambda = 0$, the generalized Poisson distribution reduces to the Poisson distribution with parameter θ, equality of mean and variance are obtained in this case. For $\lambda < 0$ underdispersion can be modelled, whereas for $\lambda > 0$ overdispersion is obtained. Figure 5.1 illustrates the density of the GP distribution for $\theta = 8$ and several values of λ. A larger value of $\lambda > 0$ corresponds to a greater overdispersion, i.e. a larger variance. A small value of λ in contrast clearly reduces the variance. In Figure 5.2 the density of the generalized Poisson distribution is shown for $\theta = 1, 3, 5$ and $\lambda = -0.2, 0.2$. The larger θ, the larger the mean, i.e. a greater shift to the right can be seen. Consul (1989) shows that the GP distribution is unimodal for all values of θ and λ. In the remainder of this thesis only the problem of overdispersion will be addressed, i.e. λ will be assumed to take only values in the interval $[0, 1)$. Similar to the NB model, the GP distribution is a mixture of Poisson distributions as has been proved by Joe and Zhu (2005).
A regression model for independent GP distributed response variables $Y_i, i = 1, .., n$ is set up by specifying the mean in terms of covariates and regression parameters by

$$E(Y_i|\boldsymbol{x_i}; \boldsymbol{\beta}, \lambda) = \mu_i > 0$$

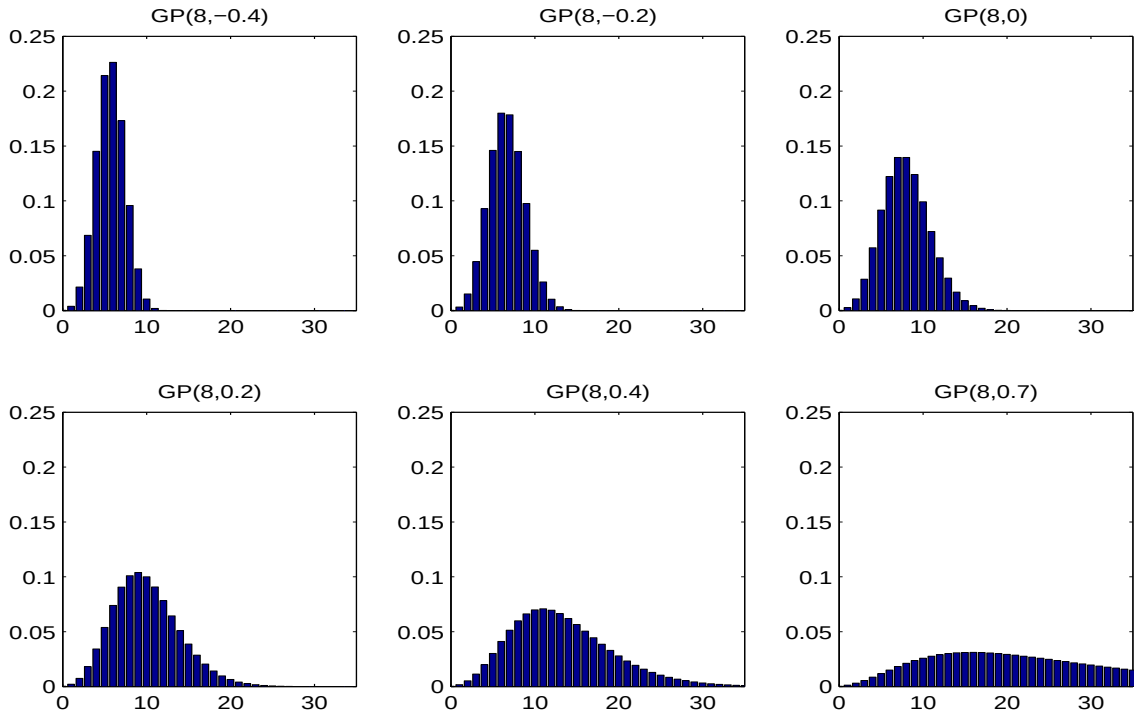

Figure 5.1: Probability mass function of the generalized Poisson distribution with $\theta = 8$ and $\lambda = -0.4, -0.2, 0, 0.2, 0.4, 0.75$.

like in the NB model. Using (5.3) yields that $\mu = \dfrac{\theta}{1-\lambda}$ and leads to another parameterisation of the GP model, denoted by $GP(\mu, \lambda)$, given by

$$P(Y = y|\mu, \lambda) = \mu[\mu(1-\lambda) + \lambda y]^{y-1} \frac{(1-\lambda)}{y!} \exp\Big[-\mu(1-\lambda) - \lambda y\Big],\ y = 0, 1, 2, ... \quad (5.4)$$

Throughout the remainder of this thesis this parameterisation of the GP distribution will be used. While the dispersion parameter in the NB regression model depends on μ_i leading to a variance function which is quadratic in μ_i, the dispersion parameter $\varphi = \frac{1}{(1-\lambda)^2}$ in the GP regression model is the same for each observation and results in a linear variance function.

5.1.3 Comparison of NB and GP distribution

In order to compare the behaviour of the NB and the GP distribution, we equate the mean and the variance of a GP(μ, λ) with the mean and the variance of a NB(r, μ) distributed random variable, i.e.

$$\frac{\mu}{(1-\lambda)^2} = \mu(1 + \frac{\mu}{r})$$

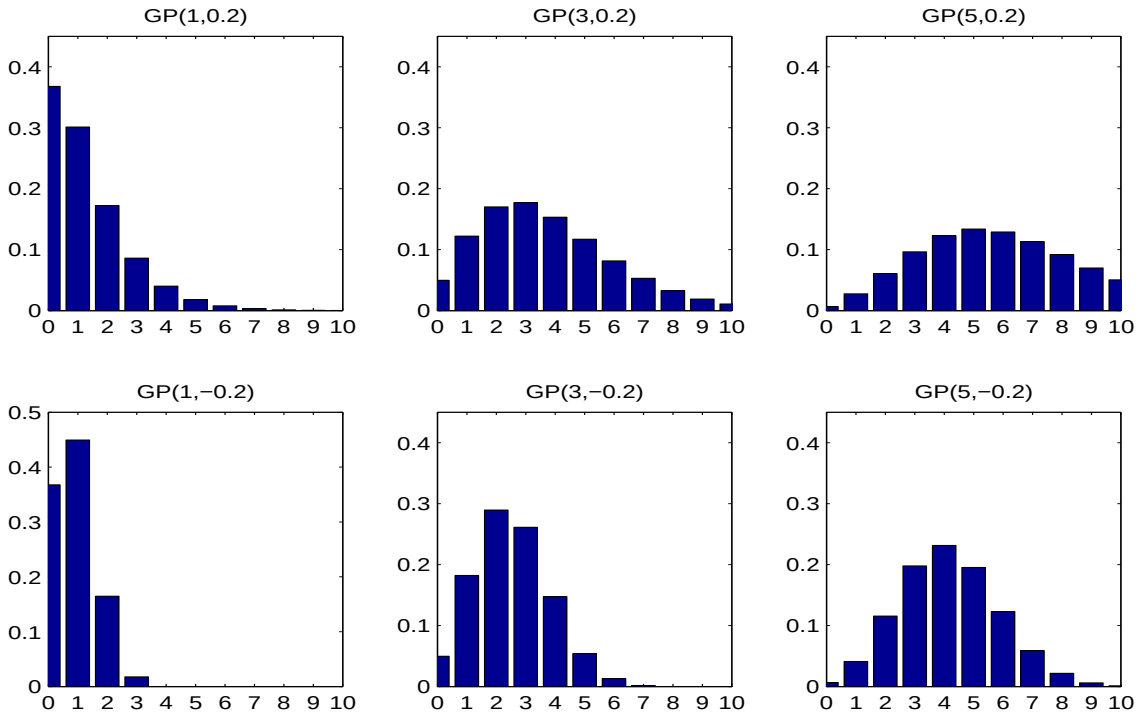

Figure 5.2: Probability mass function of the Generalized Poisson distribution with $\theta = 1, 3, 5$ and $\lambda = 0.2, -0.2$.

has to hold and the equation

$$r = \frac{\mu(1-\lambda)^2}{\lambda(2-\lambda)} \tag{5.5}$$

is obtained. In Figure 5.3 the NB distribution is plotted in comparison to the GP distribution with equal mean and variance, i.e. with μ and r chosen according to (5.5). For a better visual comparison the densities of these discrete distributions are presented as line plots. For small values of λ both distributions behave very similarly. With increasing values of λ slight differences between the two distributions can be observed which become greater when λ tends to 1. In particular, the NB distribution gives more mass to small values of y if a large overdispersion is present.

Comparison of GP and NB model via empirical variances

Assume we have Poisson data $y_i \sim Poi(\mu_i)$, $i = 1, .., n$ and fit the appropriate Poisson model. Then, due to the equality of variance and mean in the Poisson model, a plot of the estimated posterior means $\hat{\mu}_i$ against the empirical variances $Var(y_i)^{emp}$, $i = 1, .., n$

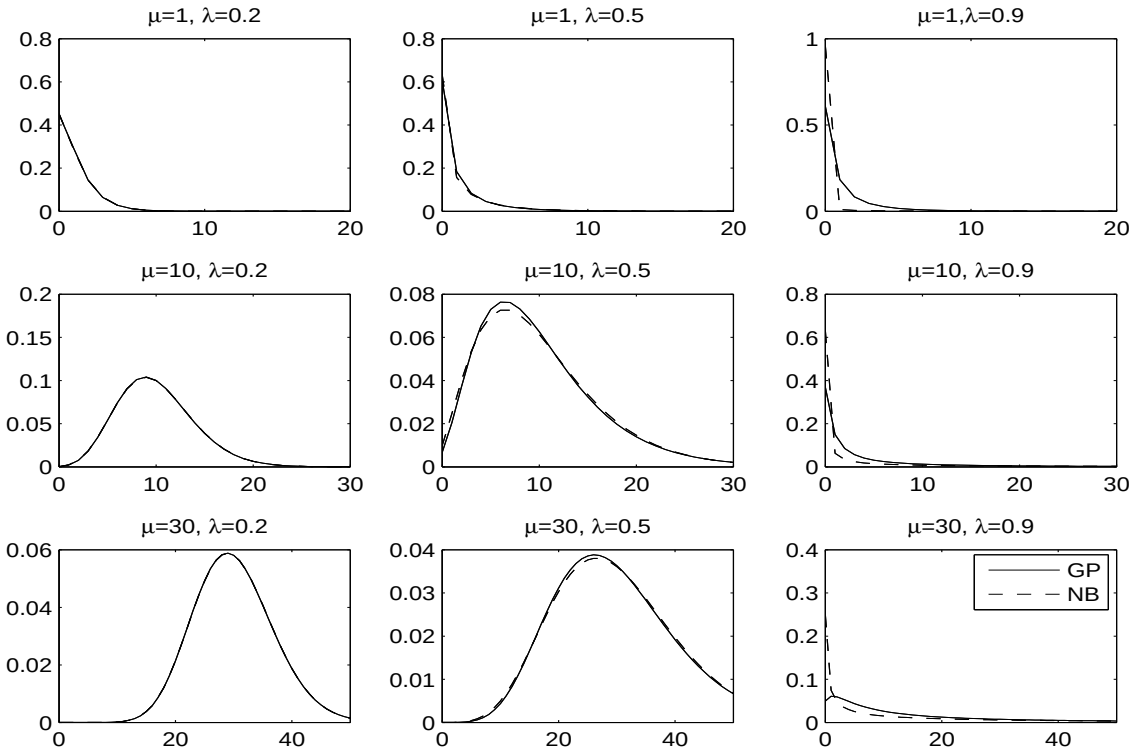

Figure 5.3: Comparison of the generalized Poisson distribution (solid line) with $\mu = 1, 10, 30$ and $\lambda = 0.2, 0.5, 0.9$ to the negative binomial distribution (dashed line) with $\mu = 1, 10, 30$ and $r = \frac{\mu(1-\lambda)^2}{\lambda(2-\lambda)}$.

should give approximately a 45 degree line. What happens if the data are generalized Poisson or negative binomial distributed but a Poisson model is assumed? Gives this plot an idea of the true underlying model? To check this we conduct a simulation study.

Simulation study

In order to get a good estimator for the empirical variance we use a replicated data set. In particular we assume three covariates, an intercept, an indicator variable $\mathbf{x}_1$ and a continuous standardized covariate $\mathbf{x}_2$, for 500 observations $\mathbf{x}_i = (1, x_{i1}, x_{i2})$, $i = 1, .., 500$ and replicate each observation 20 times. Then, we simulate both $GP(\mu_i, \lambda)$ and $NB(r, \mu_i)$ distributed response variabels $y_{ij}, i = 1, .., 500, j = 1, .., 20$ where $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ with some known $\boldsymbol{\beta}$ for $\lambda = 0, 0.2, 0.5, 0.8$ and $r = 0, 10, 100$. $\boldsymbol{\beta}$ is chosen in such a way that we obtain response values up to about 120. For each of these seven simulated data sets a Poisson model is fitted and the posterior mean $\hat{\mu}_i$ is estimated by $\hat{\mu}_i = \frac{1}{R}\sum_{j=1}^{R}\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}^j)$ where R gives the number of MCMC iterations after burnin and $\hat{\boldsymbol{\beta}}^j = (\hat{\beta}_0^j, \hat{\beta}_1^j, \hat{\beta}_2^j)$ de-

notes the j-th draw of the regression parameters. The empirical variance is estimated by $Var(y_i)^{emp} = \frac{1}{19}\sum_{j=1}^{20}(y_{ij} - \bar{y}_i)^2$ where $\bar{y}_i = \frac{1}{20}\sum_{j=1}^{20} y_{ij}$. The resulting scatter plots of $\hat{\mu}_i$ and $Var(y_i)^{emp}$, $i = 1, .., 500$ are given in Figures 5.4 and 5.5 for the GP and the NB data, respectively. The solid line in each plot is a smooth of the data whereas the dotted line gives the expected variances according to the true models, i.e. $\frac{\mu_i}{(1-\lambda)^2}$ in the GP model and $\mu_i(1 + \frac{\mu_i}{r})$ in the NB Model where for λ and r the true values are taken.

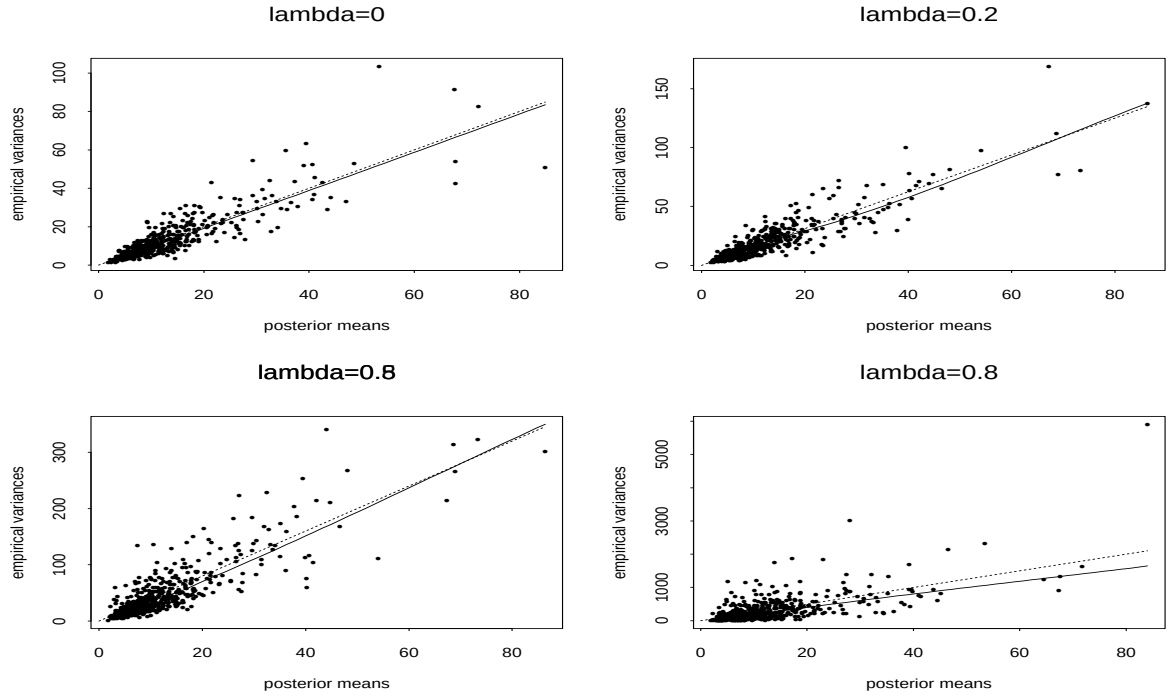

Figure 5.4: Scatter plot of $\hat{\mu}_i$ and $Var(y_i)^{emp}$, $i = 1, .., 500$ for $GP(\mu_i, \lambda)$ data with $\lambda = 0, 0.2, 0.5, 0.8$ when a Poisson Model is fitted, solid line gives a smooth of the data, the dotted line gives $\frac{\mu_i}{(1-\lambda)^2}$.

For the GP data in Figure 5.4 the smoothed curve and the dotted lines are very close for all values of λ. The slope of the smoothed curve indeed represents the degree of overdispersion present in the data. Therefore, if the scatter plot of the estimated posterior means and the empirical variances in a Poisson model significantly deviates from the 45 degree line, this is an indication for overdispersion in the data. For the NB data the empirical variances, indicated by the smooth of the data, and the variances expected from the true underlying model are reasonable close as well, especially for $r = 100$, i.e. data close to the Poisson distribution. For $r = 10, 1$ the scatter plot clearly is not along the 45 degree line, again indicating the presence of overdispersion. In comparison to the GP data where the plots scatter along a rather straight line, a slight quadratic

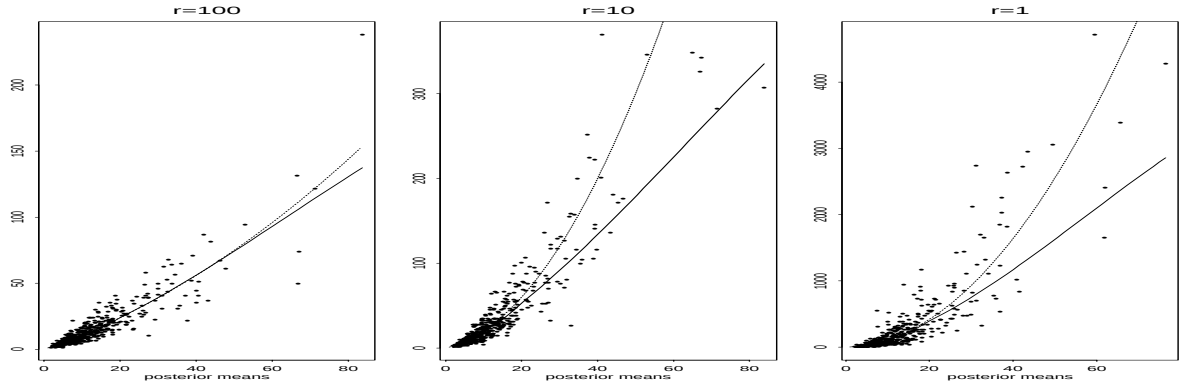

Figure 5.5: Scatter plot of $\hat{\mu}_i$ and $Var(y_i)^{emp}$, $i = 1, .., 500$ for $NB(r, \mu_i)$ data with $r = 100, 10, 1$ when a Poisson model is fitted, solid line gives a smooth of the data, the dotted line gives $\mu_i(1 + \frac{\mu_i}{r})$.

trend can be observed for the NB data. This is expected since the variance in the NB model is a quadratic form of the mean. Therefore scatterplots of the estimated posterior means and the empirical variances in a Poisson model may not only reveal the presence of overdispersion but indicate as well whether a GP or a NB model is more appropriate. Similar results where obtained for replicated data with smaller response values. However, if the data are not replicated, it is hard to find an appropriate estimator for the empirical variance. The straightforward estimator $Var(y_i)^{emp} = \frac{1}{R}\sum_{j=1}^{R}(y_i - \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}^j))^2$ does not work well, no clear pattern in the scatter plots is visible. Therefore these plots do not contain any valuable information for the real data in our application.

5.1.4 Zero Inflated (ZI) Models

For count data with an excessive number of zero observations zero inflated (ZI) models can be used. These models allow for a higher number of zeros than can be explained by standard models for count data. Additional to the zero observations arising from the supposed count data distribution, a proportion of extra zeros is assumed. ZI models have been widely used in the literature, see Winkelmann (2003) for an overview.
Let $\pi(y|\boldsymbol{\theta})$ be a distribution function for count data with unknown parameters $\boldsymbol{\theta}$. Then a zero inflated model with extra proportion $p \in [0, 1]$ of zeros is defined by (see Agarwal et al. (2002))

$$P(Y = y|p, \boldsymbol{\theta}) = \begin{cases} p + (1-p)\pi(y = 0|\boldsymbol{\theta}) & \text{if } y = 0 \\ (1-p)\pi(y|\boldsymbol{\theta}) & \text{if } y > 0 \end{cases} \tag{5.6}$$

Mean and variance are given by

$$E(Y|p,\boldsymbol{\theta}) = (1-p)E_\pi(Y|\boldsymbol{\theta}) \tag{5.7}$$

and

$$Var(Y|p,\boldsymbol{\theta}) = p(1-p)[E_\pi(Y|\boldsymbol{\theta})]^2 + (1-p)Var_\pi(Y|\boldsymbol{\theta}). \tag{5.8}$$

The introduction of latent indicator variables $\mathbf{Z} = (Z_1,..,Z_n)'$ leads to a model which is easier to handle in a Bayesian context and in particular allows a Gibbs step for p. Z_i takes the value $z_i = 0$ for all observations with $y_i > 0$. For all zero observations $y_i = 0$, the latent variable takes the value $z_i = 0$ if observation i arises from the count data distribution $\pi(y|\boldsymbol{\theta})$ and the value $z_i = 1$ if it is an extra zero. The joint distribution of Y_i and Z_i is therefore determined by

$$P(Y_i = 0, Z_i = 1|p_i,\boldsymbol{\theta}) = p_i, \quad P(Y_i = 0, Z_i = 0|p_i,\boldsymbol{\theta}) = (1-p_i)\pi(y_i = 0|\theta),$$
$$P(Y_i = y_i, Z_i = 1|p_i,\boldsymbol{\theta}) = 0, \quad P(Y_i = y_i, Z_i = 0|p_i,\boldsymbol{\theta}) = (1-p_i)\pi(y_i|\boldsymbol{\theta}), \quad y_i > 0$$

which can be written succinctly

$$P(Y_i = y_i, Z_i = z_i|p_i,\boldsymbol{\theta}) = p_i^{z_i}[(1-p_i)\pi(y_i|\boldsymbol{\theta})]^{1-z_i}.$$

Marginally, $Z_i \sim Bernoulli(p_i)$. Using the latent variables $\mathbf{Z}$, the joint likelihood of $\mathbf{Y} = (Y_1,..,Y_n)'$ and $\mathbf{Z}$ is given by

$$f(\mathbf{Y},\mathbf{Z}|p,\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{z_i}[(1-p_i)\pi(y_i|\boldsymbol{\theta})]^{1-z_i} \tag{5.9}$$
$$= \prod_{i:y_i=0} p_i^{z_i}[(1-p_i)\pi(0|\boldsymbol{\theta})]^{1-z_i} \cdot \prod_{i:y_i>0}(1-p_i)\pi(y_i|\boldsymbol{\theta}) \tag{5.10}$$

In this thesis we will focus on the zero inflated Poisson and the zero inflated generalized Poisson models, which are special cases of the ZI model (5.6). The zero inflated negative binomial distribution will not be discussed, since the GP model turned out to be more adequate than the NB model for the application considered later on.

Zero Inflated Poisson (ZIP) Distribution

Here the Poisson distribution is assumed for the underlying count data distribution, i.e. $\pi(y|\theta) := \pi(y|\mu) = \mu^y\frac{\exp(-\mu)}{y!}$. Using (5.7) and (5.8), mean and variance of the ZIP distribution, denoted by $ZIP(p,\mu)$, are specified by

$$E(Y|p,\mu) = (1-p)\mu$$

and

$$Var(Y|p,\mu) = (1-p)\mu(\mu p+1) = E(Y|p,\mu)(\mu p+1).$$

For $p > 0$ the dispersion factor $\varphi := \mu p + 1$ of the ZIP model is positive, i.e. the presence of extra zeros leads to overdispersion.

Zero Inflated Generalized Poisson (ZIGP) Distribution

The ZIGP regression model was already introduced by Famoye and Singh (2003b), in Famoye and Singh (2003a) a generalisation to k-inflated GP regression models is given. Czado and Min (2005) show asymptotic existence, consistency and asymptotic normality of the maximum likelihood estimator in a ZIGP regression model. These results remain valid for a GP and ZIP regression model. A Bayesian analysis of the ZIGP model is presented in Angers and Biswas (2003), however they do not incorporate covariates. The zero inflated generalized Poisson distribution, denoted by $ZIGP(p,\theta,\lambda)$, is obtained if the density function of the GP distribution is chosen for $\pi(y|\theta)$. The mean and the variance of the ZIGP distribution are then given by

$$E(Y|p,\mu,\lambda) = (1-p)\mu$$

and

$$\begin{aligned} Var(Y|p,\mu,\lambda) &= p(1-p)\mu^2 + (1-p)\frac{\mu}{(1-\lambda)^2} = (1-p)\mu\Big[p\mu + \frac{1}{(1-\lambda)^2}\Big] \\ &= E(Y|p,\mu,\lambda)\Big[p\mu + \frac{1}{(1-\lambda)^2}\Big]. \end{aligned}$$

The dispersion factor of the ZIGP model is therefore given by $\varphi := p\mu + \frac{1}{(1-\lambda)^2}$. Here, overdispersion can both result from the overdispersion parameter λ of the GP distribution and the extra proportion of zeros p when $p > 0$.

Zero Inflated Regression Models

In a regression model $Y_i \sim ZIP(p_i,\mu_i)$ and $Y_i \sim ZIGP(p_i,\theta_i,\lambda)$, independent for $i = 1,..,n$, respectively, a regression can be performed both for $\mathbf{p} = (p_1,..,p_n)'$ and for $\boldsymbol{\mu} = (\mu_1,..,\mu_n)'$. As in the NB and GP model the parameter μ_i is assumed to depend on covariates $\mathbf{x}_i$ and unknown regression parameters $\boldsymbol{\beta}$. For the proportion of extra zeros a logistic link might be chosen, i.e. $p_i = \frac{\exp(\tilde{\mathbf{x}}_\mathbf{i}'\boldsymbol{\alpha})}{1+\exp(\tilde{\mathbf{x}}_\mathbf{i}'\boldsymbol{\alpha})}$ with covariate vector $\tilde{\mathbf{x}}_\mathbf{i}$ and regression parameters $\boldsymbol{\alpha}$. However, for the data considered in our application no significant zero inflation is detected. An extension to ZI models with regression on p therefore seems

unnecessary and is not addressed further in this thesis.
Alternatively to ZI models hurdle models, see for example Winkelmann (2003) for an overview, could be used. The most widely used hurdle model is the zero hurdle model which assumes two separate models for zero and non-zero observations. Attention is however restricted to zero inflated models in this thesis.

5.2 Regression set up and prior specifications

In order to account for parameter uncertainty and to allow for an underlying spatial structure we consider the count data regression models discussed in Section 5.1 in a Bayesian context. MCMC will be used for parameter estimation. Assume the response variables Y_i, $i = 1, .., n$ to be observed at J regions. Besides the Poisson regression model $Poi(\mu_i)$ we consider the $NB(r, \mu_i), GP(\mu_i, \lambda), ZIP(p, \mu_i)$ and $ZIGP(p, \mu_i, \lambda)$ regression models. In each of these models the parameter μ_i, $i = 1, .., n$ is specified by

$$\mu_i = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_{R(i)}), \tag{5.1}$$

where $\mathbf{x}_i = (1, x_{i1}, .., x_{ik})$ denotes the vector of covariates and t_i gives the observation specific exposure which will be treated as an offset. The vector $\boldsymbol{\beta} = (\beta_0, .., \beta_k)$ denotes the vector of unknown regression parameters. Note, that an intercept β_0 is included in the model. To allow for geographical differences in the J regions spatial random effects $\boldsymbol{\gamma} = (\gamma_1, .., \gamma_J)$ are introduced, $R(i) \in \{1, .., J\}$ denotes the region of the i-th observation. For the zero inflated models we assume a constant p for all observations. Parameter estimation in the regression set up is done using a Bayesian approach with the prior specifications given in the following. The parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, λ, p and r respectively are taken to be a priori independent. Since we have little prior knowledge on the regression parameters $\boldsymbol{\beta}$, we assume a noninformative normal prior, in particular

$$\pi(\boldsymbol{\beta}) \sim N_{k+1}(0, \tau^2 I_{k+1})$$

with $\tau^2 = 100$. The spatial effects $\boldsymbol{\gamma}$ are modelled using the CAR model described in Section 4, i.e.

$$\pi(\boldsymbol{\gamma}|\sigma^2, \psi) \sim N(\mathbf{0}, \sigma^2 Q^{-1})$$

with Q specified as in (4.3). For the spatial hyperparameters σ^2 and ψ also proper priors are assumed, in particular we choose for σ^2 the noninformative prior

$$\sigma^2 \sim IGamma(a, b) \quad \text{with} \quad a = 1 \quad \text{and} \quad b = 0.005$$

which is a common parameter choice for vague gamma priors and for ψ the prior $\psi \sim \frac{1}{(1+\psi)^2}$ which is concentrated on small values for ψ close to 0. For the model specific parameters the following noninformative prior distributions are chosen:

- GP Regression: $\pi(\lambda) \sim U([0,1])$
- NB Regression : $\pi(r) \sim Gamma(a,b)$, i.e. $\pi(r) = \frac{b^a}{\Gamma(a)} r^{a-1} e^{-rb}$, where $a = 1$ and $\pi(b) \sim Gamma(c,d)$, i.e. $\pi(b) \propto b^{c-1} e^{-bd}$ with $c = 1$ and $d = 0.005$
- ZIP/ZIGP Regression: $\pi(p) \sim U([0,1])$

The schemes of the MCMC algorithms and details about the chosen proposal distributions for Metropolis Hastings steps can be found in the Appendix.

5.3 Application

5.3.1 Data description

In this section the proposed models will be used to analyse the number of invasive meningococcal disease cases reported in Germany during the year 2004. Meningococcal disease is caused by bacteria and can lead to serious, perilous diseases, like for example meningitis, in which case we refer to invasive meningococcal disease. In 2004, 600 cases of invasive meningococcal disease were reported in Germany. Germany is divided into 439 regions, for each of these regions the number of invasive meningococcal disease cases is given for both men and women. A histogram of the total number of cases in each region is given in Figure 5.6. A high proportion (67.2 %) of the data is equal to zero, on average 1.37 cases of meningococcal disease are observed in each region, the maximum number of cases observed in one region is 18. The variance of the data is 3.71 which is substantially higher than the mean and therefore already indicates the presence of overdispersion in the data. On a higher aggregation level Germany consists in 16 states. Besides the modelling of overdispersion in the data, an interesting issue is to detect whether there are areas with an increased risk of contracting invasive meningococcal disease. In this case, vaccination could be strongly recommended in these risky regions. Therefore we include the 16 states as covariates in our model, which will be modelled as factor covariates with state 1 as reference level. Since we are interested in relative risks, population effects are eliminated by including the expected number of cases in each region as an offset in the analysis. The expected number of cases in each region is determined by the population in each

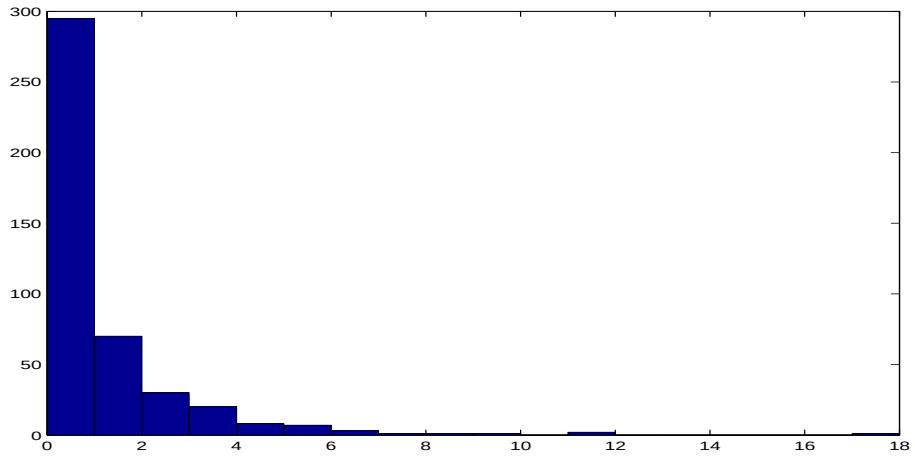

Figure 5.6: Histogram of the observations $y_i, i = 1, .., 439$.

region times the overall observed risk, i.e. the total number of cases divided by the total population.

Extra heterogeneity in the data, which might not be satisfactorily explained by the gender or the state factors, can be handled by the model specific dispersion parameters in the NB, the GP and the ZI models. While overdispersion in the NB and the GP model can be interpreted as unobserved heterogeneity among observations, zero inflated models would assume that part of the observations equal to zero are extra zeros, i.e. in some regions the occurrence of invasive meningococcal disease might not have been reported.

On the other hand, heterogeneity in the data might also be taken into account by assuming a finer geographic resolution, i.e. by including a random spatial effect for each region. We will assume the CAR prior presented in Section 4 for these spatial effects which allows for a spatial dependency structure. In contrast to this approach, the effects of the states included as factor covariates can be seen as unstructured effects on a lower resolution, since no correlation between states is allowed.

We first analysed the data set in Splus using a Poisson model without spatial effects including an intercept and as covariates the gender and the 16 states as factor covariates. No significant influence of gender could be detected, therefore we decided to model the total number of cases without distinguishing between men and women. This means, that we have only one observation $y_i, i = 1, .., 439$ for each region. The four states Nordrhein-Westfalen, Mecklenburg-Vorpommern, Sachsen-Anhalt and Thüringen were found to be significant and only the regression indicators of these states will be included in the following. This model was used as an initial model for the MCMC algorithms of the models discussed in Section 5.1.

5.3.2 Models

The MCMC algorithms for the Poisson, the GP, the NB, the ZIP and the ZIGP regression models are run for 20000 iterations. The parameter $\mu_i,\ i = 1, .., n$ is specified for all models by

$$\mu_i = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_i)$$

with the same covariates included in each model and $t_i = pop_i \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n pop_i}$ where pop_i denotes the population in region i. All models are fitted with and without spatial effects. Since we have only one observation for each region we use the simplified notation $\gamma_{R(i)} = \gamma_i$. The first 5000 iterations of the MCMC samplers are discarded as burnin, convergence is achieved well before for all models. After convergence the mixing of the samplers is satisfactorily good, the estimated empirical autocorrelations with lag 5 are in general well below 0.05 for the regression parameters in all non spatial models and both the regression parameters and the spatial effects in the spatial GP and ZIGP models. Only in the spatial Poisson, ZIP and NB models a lag of about 20 is needed in order to obtain autocorrelations of the regression parameters below 0.05, for the spatial effects the autocorrelations are below 0.05 at a lag of 5 in the Poisson and ZIP models and a lag of about 10 in the NB model. The estimated posterior means and 90 % credible intervals for the regression parameters are reported in Table 5.1 for all models. Estimation of the intercept slightly differs between the models and also changes when spatial effects are added, especially for the Poisson and ZIP models where large spatial effects are observed, see below. Estimation of the state effects is rather similar in all models.

For a comparison of the estimated overdispersion in the different models, we consider the estimated dispersion factors φ_i which are defined by $1 + \frac{\mu_i}{r}$, $\frac{1}{(1-\lambda)^2}$, $(p\mu_i + 1)$ and $p\mu_i + \frac{1}{(1-\lambda)^2}$ for the NB, GP, ZIP and ZIGP regression models, respectively. In particular, we compute the mean, minimum, maximum value and quantiles of the estimated posterior means $\hat{\varphi}_i := \frac{1}{R}\sum_{j=1}^R \hat{\varphi}_i^j$ of the dispersion factors in each model, where $\hat{\varphi}_i^j$ denotes the j-th MCMC iterate for φ_i after burnin. The results are reported in Table 5.2. Note, that the dispersion factor in the GP regression model is the same for all observations, whereas it depends on the parameter μ_i and therefore is different for each observation in the other models. Except for the ZIP model, all models exhibit a substantial degree of overdispersion with respect to the Poisson model, regardless whether spatial effects are included or not. In the non spatial NB model the average of the estimated posterior

Model	γ	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Poi	no	-0.17	0.43	0.65	0.56	0.56
		$(-0.26, -0.09)$	$(0.30, 0.57)$	$(0.28, 1.02)$	$(0.23, 0.87)$	$(0.22, 0.88)$
NB	no	-0.15	0.42	0.67	0.54	0.48
		$(-0.25, -0.04)$	$(0.23, 0.60)$	$(0.23, 1.10)$	$(0.16, 0.91)$	$(0.09, 0.85)$
GP	no	-0.16	0.42	0.65	0.45	0.50
		$(-0.26, -0.06)$	$(0.26, 0.58)$	$(0.22, 1.06)$	$(0.05, 0.84)$	$(0.09, 0.88)$
ZIP	no	-0.13	0.43	0.69	0.60	0.59
		$(-0.23, -0.03)$	$(0.27, 0.59)$	$(0.29, 1.08)$	$(0.25, 0.94)$	$(0.24, 0.95)$
ZIGP	no	-0.15	0.42	0.67	0.47	0.52
		$(-0.26, -0.04)$	$(0.23, 0.61)$	$(0.22, 1.09)$	$(0.05, 0.87)$	$(0.09, 0.92)$
Poi	yes	-0.25	0.40	0.68	0.58	0.47
		$(-0.38, -0.14)$	$(0.18, 0.61)$	$(0.21, 1.13)$	$(0.19, 0.97)$	$(0.06, 0.86)$
NB	yes	-0.18	0.41	0.68	0.55	0.48
		$(-0.29, -0.06)$	$(0.20, 0.61)$	$(0.23, 1.21)$	$(0.17, 0.93)$	$(0.08, 0.87)$
GP	yes	-0.16	0.42	0.65	0.45	0.50
		$(-0.26, -0.06)$	$(0.25, 0.58)$	$(0.21, 1.06)$	$(0.04, 0.84)$	$(0.09, 0.88)$
ZIP	yes	-0.23	0.39	0.68	0.60	0.49
		$(-0.36, -0.10)$	$(0.16, 0.63)$	$(0.22, 1.14)$	$(0.19, 0.99)$	$(0.07, 0.90)$
ZIGP	yes	-0.15	0.42	0.66	0.48	0.52
		$(-0.26, -0.04)$	$(0.23, 0.61)$	$(0.22, 1.09)$	$(0.06, 0.88)$	$(0.10, 0.92)$

Table 5.1: Posterior means and 90 % credible intervals for the regression parameters ($\beta_1, .., \beta_4$: effects for the states Nordrhein-Westfalen, Mecklenburg-Vorpommern, Sachsen-Anhalt and Thüringen, respectively) in the different models for the meningococcal disease data.

means of the dispersion parameter is given by 1.396 and drops to 1.293 when spatial effects are included. The range of the estimated spatial effects in the NB model, see Table 5.3, is considerably smaller than in the Poisson model where unexplained heterogeneity in the data is captured by the spatial effects alone. However, part of the data variability in the NB model is explained by spatial effects as well rather than the parameter r alone. This is in contrast to the GP and ZIGP model, where the estimated spatial effects are all very close to zero. Overdispersion in these models is captured by the parameter λ only, resulting in a high estimated dispersion parameter. Results are hardly affected by the inclusion of spatial effects. The extension from a GP to a ZIGP model has almost no

Parameter	γ	mean (2.5 %, 97.5 %)	$\hat{\varphi}_i$					
			mean	min	25%	50%	75%	max
r in NB	yes	6.433 (2.483, 22.552)	1.293	1.051	1.150	1.205	1.313	5.358
	no	3.836 (2.202, 6.864)	1.396	1.064	1.200	1.273	1.428	7.053
λ in GP	yes	0.162 (0.098, 0.231)	1.432					
	no	0.163 (0.098, 0.232)	1.435					
p in ZIP	yes	0.029 (0.001, 0.087)	1.041	1.007	1.020	1.028	1.044	1.536
	no	0.056 (0.006, 0.125)	1.081	1.013	1.041	1.056	1.089	2.226
p in ZIGP	yes	0.019 (0.001, 0.064)	1.434	1.412	1.421	1.426	1.436	1.819
λ in ZIGP	yes	0.155 (0.088, 0.224)						
p in ZIGP	no	0.019 (0.001, 0.063)	1.440	1.417	1.427	1.432	1.442	1.826
λ in ZIGP	no	0.157 (0.090, 0.227)						

Table 5.2: Estimated posterior means for the model specific dispersion parameters in the considered models with and without spatial effects, with the 2.5 % and 97.5 % quantiles given in brackets. Further the mean, range and quantiles of the estimated posterior means of the dispersion factors $\hat{\varphi}_i$ are given.

influence on the estimation of λ and the average dispersion parameter φ_i, the proportion of extra zeros p is estimated very close to zero.
In the non spatial ZIP model the proportion of extras zeros p is estimated as 5.6 %, resulting in an average dispersion parameter of about 1.081. According to the large 95 % credible interval for p however, no significant degree of zero inflation seems to be present. Unobserved heterogeneity still present in the data after adjusting for covariates is captured better by the GP and NB model, whereas the assumptions of extra zeros is obviously not appropriate for this data. When spatial effects are included to the ZIP model the estimated proportion of extra zeros drops even further, indicating that unexplained heterogeneity is picked up mostly by the spatial effects alone like in the Poisson model, the range of the estimated posterior means of the spatial effects in the ZIP and the Poisson model is almost the same, see Table The map plot of the estimated posterior means of the spatial effects in the Poisson model, given in Figure 5.7 roughly represents the spatial pattern of the observed relative risk in each region $\frac{y_i}{t_i}, i = 1, .., n$ which is plotted in the left panel in Figure 5.8. The estimated posterior mean and median of the spatial hyperparameter ψ in the Poisson model, see Table 5.3 are rather small, the lower bound of the 95 % credible interval is close to zero, indicating that the overall degree of spatial dependence

Model	$[\min_j \hat{\gamma}_j \max_j \hat{\gamma}_j]$	$\hat{\psi}$			$\hat{\sigma}^2$		
		mean	median	95 % CI	mean	median	95 % CI
Poisson	[−0.383, 1.059]	0.394	0.207	(0.011,0.431)	0.541	0.422	(0.149, 0.615)
NB	[−0.151, 0.345]	1.223	0.588	(0.015,1.352)	0.227	0.140	(0.012,0.281)
GP	[−0.009, 0.020]	2.604	1.194	(0.038,2.199)	0.012	0.005	(0.001,0.011)
ZIP	[−0.371, 0.998]	0.329	0.235	(0.013,0.397)	0.477	0.417	(0.169,0.529)
ZIGP	[−0.011, 0.019]	1.739	1.154	(0.044,1.905)	0.018	0.011	(0.001,0.027)

Table 5.3: Range of estimated posterior means of the spatial effects as well as estimated posterior means, medians and 95 % credible intervals for the spatial hyperparameters in the considered models for the meningococcal disease data.

is very small. This is reflected in the estimated spatial pattern which is not particularly smooth. Only some rather isolated regions, which are marked in black in the right map in Figure 5.7, have a significant positive spatial effect according to the 80 % credible intervals. In these regions the observed number of invasive meningococcal disease cases was rather high and most of them do not lie within the four states included as covariates. Therefore without spatial effects the risk in these regions is not modelled sufficiently. The estimated posterior means of the risk factor $\frac{\mu_i}{t_i}$ are plotted for the non-spatial and spatial Poisson model in the middle and right panel. In the non-spatial Poisson model geographic differences are modelled by four state indicators only. Since the risk in two of the states, which are neighbours, is about the same, visually only three states can be distinguished in this plot. The inclusion of spatial random effects gives a rather smoothed representation of the true pattern, however the rough structure is detected reasonable well.

5.3.3 Model comparison using DIC

In order to compare the presented models the DIC, reviewed in Section 3.2.1, is considered. In Table 5.4, the DIC, the posterior mean of the deviance and the effective number of parameters are given for each model. Only in the Poisson regression case a well defined normalizing constant $f(y)$ (see Section 3.2.1) exists, while in all other models the likelihood of the saturated model depends on the unknown overdispersion parameters. Therefore we make the choice of setting the normalizing function $f(y)$ to 0. Consequently $E[D(\boldsymbol{\theta}|\mathbf{y})]$ is based only on the unscaled deviance which cannot be directly interpreted as an overall goodness of fit measure of one specific model. However, $E[D(\boldsymbol{\theta}|\mathbf{y})]$ can be used for comparing the model fit of several models when the number of parameters is roughly

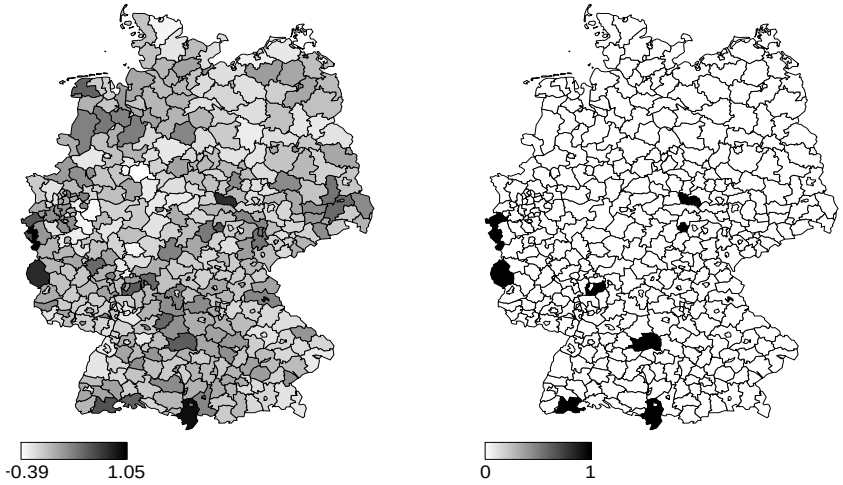

Figure 5.7: Maps of the estimated posterior means (left) and 80 % credible intervals (white: 0 included in credible interval, black: strictly negative credible interval) of the spatial effects in the Poisson regression model for the meningococcal disease data.

the same.

Model	γ	DIC	$E[D(\boldsymbol{\theta}\|\mathbf{y})]$	p_D
Poisson	no	1291.8	1286.8	5.04
NB	no	1273.9	1267.8	6.10
GP	no	1265.6	1259.6	6.01
ZIP	no	1291.8	1285.9	5.96
ZIGP	no	1267.8	1261.5	6.35
Poisson	yes	1248.7	1159.1	89.56
NB	yes	1270.8	1240.0	30.74
GP	yes	1265.7	1258.3	7.32
ZIP	yes	1255.4	1175.1	80.31
ZIGP	yes	1267.6	1260.2	8.28

Table 5.4: DIC, $E[D(\boldsymbol{\theta}|\mathbf{y})]$ and effective number of parameters p_D for the different models.

For the non spatial models the lowest value of the DIC is obtained for the GP model, while the DIC for the Poisson and the ZIP model takes the highest value. Hence, according to the DIC the GP model is considered best among the non spatial models, while the Poisson

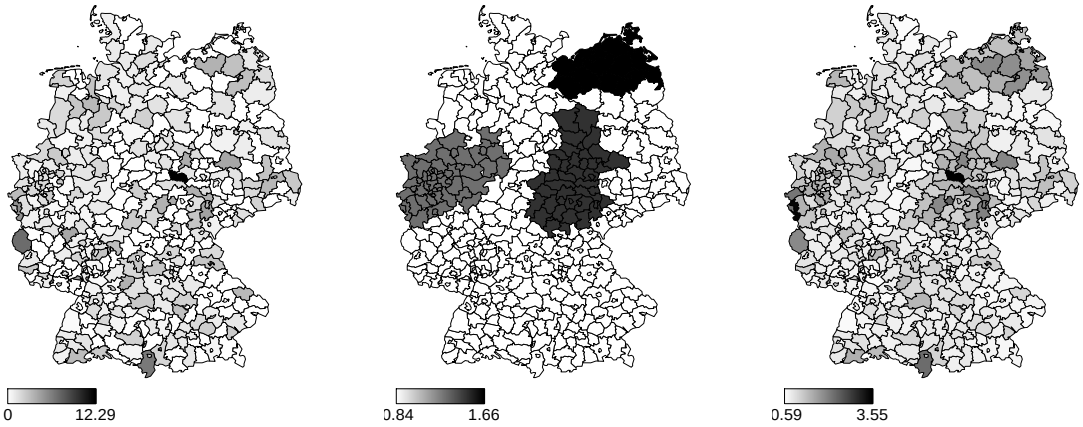

Figure 5.8: Maps of the observed risk $\frac{y_i}{t_i}$ (left) and the estimated posterior means of the relative risk factor $\frac{\mu_i}{t_i}$ in the non-spatial (middle) and spatial (right) Poisson regression model for the meningococcal disease data.

and ZIP model clearly perform worse. The effective number of parameters p_D is close to the true number of parameters which is five for the Poisson regression model, six for the NB, GP and ZIP regression models and seven for the ZIGP regression model.
When spatial effects are added, the posterior mean of the deviance and the number of effective parameters in the GP and ZIGP models hardly change. As mentioned in the previous section already, spatial effects are not significant in these models, i.e. after adjusting for covariate information, there is no further spatial heterogeneity in the data which might be captured by the spatial effects. Instead any overdispersion present in the data seems to be sufficiently captured by the model specific dispersion parameter. The DIC for the spatial NB model is slightly better than for the non spatial one, hence spatial effects improve the model. However, the spatial pattern is rather smooth as can be seen from the effective number of parameters estimated by 30.74. For the Poisson and ZIP regression model in contrast, a significant drop in the DIC is observed when spatial effects are taken into account. This shows that there is some extra variability in the data which is not sufficiently explained by the covariates only in these models. Since the Poisson model does not allow for overdispersion and the heterogeneity is not of a zero inflated nature, for these two models the unexplained variability is covered by the spatial effects. According to the DIC the spatial Poisson model gives the best fit and is to be preferred to a non spatial GP model. Note, that the DIC must be used with care here, since strictly speaking the DIC is defined for distributions of the exponential family only, see van der Linde (2005). However, the posterior mean of the deviance $E[D(\boldsymbol{\theta}|\mathbf{y})]$

which can be considered for comparing the model fit of the non spatial models where the number of parameters is very close, gives the same ranking of the models as the DIC.

5.3.4 Model checking using proper scoring rules

Apart from the DIC we also compute the Brier score and the logarithmic score presented in Section 3.2.3 for each model, results are reported in Table 5.5. These scores are based both on the posterior predictive probabilities and the true observed number of cases and therefore provide a good measure for checking which model fits the data best. The results

Model	γ	Brier score	logarithmic score
Poisson	no	-0.6937	-1.4569
NB	no	-0.6883	-1.4363
GP	no	-0.6873	-1.4272
ZIP	no	-0.6921	-1.4549
ZIGP	no	-0.6878	-1.4291
Poisson	yes	-0.6280	-1.2422
NB	yes	-0.6717	-1.3779
GP	yes	-0.6863	-1.4243
ZIP	yes	-0.6481	-1.3251
ZIGP	yes	-0.6900	-1.4529

Table 5.5: Brier score and logarithmic score for the considered models with and without spatial effects.

support the conclusions drawn in the previous section. For the non spatial models the GP regression model fits the heterogeneity in the data best, followed by the ZIGP and NB regression model. The use of a non spatial ZIP regression model does not seem to be appropriate, the gain in comparison to the non spatial Poisson model for which the lowest scores are obtained is very small. The scores for the GP and the ZIGP model hardly change by allowing for spatial effects, indicating that the model specific dispersion parameters capture the data heterogeneity well. Again a small improvement in the NB model is observed when spatial effects are included. The scores for the spatial Poisson and ZIP model however, are considerably smaller than for the other models. This confirms again, that spatial effects have a significant influence in these models and that a spatial Poisson model gives the best fit to the data.

5.4 Conclusions

We have presented several regression models for count data allowing for overdispersion. Overdispersion is either modelled by the introduction of an additional parameter as in the NB and GP model, by allowing for an extra proportion of zero observations using zero inflated models or by combining zero inflated models with overdispersed distributions.
Further, additionally spatial random effects are included in the models in order to account for unobserved spatial heterogeneity in the data. This approach allows for spatial correlations between observations.
These models were applied to analyse the number of invasive meningococcal disease cases in Germany in the year 2004. The DIC, the Brier and the logarithmic score were used for model comparison. The models allowing for overdispersion gave a significantly better fit than an ordinary non spatial Poisson regression model. Among these non spatial models, the GP model fitted the data best, while the overdispersion present did not seem to be caused by the presence of extra zeros in the data. For the GP and the ZIGP model the inclusion of spatial effects did not improve the models, in the NB model still some significant spatial variation was detected. For the Poisson model which does not allow for overdispersion and the ZIP model which is not modelling the nature of the overdispersion appropriately, the inclusion of spatial effects led to a significant improvement. According to the considered criterions the spatial Poisson model is to be preferred to all other models. But we would like to note that the spatial model fitted shows no smooth surface structure, it rather indicates isolated specific regions where the covariates provide no adequate fit.
Instead of analysing the number of cases of invasive meningococcal disease for one year only, it might be interesting to include data over several years in the analysis. Space-time interactions could be included in order to examine whether the spatial pattern changed over the years.

5.5 Algorithmic schemes

Details on the MCMC algorithms for the presented models are given in the Appendix B. In this section the algorithmic schemes we finally used are summarized. Most update steps are performed using a single component Metropolis Hastings (MH) step. For the proposal distributions either a symmetric random walk proposal or an independence proposal is used. In particular, for the independence proposal we take a t-distribution with $v = 20$

degrees of freedom with the same mode and the same inverse curvature at the mode as the target distribution, see Section 2.2.2.

5.5.1 GP regression model

- Sample $\lambda|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}$
- Sample $\beta_j|\mathbf{y}, \lambda, \boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}, \qquad j = 0, .., k$
- Update of spatial effects
  - Sample $\frac{1}{\sigma^2}|\boldsymbol{\gamma}, \psi \sim Gamma$
  - Sample $\psi|\boldsymbol{\gamma}, \sigma$
  - Sample $\gamma_j|\mathbf{y}, \lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}_{-j}, \psi, \sigma, \qquad j = 1, .., J$

Since the full conditional of σ^2 is Inverse Gamma, σ^2 can be sampled directly using a Gibbs step. For the remaining parameters a MH step is used. In particular, λ, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are updated component by component using an independence proposal distribution. The spatial hyperparameter ψ is updated using a random walk proposal. For the Poisson regression model the algorithmic scheme is the same, but with λ set fix to 0.

5.5.2 NB regression model

- Sample $r|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}$
- Sample $\beta_j|\mathbf{y}, r, \boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}, \qquad j = 0, .., k$
- Update of spatial effects
  - sample spatial hyperparameters $\frac{1}{\sigma^2}$ and ψ as in 5.5.1
  - Sample $\gamma_j|\mathbf{y}, r, \boldsymbol{\beta}, \boldsymbol{\gamma}_{-j}, \psi, \sigma, \qquad j = 1, .., J$

In the NB regression model $r, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are updated component by component using a MH step. For r a random walk proposal is used, while $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are updated using an independence sampler.

5.5.3 ZI models

To avoid convergence problems in the ZI models which arose in simulated data due to correlation between the intercept β_0, p and λ, we use collapsed algorithms, in particular β_0, p and λ are updated with the latent variables $\mathbf{z}$ integrated out, i.e. based on model (5.6). Doing so convergence and mixing of the samplers was improved a lot.

ZIP model with constant p

- Updates with $\mathbf{z}$ integrated out
  - Sample $\beta_0|\mathbf{y}, p, \boldsymbol{\beta}_{-0}, \boldsymbol{\gamma}$
  - Sample $p|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}$
- Sample $z_i|\mathbf{y}, p, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim Bernoulli\Big(\frac{p}{p + (1-p)\exp(-\mu_i)}\Big) \quad \forall i \text{ with } y_i = 0$
- Sample $\beta_j|\mathbf{y}, \boldsymbol{\beta}_{-j}, \mathbf{z}, \boldsymbol{\gamma}, \qquad j = 1, .., k$
- Update of spatial effects
  - sample spatial hyperparameters $\frac{1}{\sigma^2}$ and ψ as in 5.5.1
  - Sample $\gamma_j|\mathbf{y}, \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\gamma}_{-j}, \psi, \sigma, \qquad j = 1, .., J$

The latent variables $\mathbf{z}$ can be updated using a Gibbs step. Since the full conditional of p is log concave, adaptive rejection sampling (ARS) introduced by Gilks and Wild (1992) is used to update p. For the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ a MH step using an independence proposal distribution is performed.

ZIGP model with constant p

- Updates with $\mathbf{z}$ integrated out
  - Sample $\beta_0|\mathbf{y}, p, \lambda, \boldsymbol{\beta}_{-0}, \boldsymbol{\gamma}$
  - Sample $p|\mathbf{y}, \lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}$
  - Sample $\lambda|\mathbf{y}, p, \boldsymbol{\beta}, \boldsymbol{\gamma}$
- Sample $z_i|\mathbf{y}, \lambda, p, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim Bernoulli\Big(\frac{p}{p + (1-p)\exp(-\mu_i)}\Big) \quad \forall i \text{ with } y_i > 0$
- Sample $\beta_j|\mathbf{y}, \lambda, \boldsymbol{\beta}_{-j}, \mathbf{z}, \boldsymbol{\gamma}, \qquad j = 1, .., k$

- Update of spatial effects
  - sample spatial hyperparameters $\frac{1}{\sigma^2}$ and ψ as in 5.5.1
  - Sample $\gamma_j|\mathbf{y}, \lambda, \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\gamma}_{-j}, \psi, \sigma, \qquad j = 1, .., J$

For the ZIGP model the same proposal distributions as in the ZIP model are used. For λ an independence proposal is taken.

Chapter 6

A Gibbs sampler for spatial Poisson regression models

In this chapter we present a straightforward Gibbs sampler for spatial Poisson regression models using data augmentation techniques. In particular, we aim to investigate whether this Gibbs sampler is found to be superior to a conventional single site Metropolis Hastings (MH) sampler. The issue of model parameterisation and several update schemes for the parameters in the Gibbs sampler is thoroughly addressed. The performance of the developed Gibbs sampler schemes and the MH sampler is investigated in two simulation studies as well as on real data from a German car insurance company. Performance of the samplers is measured in the computational costs required to obtain the same precision of the posterior means of the parameters.

Since the full conditional distributions of a spatial Poisson regression model do not follow any standard distribution, often single site MH steps are performed in a MCMC setting, see for example Diggle et al. (1998), Dimakos and Frigessi (2002) or Chapter 5 in this thesis. However, this requires the choice of appropriate proposal distributions in order to achieve reasonable acceptance rates and a good mixing of the MCMC chains. Advanced independence proposals, like for example a normal proposal with the same mode and inverse curvature at the mode as the target distribution which have been used for the models presented in the previous chapter, can lead to high acceptance rates and low autocorrelations but involve considerable computational efforts.

Frühwirth-Schnatter and Wagner (2004a) developed a Gibbs sampler for Poisson regression models for small counts. They show that by data augmentation via the introduction of two sequences of latent variables a linear normal model is obtained. In Frühwirth-Schnatter and Wagner (2004b) an application of this Gibbs sampler to state space models

is given, in Frühwirth-Schnatter and Wagner (2004a) the same methodology is applied for standard Poisson regression models and Poisson regression models with overdispersion. Using similar techniques, a Gibbs sampler for logistic models is developed in Frühwirth-Schnatter and Waldl (2004).

The aim of this chapter is to show that this methodology can be extended to spatial Poisson regression models in a straightforward manner allowing for a Gibbs update of both regression parameters and spatial effects. Although we only consider spatial Poisson data distributed on regions in this thesis, the presented methodology could also be applied on geostatistical Poisson models, see Diggle et al. (1998).

It is well known, that mixing and convergence of the Gibbs sampler depends crucially on several implementation issues, see for example Roberts and Sahu (1997) for a detailed discussion. High autocorrelations can be reduced by updating several parameters in one block or using collapsed algorithms (see Section 2.3), another important issue is model parameterisation. Gelfand et al. (1995) discuss the efficiency of centered and non-centered parameterisations for hierarchical normal linear models, Papaspiliopoulos et al. (2003) address parameterisation issues for several classes of hierarchical models and introduce partially non-centered parameterisations. Christensen et al. (2005) propose the standardization and orthogonalization of all model components leading to efficient and robust MCMC algorithms.

In this chapter both centered and non-centered model parameterisations are considered, various algorithmic schemes, in particular a joint block update of the intercept and the spatial effects as well as collapsed algorithms, are discussed. The performance of the samplers is examined and compared to a single site MH sampler with independence proposals in two simulation studies. In the first study, the samplers are applied on data with both large and small spatial effects, while the second study considers the influence of the data heterogeneity on the performance of the samplers. The performance of the samplers is measured in the computational costs required in order to obtain a certain precision of the posterior means of the regression parameters and spatial effects. This is done by taking both the Monte Carlo error of the posterior means of the parameters and the computational time required for one iteration into account. A very similar approach for comparing the performance of MCMC samplers is conducted by Christensen and Waagepetersen (2002). Among the Gibbs samplers collapsed algorithms perform best. In particular for data with small spatial effects, the Monte Carlo errors of the spatial effects are considerably reduced when collapsed samplers and model parameterisations with non-centered scale or variance are used. The Monte Carlo errors of the regression parameters however

are rather high, especially for data with low heterogeneity. The MH independence sampler in contrast, exhibits very low Monte Carlo errors and good mixing for both regression and spatial effects in all settings. Although the MH sampler requires a higher computational effort, this drawback is compensated by the high precision of the posterior means of the parameters. In order to obtain the same precision the Gibbs samplers would have to be run considerably longer, diminishing the computational advantage in comparison to the MH sampler. Therefore we have to conclude that the proposed Gibbs sampler for spatial Poisson regression models can not outperform a single site MH sampler using independence proposals.

This chapter is organized as follows. In Section 6.1 the spatial Poisson regression model is specified and the two steps of the data augmentation scheme are described for this specific model. Details on several algorithmic schemes for updating the regression and spatial effects are given in Section 6.2. In Section 6.3 the developed Gibbs sampler schemes are examined and compared to a single component MH sampler with independence proposals in two simulation studys. We also apply the Gibbs samplers to model the expected number of claims in a real data set from a German car insurance company. Section 6.4 gives a summary and draws conclusions.

6.1 Data augmentation and Gibbs sampler for spatial Poisson regression models

We assume that observations Y_i, $i = 1, .., n$ observed at J regions follow a Poisson model

$$y_i \sim Poisson(\mu_i). \tag{6.1}$$

The mean μ_i is specified by

$$\mu_i = t_i \exp(\mathbf{z}_i'\boldsymbol{\alpha}) := t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{v}_i'\boldsymbol{\gamma}) = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_{R(i)}) \tag{6.2}$$

where $\mathbf{z}_i' = (\mathbf{x}_i', \mathbf{v}_i')$ denotes the covariate vector $\mathbf{x}_i = (1, x_{i1}, .., x_{ip})'$ and the incidence vector $\mathbf{v}_i = (v_{i1}, .., v_{iJ})'$ for the regions, i.e. $v_{ij} = \begin{cases} 1, & \text{if} \quad R(i) = j \\ 0 & \text{otherwise} \end{cases}$, with $R(i) \in \{1, .., J\}$ denoting the region of the i-th observation. Further $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma})'$ denotes the vector of regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, .., \beta_p)$ and spatial random effects $\boldsymbol{\gamma} = (\gamma_1, ..\gamma_J)$. By the inclusion of spatial effects we allow for geographical differences in the J regions. The quantity t_i gives the exposure time for the i-th observation.

We assume a normal prior distribution centered around zero with a large standard deviation for the regression parameters $\boldsymbol{\beta}$, in particular

$$\boldsymbol{\beta} \sim N_{p+1}(0, V_0)$$

where $V_0 = \tau^2 I_{p+1}$ with $\tau^2 = 100$. Here $N_p(\mu, \Sigma)$ denotes the p-variate Normal distribution with mean μ and covariance matrix Σ. For the spatial effects the conditional autoregressive (CAR) prior

$$\boldsymbol{\gamma}|\psi, \sigma^2 \sim N_J(0, \sigma^2 Q^{-1})$$

is used where the elements of the precision matrix $Q = (Q_{ij})$, $i, j = 1, .., J$ are defined as in (4.3), see Section 4. Therefore, we have a multivariate normal prior distribution for the regression and spatial parameters $\boldsymbol{\alpha}$ which is given by

$$\alpha|\boldsymbol{\theta} \sim N_{p+1+J}(0, \Sigma) \tag{6.3}$$

with $\Sigma = \begin{pmatrix} V_0 & 0 \\ 0 & \sigma^2 Q^{-1} \end{pmatrix}$. For the spatial hyperparameters $\boldsymbol{\theta} = (\psi, \sigma^2)$ the proper prior distributions

$$\psi \sim \frac{1}{(1+\psi)^2} \quad \text{and} \quad \sigma^2 \sim IGamma(1, 0.005)$$

are assumed. The parameterisation of this model described by Observation Equation (6.2) and Prior Specification (6.3) is called non-centered in the mean, since the intercept β_0 appears in the observation equation, but not in the spatial prior formulation. Other possible model parameterisations include parameterisations additionally non-centered in the scale and variance of the spatial prior as well as a centered parameterisation, where the intercept β_0 only appears as the mean of the spatial prior. These parameterisations are summarized in Table 6.1. For a summary on existing parameterisation techniques see for example Frühwirth-Schnatter (2004). Initially, our investigations are based on the non-centered mean parameterisation given by (6.2) and (6.3). Necessary changes when other parameterisations are used will be indicated specifically.

6.1.1 Step 1: Introduction of hidden inter-arrival times

The basic idea of the data augmentation scheme developed by Frühwirth-Schnatter and Wagner (2004b) is to regard the Poisson observations $y_i, i = 1, .., n$, as the number of jumps of an unobserved Poisson process with intensity μ_i within the unit interval. In the

parameterisation	spatial prior	observation equation
centered	$\boldsymbol{\gamma}^c \sim N(\beta_0, \sigma^2 Q^{-1})$	$\mu_i = t_i \exp(\mathbf{x}'_{i-0}\boldsymbol{\beta}_{-0} + \mathbf{v}'_i\boldsymbol{\gamma}^c)$
non-centered mean	$\boldsymbol{\gamma} \sim N(0, \sigma^2 Q^{-1})$	$\mu_i = t_i \exp(\beta_0 + \mathbf{x}'_{i-0}\boldsymbol{\beta}_{-0} + \mathbf{v}'_i\boldsymbol{\gamma})$
non-centered mean and scale	$\boldsymbol{\gamma}^* \sim N(0, Q^{-1})$	$\mu_i = t_i \exp(\beta_0 + \mathbf{x}'_{i-0}\boldsymbol{\beta}_{-0} + \sigma\mathbf{v}'_i\boldsymbol{\gamma}^*)$
non-centered mean and variance	$\boldsymbol{\gamma}^{**} \sim N(0, I)$	$\mu_i = t_i \exp(\beta_0 + \mathbf{x}'_{i-0}\boldsymbol{\beta}_{-0} + \sigma\mathbf{v}'_i L\boldsymbol{\gamma}^{**})$ where $LL' = Q^{-1}$

Table 6.1: Spatial prior and observation equation for different model parameterisations, where $\mathbf{x}_{i-0} := (x_{i1}, .., x_{ip})'$ and $\boldsymbol{\beta}_{-0} := (\beta_1, .., \beta_p)$

first step of the data augmentation, $y_i + 1$ hidden inter-arrival times $\tau_{ij}, j = 1, .., y_i + 1$ are introduced for each observation y_i. From the properties of a Poisson process, see for example Mikosch (2004), it is well known that the inter-arrival times are independent and follow an exponential distribution with parameter μ_i, i.e.

$$\tau_{ij}|\boldsymbol{\alpha} \sim Exponential(\mu_i) = \frac{Exponential(1)}{\mu_i}.$$

Taking the logarithm we obtain

$$\begin{aligned} \log \tau_{ij}|\boldsymbol{\alpha} &= -\log \mu_i + \epsilon_{ij} \\ &= -\log t_i - \mathbf{z}'_i\boldsymbol{\alpha} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \log(Exponential(1)). \end{aligned} \tag{6.4}$$

Denote by $\boldsymbol{\tau} = \{\tau_{ij}, i = 1, .., n, j = 1, .., y_i + 1\}$ the collection of all inter-arrival times. Then the posterior distribution of $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\tau}$

$$p(\boldsymbol{\alpha}|\boldsymbol{\theta}, \mathbf{y}, \boldsymbol{\tau}) = p(\boldsymbol{\alpha}|\boldsymbol{\theta}, \boldsymbol{\tau})$$

is independent of $\mathbf{y}$. Conditional on $\boldsymbol{\tau}$ we are now dealing with model (6.4) which is linear in the parameters $\boldsymbol{\alpha}$, but still has a non-normal error term.

6.1.2 Step 2: Mixture approximation for error term

The second step of the data augmentation scheme eliminates the non-normality of model (6.4). The error term in (6.4) can be approximated by a mixture of K normal distributions with mean m_r, variance s_r^2 and weight $w_r, r = 1, .., K$, i.e.

$$p(\epsilon_{ij}) = \exp(\epsilon_{ij} - \exp(\epsilon_{ij})) \approx \sum_{r=1}^{K} w_r f_N(\epsilon_{ij}; m_r, s_r^2),$$

where $f_N(\cdot; m_r, s_r^2)$ denotes the density of the normal distribution with mean m_r and variance s_r^2. Frühwirth-Schnatter and Wagner (2004b) show that $K = 5$ is sufficient to obtain a close approximation to the normal distribution. They also give the corresponding values for m_r, s_r^2 and w_r. The second step of the data augmentation then consists in the introduction of the component indicators $r_{ij} \in \{1, .., 5\}$ as latent variables. We denote the set of all component indicators by $\mathbf{R} = \{r_{ij}, i = 1, .., n, j = 1, .., y_i + 1\}$. Conditional on $\mathbf{R}$ we have

$$\log \tau_{ij} | \boldsymbol{\alpha}, r_{ij} = -\log t_i - \mathbf{z}_i' \boldsymbol{\alpha} + m_{r_{ij}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, s_{r_{ij}}^2),$$

i.e.

$$(\log \tau_{ij} + \log t_i - m_{r_{ij}}) | \boldsymbol{\alpha}, r_{ij} \sim N(-\mathbf{z}_i' \boldsymbol{\alpha}, s_{r_{ij}}^2). \tag{6.5}$$

Therefore we are dealing with a normal model which is linear in $\boldsymbol{\alpha}$ now. The posterior distribution of $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\tau}$ and $\mathbf{R}$ is given by

$$p(\boldsymbol{\alpha} | \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}) \propto \pi(\boldsymbol{\alpha} | \boldsymbol{\theta}) \prod_{i=1}^{n} \prod_{j=1}^{y_i+1} \frac{1}{s_{r_{ij}}} \exp\Big[-\frac{1}{2 s_{r_{ij}}^2} (\log \tau_{ij} + \log t_i - m_{r_{ij}} + \mathbf{z}_i' \boldsymbol{\alpha})^2 \Big].$$

Since the prior distribution $\pi(\boldsymbol{\alpha} | \boldsymbol{\theta})$ is normal as well, the resulting posterior distribution is multivariate normal and a Gibbs sampler can be applied. Note, that by performing this data augmentation we are no longer dealing with n but with $\sum_{i=1}^{n} (y_i + 1)$ observations. Therefore this Gibbs Sampler is mainly useful for count data with small counts only, otherwise the data set might get very large.

6.1.3 Algorithmic scheme

The algorithmic scheme for the above Gibbs Sampler is the following:
Choose appropriate starting values for the component indicators $\mathbf{R}$ and the inter-arrival times $\boldsymbol{\tau}$.

(1) sample regression and spatial parameters $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma})'$ given $\boldsymbol{\tau}, \mathbf{R}, \boldsymbol{\theta}$

(2) sample spatial hyperparameters $\boldsymbol{\theta} = (\psi, \sigma^2)$ given $\boldsymbol{\alpha}$

(3) sample the inter-arrival times τ_{ij} given $\boldsymbol{\alpha}, \mathbf{y}$

(4) sample the component indicators r_{ij} given $\boldsymbol{\tau}, \boldsymbol{\alpha}$

Step (1) consists of sampling from a multivariate normal distribution. This can be done in one block, however it might be computationally more efficient to perform an update in several smaller blocks. We will consider several update strategies for step (1) later in more detail. The spatial hyperparameter ψ is updated using a MH step, whereas σ^2 can be updated using a Gibbs step. Steps (3) and (4), elaborated in Frühwirth-Schnatter and Wagner (2004b), are described in the following sections.

6.1.4 Sampling the inter-arrival times

Given $\mathbf{y}$ and $\boldsymbol{\alpha}$, the inter-arrival times for different observations $i = 1, .., n$ are independent. For fixed i however, $\tau_{i1}, .., \tau_{i,y_i+1}$ are stochastically dependent, but independent of the component indicators $\mathbf{R}$. The inter-arrival times $\tau_{i1}, .., \tau_{iy_i}$ are independent of $\boldsymbol{\alpha}$ and only depend on the number of jumps, whereas τ_{i,y_i+1} depends on the model parameters. Using this we have

$$\begin{aligned} p(\tau|\mathbf{y}, \boldsymbol{\alpha}, \mathbf{R}) &= \prod_{i=1}^{n} p(\tau_{i1}, .., \tau_{iy_i}, \tau_{i,y_i+1}|y_i, \boldsymbol{\alpha}) \\ &= \prod_{i=1}^{n} p(\tau_{i,y_i+1}|y_i, \boldsymbol{\alpha}, \tau_{i1}, .., \tau_{iy_i}) p(\tau_{i1}, .., \tau_{iy_i}|y_i) \end{aligned}$$

It is well known that, given $y_i = n$, the n arrival times of a Poisson process are distributed as the order statistics of n $U([0,1])$ distributed random variables, see for example Mikosch (2004). The last inter-arrival time τ_{i,y_i+1}, given $y_i, \tau_{i1}, .., \tau_{iy_i}$, is exponentially distributed with mean $\frac{1}{\mu_i} = \frac{1}{t_i \exp(\mathbf{z}_i'\boldsymbol{\alpha})}$ conditionally on being greater than $1 - \sum_{j=1}^{y_i} \tau_{ij}$. Using the lack of memory property of the exponential distribution this corresponds to sampling τ_{i,y_i+1} from an exponential distribution with mean $\frac{1}{\mu_i}$ plus an "offset" $1 - \sum_{j=1}^{y_i} \tau_{ij}$. Therefore the inter-arrival times can be sampled as follows:

- If $y_i > 0$
  - sample y_i random numbers $u_{i1}, .., u_{iy_i} \sim U([0,1])$
  - sort these random numbers: $u_{i,(1)}, .., u_{i,(y_i)}$
  - define τ_{ij} as the increments $\tau_{ij} = u_{i,(j)} - u_{i,(j-1)}, \quad j = 1, .., y_i$ where $u_{j,(0)} := 0$
  - sample $\tau_{i,y_i+1} = 1 - \sum_{j=1}^{y_i} \tau_{ij} + \zeta_i$, where $\zeta_i \sim Exponential(\mu_i)$
- If $y_i = 0$ sample $\tau_{i1} = 1 + \zeta_i$, where $\zeta_i \sim Exponential(\mu_i)$

6.1.5 Sampling the component indicators

The component indicators $\mathbf{R}$ are mutually independent given $\boldsymbol{\tau}, \boldsymbol{\alpha}$, therefore

$$p(\mathbf{R}|\boldsymbol{\tau}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \prod_{j=1}^{y_i+1} p(r_{ij}|\tau_{ij}, \boldsymbol{\alpha})$$

Further

$$\begin{aligned} p(r_{ij} = k|\tau_{ij}, \boldsymbol{\alpha}) &= \frac{p(r_{ij} = k, \tau_{ij}, \boldsymbol{\alpha})}{p(\tau_{ij}, \boldsymbol{\alpha})} = \frac{p(\tau_{ij}|r_{ij} = k, \boldsymbol{\alpha})p(r_{ij} = k)}{p(\tau_{ij}|\boldsymbol{\alpha})} \\ &\propto p(\tau_{ij}|r_{ij} = k, \boldsymbol{\alpha})w_k \end{aligned} \tag{6.6}$$

since $w_k = p(r_{ij} = k)$. Since $\log \tau_{ij}|\boldsymbol{\alpha}, r_{ij} \sim N(-\log \mu_i + m_{r_{ij}}, s^2_{r_{ij}})$, τ_{ij} is log normal distributed, i.e.

$$p(\tau_{ij}|r_{ij} = k, \boldsymbol{\alpha}) \propto \frac{1}{s_k \tau_{ij}} \exp\Big[-\frac{1}{2}\Big(\frac{\log(\tau_{ij}) + \log \mu_i - m_k}{s_k}\Big)^2\Big].$$

r_{ij} can therefore be sampled from the discrete distribution (6.6) with $K = 5$ categories.

6.1.6 Starting values

Starting values for the component indicators r_{ij} are obtained by drawing random numbers from 1 to K. For τ_{ij} starting values are generated according to the sampling procedure described in Section 6.1.4. For observations equal to zero we sample $\zeta_i \sim Exponential(0.1)$, for observations greater than zero $\zeta_i \sim Exponential(y_i)$, as suggested in Frühwirth-Schnatter and Wagner (2004b).

6.2 Updating schemes for the regression parameters and spatial effects

For $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ several update schemes are possible and will be discussed in this section. For notational convenience we define with $N := \sum_{i=1}^{n}(y_i + 1)$

$$\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, .., \tilde{\tau}_N) := (\tau_{11}, .., \tau_{1,y_1+1}, \tau_{21}, .., \tau_{2,y_2+1}, .., \tau_{n1}, .., \tau_{n,y_n+1}),$$

$$\tilde{\boldsymbol{\epsilon}} = (\tilde{\epsilon}_1, .., \tilde{\epsilon}_N) := (\epsilon_{11}, .., \epsilon_{1,y_1+1}, \epsilon_{21}, .., \epsilon_{2,y_2+1}, .., \epsilon_{n1}, .., \epsilon_{n,y_n+1}),$$

$$\tilde{\boldsymbol{m}} = (\tilde{m}_1, .., \tilde{m}_N) := (m_{r_{11}}, .., m_{r_{1,y_1+1}}, m_{r_{21}}, .., m_{r_{2,y_2+1}}, .., m_{r_{n1}}, .., m_{r_{n,y_n+1}})$$

and

$$\tilde{\boldsymbol{s}}^{\mathbf{2}} = (\tilde{s}_1^2, .., \tilde{s}_N^2) := (s_{r11}^2, .., s_{r_{1,y_1+1}}^2, s_{r21}^2, .., s_{r_{2,y_2+1}}^2, .., s_{rn1}^2, .., s_{r_{n,y_n+1}}^2).$$

Let $\tilde{\boldsymbol{t}} = (\tilde{t}_1, .., \tilde{t}_N)$ denote the vector where t_i is repeated $y_i + 1$ times. Further define

$$\tilde{\mathbf{y}} = (\tilde{y}_1, .., \tilde{y}_N) := (\log \tilde{\tau}_1 - \tilde{m}_1 + \log \tilde{t}_1, .., \log \tilde{\tau}_N - \tilde{m}_N + \log \tilde{t}_N).$$

Using this notation (6.5) is written as

$$\tilde{y}_i | \boldsymbol{\alpha}, \mathbf{R} \sim N(-\tilde{\mathbf{z}}_i' \boldsymbol{\alpha}, \tilde{s}_i^2)$$

where $\tilde{\mathbf{z}} = \begin{pmatrix} \tilde{\mathbf{z}}_1' \\ \vdots \\ \tilde{\mathbf{z}}_N' \end{pmatrix}$ is a $N \times (p + 1 + J)$-matrix where $\mathbf{z}_i$ is repeated $y_i + 1$ times.

6.2.1 Block update of regression parameters and spatial effects

For a joint update of the regression parameters $\boldsymbol{\beta}$ and the spatial effects $\boldsymbol{\gamma}$ in one block we have to consider the full conditional of $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma})'$ which is given by

$$\begin{aligned} p(\boldsymbol{\alpha}|\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}) &\propto \pi(\boldsymbol{\alpha}|\boldsymbol{\theta}) \prod_{i=1}^{N} \exp\Big(-\frac{1}{2\tilde{s}_i^2}(\tilde{y}_i + \tilde{\mathbf{z}}_i' \boldsymbol{\alpha})^2\Big) \\ &\propto \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\alpha}' \Sigma^{-1} \boldsymbol{\alpha} + \sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2}(\tilde{y}_i + \tilde{\mathbf{z}}_i' \boldsymbol{\alpha})^2\Big]\Big\} \\ &\propto \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\alpha}' \Sigma_\alpha \boldsymbol{\alpha} - 2\boldsymbol{\alpha}' \boldsymbol{\mu}_\alpha\Big]\Big\}, \end{aligned}$$

where $\Sigma_\alpha := \Sigma^{-1} + \sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2} \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i'$ and $\boldsymbol{\mu}_\alpha := -\sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2} \tilde{\mathbf{z}}_i \tilde{y}_i$.
Hence,

$$\boldsymbol{\alpha}|\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R} \sim N_{p+1+J}(\Sigma_\alpha^{-1} \boldsymbol{\mu}_\alpha, \Sigma_\alpha^{-1}).$$

6.2.2 Separate update of regression parameters and spatial effects

The calculation of the posterior covariance matrix Σ_α^{-1} in Section 6.2.1 can be computationally expensive if the number of regression parameters and spatial effects is large as is the case in most spatial applications. Therefore it might be more efficient to update $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in two separate blocks. The full conditional distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are given by

$$\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R} \sim N_{p+1}(\Sigma_\beta^{-1} \boldsymbol{\mu}_\beta, \Sigma_\beta^{-1}) \quad \text{and} \quad \boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R} \sim N_J(\Sigma_\gamma^{-1} \boldsymbol{\mu}_\gamma, \Sigma_\gamma^{-1}).$$

The explicit formulas for Σ_β, $\boldsymbol{\mu}_\beta$, Σ_γ and $\boldsymbol{\mu}_\gamma$ are given in Table 6.2.

Section	
6.2.2	$\Sigma_\beta := V_0^{-1} + \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'$ $\boldsymbol{\mu}_\beta := -\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i(\tilde{y}_i + \gamma_{R(i)})$ $\Sigma_\gamma := \frac{1}{\sigma^2}Q + \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i'$ $\boldsymbol{\mu}_\gamma := -\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta})$
6.2.3 (*block*)	$\Sigma_{\beta_{-0}} := V_{0\beta_{-0}}^{-1} + \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_{\beta_{-0}i}\tilde{\mathbf{x}}_{\beta_{-0}i}'$ $\boldsymbol{\mu}_{\beta_{-0}} := -\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_{\beta_{-0}i}(\tilde{y}_i + \gamma_{R(i)} + \beta_0)$ $\Sigma_{\gamma\beta_0} := \begin{pmatrix} \tau^{-2} & 0 \\ 0 & \frac{1}{\sigma^2}Q \end{pmatrix} + \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}(1,\tilde{\mathbf{v}}_i)(1,\tilde{\mathbf{v}}_i)'$ $\boldsymbol{\mu}_{\gamma\beta_0} := -\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}(1,\tilde{\mathbf{v}}_i)(\tilde{y}_i + \tilde{\mathbf{x}}_{\beta_{-0}i}'\boldsymbol{\beta}_{-0})$ $V_{0\beta_{-0}} = \tau^2 I_p$ $\tilde{\mathbf{x}}_{\beta_{-0}i} = (\tilde{x}_{i1}, .., \tilde{x}_{ip})$
6.2.4 (*coll1*)	$\Sigma_{col} := \tau^{-2}I + \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i' - (\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')'A^{-1}(\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')$ $\boldsymbol{\mu}_{col} := (\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')'A^{-1}(\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{y}_i) - \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{y}_i$ $A := \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i' + \sigma^{-2}Q$
6.2.5 (*coll2*)	$\Sigma_{col}^* := \tau^{-2}I + \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i' - (\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')'(A^*)^{-1}(\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')$ $\boldsymbol{\mu}_{col}^* := (\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')(A^*)^{-1}(\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{y}_i') - \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{y}_i$ $A^* := \sigma^2\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i' + Q$ $\Sigma_\gamma^* := \sigma^2\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i' + Q$ $\boldsymbol{\mu}_\gamma^* := -\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta})$ $\Sigma_\sigma^* := \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}(\gamma_{R(i)}^*)^2 + \tau_\sigma^{-2}$ $\mu_\sigma^* := -\sum_{i=1}^N \gamma_{R(i)}^* \frac{1}{\tilde{s}_i^2}(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta})$
6.2.6 (*coll3*)	$\Sigma_{col}^{**} := \tau^{-2}I + \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i' - (\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{L'\mathbf{v}}_i\tilde{\mathbf{x}}_i')'(A^{**})^{-1}(\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{L'\mathbf{v}}_i\tilde{\mathbf{x}}_i')$ $\boldsymbol{\mu}_{col}^{**} := (\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}L'\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')(A^{**})^{-1}(\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{L'\mathbf{v}}_i\tilde{y}_i') - \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{y}_i$ $A^{**} := \sigma^2\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}L'\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i'L + I$ $\Sigma_\gamma^{**} := \sigma^2\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}L'\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i'L + I$ $\boldsymbol{\mu}_\gamma^{**} := -\sigma\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}L'\tilde{\mathbf{v}}_i(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta})$ $\Sigma_\sigma^{**} := \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}(\boldsymbol{\gamma}^{**\prime}(\sum_{i=1}^N \frac{1}{\tilde{s}_i^2}L'\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i'L)\boldsymbol{\gamma}^{**} + \tau_\sigma^{-2}$ $\mu_\sigma^{**} := -\sum_{i=1}^N \mathbf{v}_i'L\boldsymbol{\gamma}^{**}\frac{1}{\tilde{s}_i^2}(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta})$
6.2.7 (*centered*)	$\boldsymbol{\mu}_\gamma^{cent} := \frac{\beta_0}{\sigma^2}Q\mathbf{1} - \sum_{i=1}^N \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i(\tilde{y}_i + \tilde{\mathbf{x}}_{\beta_{-0}i}'\boldsymbol{\beta}_{-0})$ $\Sigma_{\beta_0} := \frac{1}{\sigma^2}\sum_{i,j=1}^J Q_{ij} + \frac{1}{\tau^2}$ $\mu_{\beta_0} := \frac{1}{\sigma^2}\mathbf{1}'Q\boldsymbol{\gamma}^c$

Table 6.2: Covariance and mean specifications for the update strategies in Sections 6.2.2-6.2.7.

6.2.3 Block update of the intercept and the spatial effects (*block*)

Due to identifiability problems between the intercept β_0 and the spatial effects $\boldsymbol{\gamma}$ mixing and convergence is not very good when $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are updated in two separate blocks. Better results are achieved if a joint block update of β_0 and $\boldsymbol{\gamma}$ is performed, whereas the remaining parameters $\boldsymbol{\beta}_{-0} = (\beta_1, .., \beta_p)$ are still updated in one separate block. With this setting the posterior distributions are given by

$$\boldsymbol{\beta}_{-0}|\beta_0, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R} \sim N_p(\Sigma_{\beta_{-0}}^{-1}\boldsymbol{\mu}_{\beta_{-0}}, \Sigma_{\beta_{-0}}^{-1})$$

and

$$\boldsymbol{\gamma}, \beta_0|\boldsymbol{\beta}_{-0}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R} \sim N_{J+1}(\Sigma_{\gamma\beta_0}^{-1}\boldsymbol{\mu}_{\gamma\beta_0}, \Sigma_{\gamma\beta_0}^{-1})$$

with $\Sigma_{\beta_{-0}}$, $\boldsymbol{\mu}_{\beta_{-0}}$, $\Sigma_{\gamma\beta_0}$ and $\boldsymbol{\mu}_{\gamma\beta_0}$ as given in Table 6.2.

6.2.4 Collapsed algorithm for a model parameterisation with a non-centered mean (*coll1*)

Another possibility is to use a collapsed algorithm. This means, that particular components of the posterior are integrated out and an update based on the marginal distribution is performed. In our context the joint posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be written as

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}) \propto p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R})p(\boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R})$$

where $p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R}) = \int p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R})d\boldsymbol{\gamma}$ is the marginalised posterior density of $\boldsymbol{\beta}$ with $\boldsymbol{\gamma}$ integrated out. It is shown in the Appendix C that

$$\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R} \sim N_{p+1}(\Sigma_{col}^{-1}\boldsymbol{\mu}_{col}, \Sigma_{col}^{-1})$$

with Σ_{col} and $\boldsymbol{\mu}_{col}$ as given in Table 6.2.

Step (1) in the algorithmic scheme presented in Section 6.1.3 is then the following for the collapsed algorithm:

- sample $\boldsymbol{\beta}$ from $N_{p+1}(\Sigma_{col}^{-1}\boldsymbol{\mu}_{col}, \Sigma_{col}^{-1})$
- sample $\boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}$ as in Section 6.2.2

6.2.5 Collapsed algorithm for a model parameterisation with a non-centered mean and scale (*coll2*)

Up to now, we only considered models with the non-centered mean parameterisation specified by (6.2) and the spatial prior $\boldsymbol{\gamma}|\psi, \sigma \sim N_J(0, \sigma^2 Q^{-1})$. In this section we consider a model where the prior of the spatial effects is not only non-centered in the mean, but in the scale as well, i.e. the third model parameterisation given in Table 6.1. By assuming $\boldsymbol{\gamma}^*|\psi \sim N_J(0, Q^{-1})$, σ appears as an unknown parameter in the observation equation, in particular we have

$$\mu_i = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \sigma\gamma^*_{R(i)}).$$

For this parameterisation and $\pi(\cdot)$ denoting the prior distributions, the joint posterior of $\boldsymbol{\beta}, \boldsymbol{\gamma}^*, \psi$ and σ is given by

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}^*, \psi, \sigma|\tilde{y}, \boldsymbol{\tau}, \mathbf{R}) \propto \exp\Big\{-\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\tilde{s}_i^2}(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta} + \sigma\tilde{\mathbf{v}}_i'\boldsymbol{\gamma}^*)^2\Big\}\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}^*|\psi)\pi(\psi)\pi(\sigma).$$

Following the lines of Section 6.2.4 we obtain for $\boldsymbol{\beta}$ the marginalized posterior distribution

$$\boldsymbol{\beta}|\sigma, \boldsymbol{\tau}, \mathbf{R} \sim N_{p+1}((\Sigma^*_{col})^{-1}\boldsymbol{\mu}^*_{col}, (\Sigma^*_{col})^{-1}).$$

The full conditional distribution for $\boldsymbol{\gamma}^*$ is given by

$$\boldsymbol{\gamma}^*|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{R}, \sigma, \psi \;\sim\; N_J((\Sigma^*_{\gamma})^{-1}\boldsymbol{\mu}^*_{\gamma}, (\Sigma^*_{\gamma})^{-1}).$$

The definitions of Σ^*_{col}, $\boldsymbol{\mu}^*_{col}$, Σ^*_{γ} and $\boldsymbol{\mu}^*_{\gamma}$ can be found in Table 6.2. The spatial hyperparameter ψ is again updated using a MH step since the full conditional distribution given by

$$p(\psi|\boldsymbol{\gamma}^*) \propto |Q|^{\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{\gamma}^{*\prime}Q\boldsymbol{\gamma}^*)\pi(\psi)$$

can not be sampled from directly. For this model parameterisation we choose a normal prior for σ, in particular $\sigma \sim N(0, \tau_\sigma^2)$. Note, that σ is not restricted to take positive values, leading to unidentifiability, since the same likelihood results for $(\sigma, \boldsymbol{\gamma}^*)$ and $(-\sigma, -\boldsymbol{\gamma}^*)$. However, as pointed out by Frühwirth-Schnatter (2004), this leads to an improved mixing for models with small scales σ since boundary problems for σ are avoided. The full conditional distribution of σ is then again normal, in particular

$$\sigma|\boldsymbol{\beta}, \boldsymbol{\gamma}^*, \boldsymbol{\tau}, \mathbf{R} \;\sim\; N((\Sigma^*_{\sigma})^{-1}\mu^*_{\sigma}, (\Sigma^*_{\sigma})^{-1}),$$

see Table 6.2 for details on Σ^*_{σ} and μ^*_{σ}.

6.2.6 Collapsed algorithm for a model parameterisation with a non-centered mean and variance (*coll3*)

In this section we consider the model parameterisation non-centered in both mean and variance, also given in Table 6.1. In contrast to the non-centered parameterisation in scale only considered in the previous section, we now assume the prior

$$\boldsymbol{\gamma}^{**} \sim N_J(0, I).$$

The spatial structure incorporated in the precision matrix Q is now moved to the observation equation given by

$$\mu_i = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \sigma \mathbf{v}_i' L \boldsymbol{\gamma}^{**}),$$

where L is a lower triangular matrix resulting from the Cholesky decomposition $Q^{-1} = LL'$. The resulting joint posterior distribution of $\boldsymbol{\beta}, \boldsymbol{\gamma}^{**}, \psi$ and σ is given by

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}^{**}, \psi, \sigma | \tilde{y}, \boldsymbol{\tau}, \mathbf{R}) \propto \exp\Big\{-\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\tilde{s}_i^2}(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta} + \sigma \tilde{\mathbf{v}}_i' L \boldsymbol{\gamma}^{*})^2\Big\} \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\gamma}^{**}) \pi(\psi) \pi(\sigma).$$

The marginalized posterior distribution of $\boldsymbol{\beta}$ changes to

$$\boldsymbol{\beta} | \sigma, \boldsymbol{\tau}, \mathbf{R} \sim N_{p+1}((\Sigma_{col}^{**})^{-1} \boldsymbol{\mu}_{col}^{**}, (\Sigma_{col}^{**})^{-1}),$$

the full conditional distribution of $\boldsymbol{\gamma}^{**}$ is given by

$$\boldsymbol{\gamma}^{**} | \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{R}, \sigma, \psi \quad \sim \quad N_J((\Sigma_{\gamma}^{**})^{-1} \boldsymbol{\mu}_{\gamma}^{**}, (\Sigma_{\gamma}^{**})^{-1}),$$

with Σ_{col}^{**}, $\boldsymbol{\mu}_{col}^{**}$, Σ_{γ}^{**} and $\boldsymbol{\mu}_{\gamma}^{**}$ as given in Table 6.2. The full conditional distribution of ψ is given by

$$\begin{aligned} p(\psi | \boldsymbol{\beta}, \boldsymbol{\gamma}^{**}, \sigma, \boldsymbol{\tau}, \mathbf{R}) \quad &\propto \quad \exp\Big\{-\frac{1}{2}\Big(\boldsymbol{\gamma}^{**\prime}[\sigma^2 \sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2} L' \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_i' L] \boldsymbol{\gamma}^{**} + 2\sigma \sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2} \tilde{\mathbf{v}}_i' L \boldsymbol{\gamma}^{**} (\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta})\Big)\Big\} \\ &\cdot \quad \pi(\psi). \end{aligned}$$

While ψ is again updated using a MH step, the full conditional distribution of σ is given by

$$\sigma | \boldsymbol{\beta}, \boldsymbol{\gamma}^{**}, \psi, \boldsymbol{\tau}, \mathbf{R} \quad \sim \quad N((\Sigma_{\sigma}^{**})^{-1} \mu_{\sigma}^{**}, (\Sigma_{\sigma}^{**})^{-1}),$$

see Table 6.2 for details on Σ_{σ}^{**} and μ_{σ}^{**}. Here again the normal prior $\sigma \sim N(0, \tau_{\sigma}^2)$ is assumed.

6.2.7 Centered CAR-Model (*centered*)

Alternatively, the centered spatial prior $\boldsymbol{\gamma}^c|\beta_0 \sim N(\beta_0, \sigma^2 Q^{-1})$ with $\beta_0 \sim N(0, \tau^2)$ and $\boldsymbol{\beta}_{-0} \sim N(0, \tau^2 I_p)$ can be used. For this model the posterior distribution for $\boldsymbol{\beta}_{-0}$ is the same as in Section 6.2.3 but with $\boldsymbol{\mu}_{\beta_{-0}}$ replaced by $-\sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2} \tilde{\mathbf{x}}_{\beta_{-0}i}(\tilde{y}_i + \gamma^c_{R(i)})$.
The posterior distribution for $\boldsymbol{\gamma}^c$ is given by

$$\boldsymbol{\gamma}^c|\beta_0, \boldsymbol{\beta}_{-0}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}, \mathbf{y} \sim N_J(\Sigma_\gamma^{-1}\boldsymbol{\mu}_\gamma^{cent}, \Sigma_\gamma^{-1})$$

where Σ_γ is the same as in Section 6.2.2 and $\boldsymbol{\mu}_\gamma^{cent}$ is given in Table 6.2.
β_0 is updated in an extra Gibbs step, in particular

$$\beta_0|\boldsymbol{\beta}_{-0}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}, \mathbf{y} \sim N(\Sigma_{\beta_0}^{-1}\mu_{\beta_0}, \Sigma_{\beta_0}^{-1})$$

with Σ_{β_0} and μ_{β_0} defined as in Table 6.2.

6.3 Simulation studies and application

We aim to apply the developed Gibbs samplers to analyse the expected number of claims in a data set from a German car insurance company. The data include 16307 policyholders in Bavaria with full comprehensive car insurance within one year and contain information on several covariates like age and gender of the policyholders, kilometers driven per year and the geographical region each policyholder is living in. Bavaria is divided into 96 regions. The variability of these data is very small, 95% of the observations are zero observations, the highest number of claims observed is only four. The data have been already analysed by Gschlößl and Czado (2005b) who considered both a spatial Poisson regression model as well as spatial models taking overdispersion into account. They show that the spatial effects are very small for these data and have no significant contribution to explaining the expected claim number.
In this section, the performance of the Gibbs sampler schemes developed in Sections 6.1 and 6.2 will be examined on simulated data first. For comparison, we additionally use a single site Metropolis Hastings algorithm for spatial Poisson regression models with an independence proposal where both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are updated component by component. In particular, we use a t-distribution with 20 degrees of freedom as proposal which has the same mode and inverse curvature at the mode as the target distribution.
The performance of the samplers is measured in terms of the computation time required in order to obtain a certain precision of the estimated posterior means of the parameters.

The posterior mean of a variable θ is given by $\bar{\theta} := \sum_{j=1}^{R} \hat{\theta}^j$ with $\hat{\theta}^j, j = 1, .., R$ denoting the MCMC iterates of θ after burnin. The precision of $\bar{\theta}$ is given by the Monte Carlo standard error of $\bar{\theta}$ which is defined as $\sigma_{MC}(\bar{\theta}) := \frac{\sigma_{asy}(\bar{\theta})}{\sqrt{R}}$ where

$$\sigma^2_{asy}(\bar{\theta}) := Var(\theta)\Big(1 + 2\sum_{k=1}^{\infty} \rho_k(\theta)\Big)$$

denotes the asymptotic variance of $\bar{\theta}$, $Var(\theta)$ the sample variance and $\rho_k(\theta)$ the autocorrelation of the MCMC iterates $\hat{\theta}^1, .., \hat{\theta}^R$ at lag k. The asymptotic variance will be estimated using the initial monotone sequence estimator (see Geyer (1992)), defined by

$$\hat{\sigma}^2_{asy}(\bar{\theta}) := \hat{Var}(\theta)(1 + 2\sum_{j=1}^{2m+1} \hat{\rho}_k(\theta)),$$

where m is chosen to be the largest integer such that the sequence $\Gamma_m = \hat{\rho}_{2m}(\theta) + \hat{\rho}_{2m+1}(\theta)$ is positive and monotone. Here $\hat{Var}(\theta) := \hat{\gamma}_0$, $\hat{\rho}_k(\theta) := \dfrac{\hat{\gamma}_k}{\hat{\gamma}_0}$, $\hat{\gamma}_k := \frac{1}{R}\sum_{j=1}^{R-k}(\hat{\theta}^j - \bar{\theta})(\hat{\theta}^{j+k} - \bar{\theta})$. We additionally require the estimated empirical autocorrelations $\hat{\rho}_{2m+1}(\theta)$ to fall below 0.1.

In order to obtain a certain precision k, $R = \frac{\hat{\sigma}^2_{asy}}{k^2}$ samples are needed. Hence, the computation time required to obtain a precision k for an algorithm with computational costs m per iteration, is given by $R \cdot m$. For a direct comparison of the Gibbs sampler schemes to the MH independence sampler we consider the computational costs relative to the costs of the MH sampler required to obtain the same precision of the posterior means of the parameters. This is given by $R_{rel} \cdot m_{rel} := \frac{\hat{\sigma}^2_{asy}}{\hat{\sigma}^2_{asy,ind}} \cdot \frac{m}{m_{ind}}$, where $\hat{\sigma}^2_{asy,ind}$ and m_{ind} denote the estimated asymptotic variance and the computational costs for one iteration of the MH independence sampler.

We consider two studies. In the first study the influence of the size of the spatial effects on mixing behaviour is examined, while in the second study the impact of data heterogeneity is investigated. In both studies the Gibbs samplers described in Sections 6.2.3-6.2.7, i.e. the following model parameterisations and update schemes are assumed:

- non-centered mean:
  - block update of $\boldsymbol{\beta}_{-0}|\beta_0, \boldsymbol{\gamma}$ and $(\beta_0, \boldsymbol{\gamma})|\boldsymbol{\beta}_{-0}$ given in Section 6.2.3 (*block*)
  - collapsed algorithm given in Section 6.2.4 (*coll1*)
- non-centered mean and scale: collapsed algorithm given in Section 6.2.5 (*coll2*)
- non-centered mean and variance: collapsed algorithm given in Section 6.2.6 (*coll3*)

- centered parameterisation: algorithm given in Section 6.2.7 (*centered*)

In the following we will refer to these samplers as *block*, *coll1*, *coll2*, *coll3* and *centered*.

6.3.1 Computational costs

Recall, that by using the data augmentation scheme described above, we are no longer dealing with n observations, but with $N = \sum_{i=1}^{n}(y_i + 1)$ latent inter-arrival times τ_{ij} and mixture component indicators r_{ij}. Both $\boldsymbol{\tau}$ and $\mathbf{R}$ have to be updated, therefore the number of variables to sample from in each iteration is $2N + J + p + 1(+2$ hyperparameters) in comparison to $J+p+1(+2$ hyperparameters) variables in the MH independence sampler. The MH independence sampler in contrast requires the calculation of the posterior mode and the inverse curvature at the posterior mode for each of the $J + p + 1$ components in every iteration. The posterior mode may be obtained using the bisection method for example. In our simulation studies, except the sampler *coll3*, the Gibbs samplers are always faster than the MH independence sampler. However, the computational advantage of the Gibbs samplers depends on the complexity of the model. The computational costs m_{rel} relative to the costs of the MH sampler for one iteration are reported in Table 6.3. For the setting in Study 1 with 5000 observations, an intercept and two covariates for example, the *centered* Gibbs sampler only takes 0.86 times as long as the MH independence sampler. For the setting in Study 2 with a larger data set the *centered* Gibbs sampler even takes only 0.26 times as long. Among the Gibbs samplers the *centered* Gibbs sampler is the fastest, followed closely by the Gibbs sampler using a block update. The collapsed Gibbs samplers non-centered in the mean (*coll1*) and non-centered in mean and scale (*coll2*) require slightly more time than the *centered* Gibbs sampler. The computational effort for the Gibbs sampler in the model parameterisation non-centered in the mean and the variance (*coll3*) however is more than twice as large. In this algorithm a Cholesky decomposition of the precision matrix Q^{-1} has to be performed in every iteration.

6.3.2 Study 1: Influence of the size of the spatial effects

We consider two simulated data sets of size 5000 with $y_i \sim Poisson(\mu_i), i = 1, .., 5000$. For both data sets the mean μ_i is specified by

$$\mu_i = \exp(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \gamma_{R(i)})$$

where $\mathbf{x}_1$ is an indicator variable and $\mathbf{x}_2$ a continuous standardized variable. The exposure is assumed to be $t_i = 1$ for all observations. We assume a simple spatial structure, namely

sampler	Study 1	Study 2
independence	1	1
block	0.87	0.27
centered	0.86	0.26
coll1	0.96	0.30
coll2	0.99	0.31
coll3	2.18	0.57

Table 6.3: Computation times m_{rel} for the different samplers relative to the MH independence sampler for the settings in Study 1 and Study 2.

100 regions on a 10×10 grid. The spatial effects $\boldsymbol{\gamma}$ are generated according to the CAR prior $\boldsymbol{\gamma} \sim N(0, \sigma^2 Q^{-1})$ with spatial dependence parameter $\psi = 3$. For the first simulated data set $\mathbf{y}_1$ we assume $\sigma^2 = 1$ resulting in a range of $[\min(\boldsymbol{\gamma}) \max(\boldsymbol{\gamma})] = [-0.86, 0.85]$ for the spatial effects, whereas for the second data set $\mathbf{y}_2$ we take $\sigma^2 = 0.01$ resulting in a range of $[\min(\boldsymbol{\gamma}) \max(\boldsymbol{\gamma})] = [-0.08, 0.08]$. The Gibbs samplers *block*, *coll1*, *coll2*, *coll3* and *centered* as well as the independence MH sampler are run for 5000 iterations, a burnin of 1000 iterations is taken. As described above, the performance of the samplers is measured in terms of the Monte Carlo standard error of the posterior means of the parameters and the required computation times. Since estimation of the Monte Carlo error is based on the estimated empirical autocorrelations, this quantity also depends on the mixing of the samplers. For a fair comparison of the Monte Carlo error of the spatial effects the model parameterisation of each sampler has to be taken into account. Therefore we compute the Monte Carlo error for $\beta_0 + \boldsymbol{\gamma}$ for the MH independence sampler and the samplers *block* and *collapsed*, while for the *centered* sampler the standard error of $\boldsymbol{\gamma}$ is considered since here the intercept is the spatial prior mean and therefore already included in $\boldsymbol{\gamma}$. For the *coll2* and *coll3* samplers the Monte Carlo errors for $\beta_0 + \sigma\boldsymbol{\gamma}$ and $\beta_0 + \sigma L\boldsymbol{\gamma}$, respectively, are computed.

In the left panel of Table 6.4, for each sampler the Monte Carlo standard errors and the performance relative to the MH independence sampler $R_{rel} \cdot m_{rel}$ are reported for the regression parameters β_1, β_2 and the spatial effects in data set $\mathbf{y}_1$. For the spatial effects the average error, taken over all J components, is given. Additionally plots of the empirical estimated autocorrelations are presented in Figure 6.1. In the left panel the autocorrelations for 25 of the spatial effects, in the right panel autocorrelations for the regression effects are plotted. Mixing for all Gibbs samplers is reasonable well, in average autocorre-

lations of the spatial effects are below 0.1 at a lag of about 16 to 18. The average Monte Carlo error for the spatial effects is around 0.01 for all Gibbs samplers. The Monte Carlo error of the regression parameters however is lower for the collapsed Gibbs samplers, for the *block* and the *centered* Gibbs sampler especially the autocorrelations of β_1 decrease rather slowly.

The independence MH sampler in contrast, displays the smallest Monte Carlo error for both spatial effects and regression parameters. In average the autocorrelations of $\beta_0+\gamma_j$ are below 0.1 at a lag 3 already, the autocorrelations for the regression parameters decrease rapidly as well. Considering the computational effort relative to the MH independence sampler, given by $R_{rel} \cdot m_{rel}$, the MH independence sampler outperforms the Gibbs samplers considerably. The computational effort required to obtain the same precision of the posterior means of the spatial effects is more than 5 times as large for the Gibbs samplers compared to the independence sampler.

sampler	Data set $\mathbf{y}_1$			Data set $\mathbf{y}_2$		
	spatial effects	β_1	β_2	spatial effects	β_1	β_2
independence	0.0041	0.0015	0.0032	0.0021	0.0013	0.0030
	1	1	1	1	1	1
block	0.0100	0.0039	0.0130	0.0031	0.0042	0.0108
	5.18	5.88	14.36	1.90	9.08	11.28
centered	0.0102	0.0045	0.0115	0.0061	0.0078	0.0279
	5.32	7.74	11.11	7.26	30.96	74.38
coll1	0.0101	0.0031	0.0117	0.0022	0.0024	0.0097
	5.83	4.10	12.83	1.05	3.27	10.04
coll2	0.0099	0.0027	0.0105	0.0025	0.0029	0.0114
	5.77	3.21	10.66	1.40	4.93	14.30
coll3	0.0101	0.0024	0.0102	0.0023	0.0026	0.0133
	13.23	5.58	22.15	2.62	8.72	42.85

Table 6.4: Estimated $\hat{\sigma}_{MC}$ (upper row) for the regression parameters β_1, β_2 and average estimated $\hat{\sigma}_{MC}$ for the spatial effects $\boldsymbol{\gamma}+\beta_0$ in the MH independence, *block*, *coll1* sampler, $\boldsymbol{\gamma}$ in the *centered*, $\beta_0+\sigma\boldsymbol{\gamma}$ in the *coll2* and $\beta_0+\sigma L\boldsymbol{\gamma}$ in the *coll3* sampler, as well as $R_{rel}\cdot m_{rel}$ (lower row) for all parameters for data set $\mathbf{y}_1$ and $\mathbf{y}_2$ using different update strategies in Study 1.

The corresponding results for data set $\mathbf{y}_2$ with small spatial effects are reported in the right panel in Table 6.4, plots of the estimated empirical autocorrelations are given in Figure 6.2. Here, clearly the lowest precision and worst mixing is obtained if the Gibbs sampler based on the centered model parameterisation is used. This confirms the results given in Gelfand et al. (1995). They show that for a hierarchical normal linear model with random effects the centered parameterisation is efficient if the variance of the random effects dominates the variance in the data. However, if the variance of the random effects is very small in contrast to the variability of the data (as it is the case in data set $\mathbf{y}_2$), high posterior correlations result. For the *block* and particularly the collapsed Gibbs samplers a considerably lower Monte Carlo error is obtained. The average Monte Carlo error of the spatial effects in the collapsed sampler *coll1* is almost as small as in the MH independence sampler. For the regression effects however, the MH independence sampler exhibits lower Monte Carlo standard errors. The computational costs $R_{rel} \cdot m_{rel}$ relative to the MH sampler, which are required to obtain the same precision of the posterior means of the parameters are greater 1 for all Gibbs samplers for both spatial effects and regression parameters. Hence, the independence sampler gives the best performance for data set $\mathbf{y}_2$ as well.

The variance of the two simulated data sets $\mathbf{y}_1$ and $\mathbf{y}_2$ takes the values $\text{var}(\mathbf{y}_1) = 0.51$ and $\text{var}(\mathbf{y}_2) = 0.49$. However, the variability of our real data from a car insurance company is very small, the variance of these data is only 0.05. Therefore we conduct a second simulation study where we examine whether the heterogeneity of the data influences the performance of the samplers.

6.3.3 Study 2: Influence of data heterogeneity

We simulate two data sets based on the design of the real data where, according to Gschlößl and Czado (2005b), eight covariates significant for explaining the expected claim number y_i were observed, i.e. $y_i \sim Poisson(\mu_i), i = 1, .., 16307$ with

$$\mu_i = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_{R(i)}).$$

Here $\mathbf{x}_i = (1, x_{i1}, .., x_{i8})$ and $x_{ik}, k = 1, .., 8$ are standardized categorical and metrical covariates, the observation specific exposure t_i takes values up to one year. In this setting we have 96 irregular regions in Bavaria. The spatial effects $\boldsymbol{\gamma}$ again are generated according to the CAR prior $\boldsymbol{\gamma} \sim N(0, \sigma^2 Q^{-1})$ with $\psi = 8$ and $\sigma^2 = 0.01$. This results in small spatial effects with a range of $[-0.06 \quad 0.08]$, i.e. spatial effects similar to the ones observed in

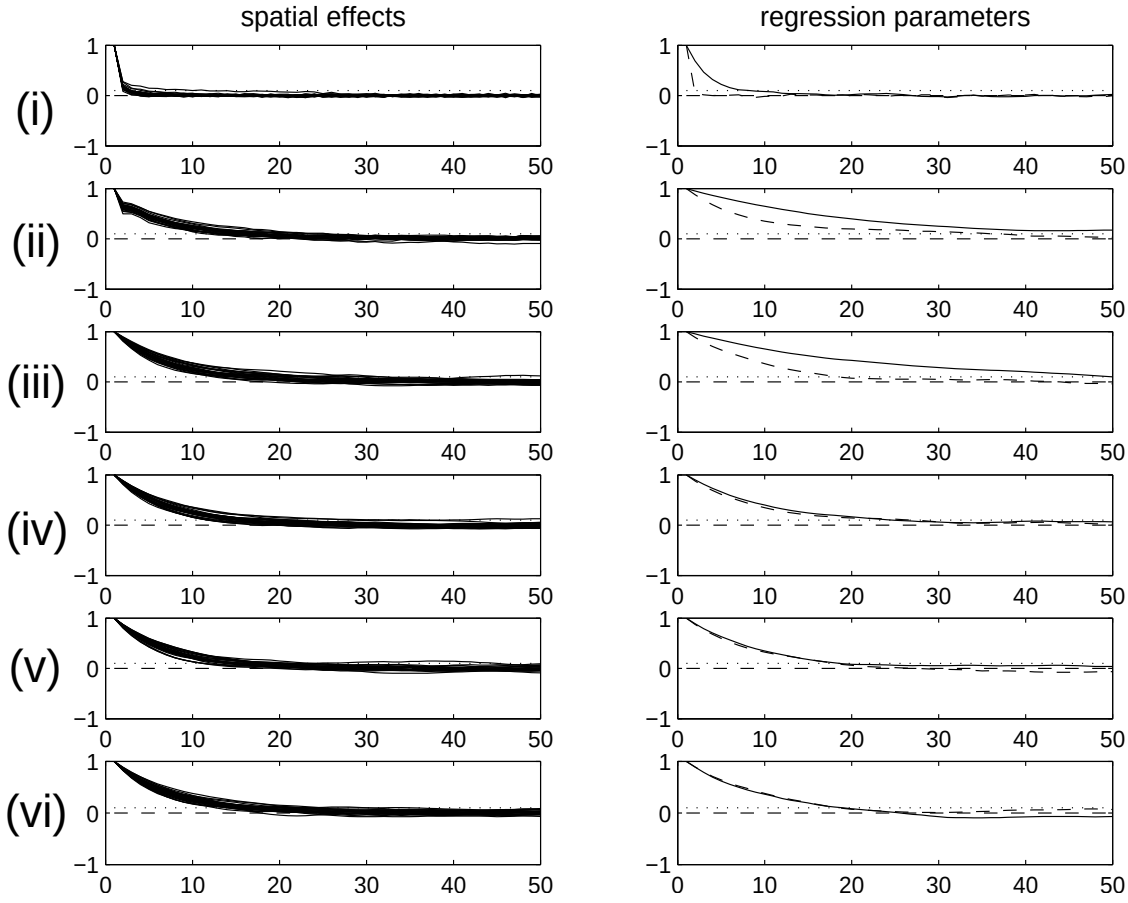

Figure 6.1: Estimated empirical autocorrelations for the spatial effects (left panel) and the regression parameters β_1 (solid),β_2 (dashed) (right panel) for the independence MH sampler (i), the *block* (ii), *centered* (iii), *coll1* (iv), *coll2* (v) and *coll3* (vi) Gibbs samplers for data set $\mathbf{y}_1$.

our real data set. For the first data set $\mathbf{y}_3$ the intercept β_0 is taken to be -1, whereas for the second data set $\mathbf{y}_4$ we take $\beta_0 = -2.5$. For the remaining regression parameters the same values are assumed for both data sets. The resulting variances of $\mathbf{y}_3$ and $\mathbf{y}_4$ are $Var(\mathbf{y}_3) = 0.46$ and $Var(\mathbf{y}_4) = 0.05$, i.e. data set $\mathbf{y}_4$ has very low heterogeneity and is close to our real data. The variance of data set $\mathbf{y}_3$ is not particularly high either, but in comparison to data set $\mathbf{y}_4$ we will refer to this data set as data with high heterogeneity. The *block*, *centered*, *coll1*, *coll2* and *coll3* Gibbs samplers are run for 5000 iterations, the first 1000 iterations are discarded for burnin. For comparison again the MH independence sampler is applied. The Monte Carlo errors for the posterior means of the regression parameters $\beta_1, .., \beta_8$, the spatial effects $\boldsymbol{\gamma}$ in the centered, $\beta_0 + \boldsymbol{\gamma}$ in the non-centered mean, $\beta_0 + \sigma\boldsymbol{\gamma}$ in the non-centered mean and scale and $\beta_0 + \sigma L\boldsymbol{\gamma}$ in the non-centered mean

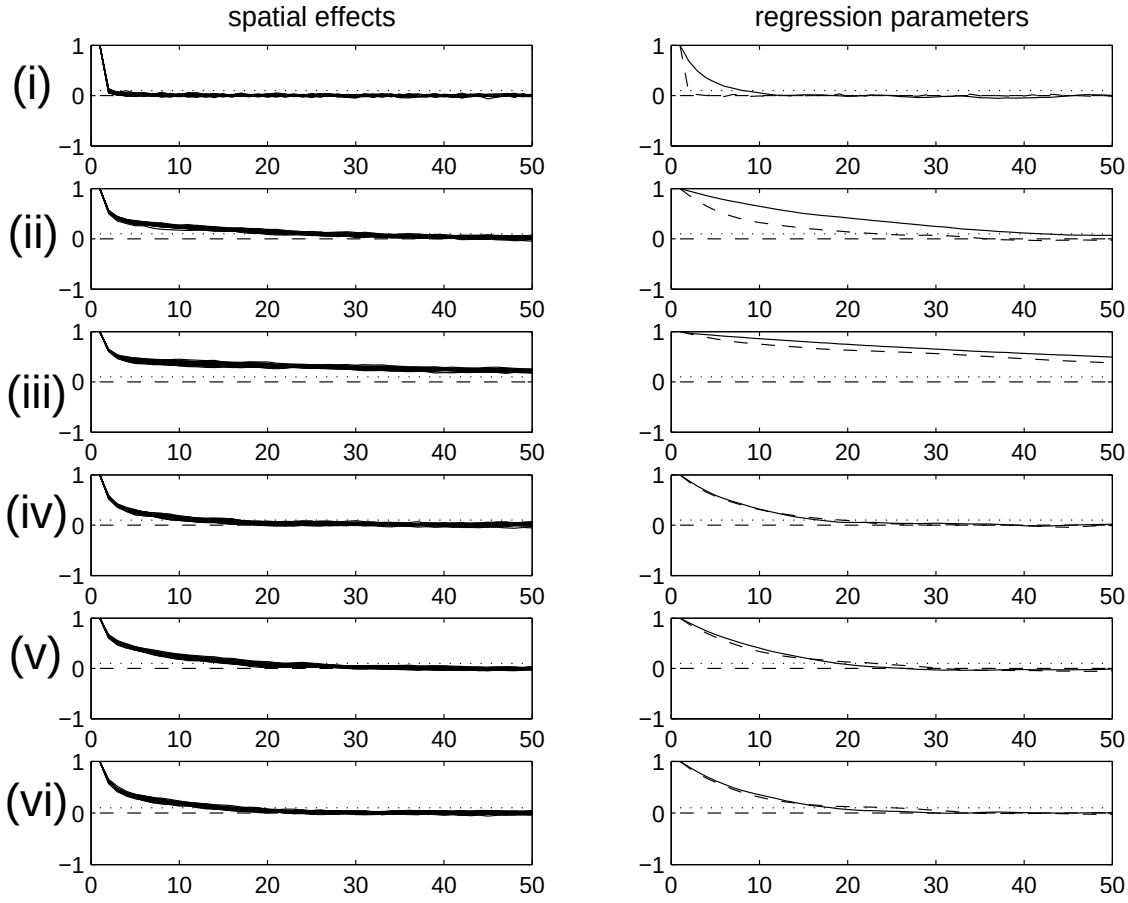

Figure 6.2: Estimated empirical autocorrelations for the spatial effects (left panel) and the regression parameters β_1 (solid),β_2 (dashed) (right panel) for the independence MH sampler (i), the *block* (ii), *centered* (iii), *coll1* (iv), *coll2* (v) and *coll3* (vi) Gibbs samplers for data set $\mathbf{y}_2$.

and variance model parameterisation and the quantities $R_{rel} \cdot m_{rel}$ are reported in Table 6.5. For the high heterogeneity data set $\mathbf{y}_3$ the collapsed Gibbs samplers *coll2* and *coll3* exhibit the lowest Monte Carlo errors for the spatial effects among the Gibbs samplers. The sampler *coll2* even only requires 38 % of the computational effort of the MH sampler in order to obtain the same precision for the spatial effects. The precision and autocorrelations (see Figure 6.3) of the regression effects however are considerably smaller in the independence sampler compared to all Gibbs samplers. In order to achieve a high precision like in the MH sampler for all parameters, for each Gibbs sampler the maximum relative effort $R_{rel} \cdot m_{rel}$, occurring for spatial and regression parameters, is required. Since the maximum values $R_{rel} \cdot m_{rel}$ are considerably greater than 1 for each Gibbs sampler, the MH sampler is clearly superior to the Gibbs samplers.

The average Monte Carlo error for the spatial effects in data set $\mathbf{y}_4$ with low hetero-

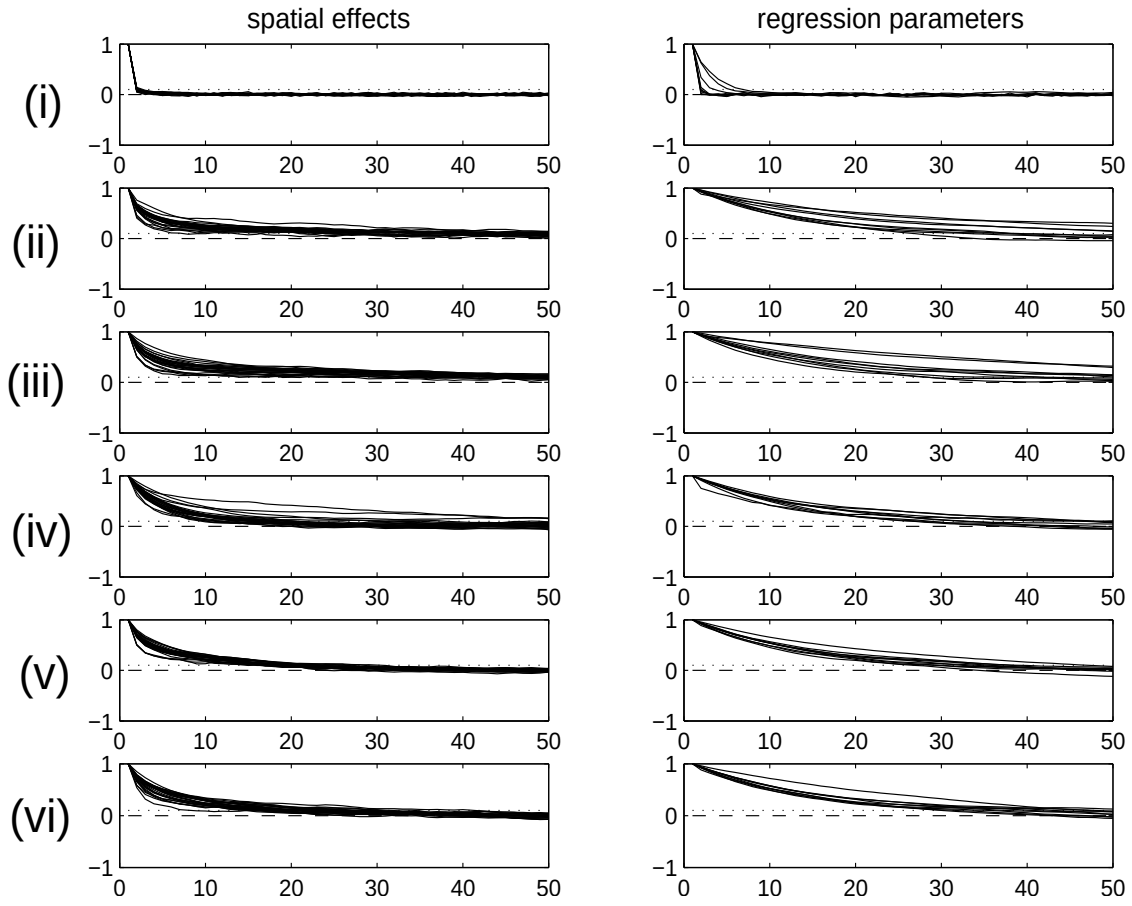

Figure 6.3: Estimated empirical autocorrelations for the spatial effects (left panel) and the regression parameters $\beta_1, .., \beta_8$ (right panel) for the independence MH sampler (i), the *block* (ii), *centered* (iii), *coll1* (iv), *coll2* (v) and *coll3* (vi) Gibbs samplers for data set $\mathbf{y}_3$.

geneity is rather high for the three Gibbs sampler schemes *block*, *centered* and *coll1* for both spatial effects and regression parameters, the estimated empirical autocorrelations plotted in Figure 6.4 decrease very slowly. While for the high heterogeneity data $\mathbf{y}_3$ the computational costs in order to obtain the same precision for the spatial effects of the *block* Gibbs sampler are only 0.65 times as large as of the MH sampler, for the data $\mathbf{y}_4$ the performance of the Gibbs sampler is clearly worse with $R_{rel} \cdot m_{rel} = 5.37$. Results are improved for the collapsed algorithms based on the model parameterisations non-centered in the scale (*coll2*) and in the variance (*coll3*). The sampler *coll2* performs even better than the MH sampler ($R_{rel} \cdot m_{rel} = 0.45$). As indicated in Section 6.2.5, the model parameterisation with non-centered scale is supposed to improve mixing particularly for models with small scale σ^2 which is the case for data sets $\mathbf{y}_3$ and $\mathbf{y}_4$. However, the Monte Carlo

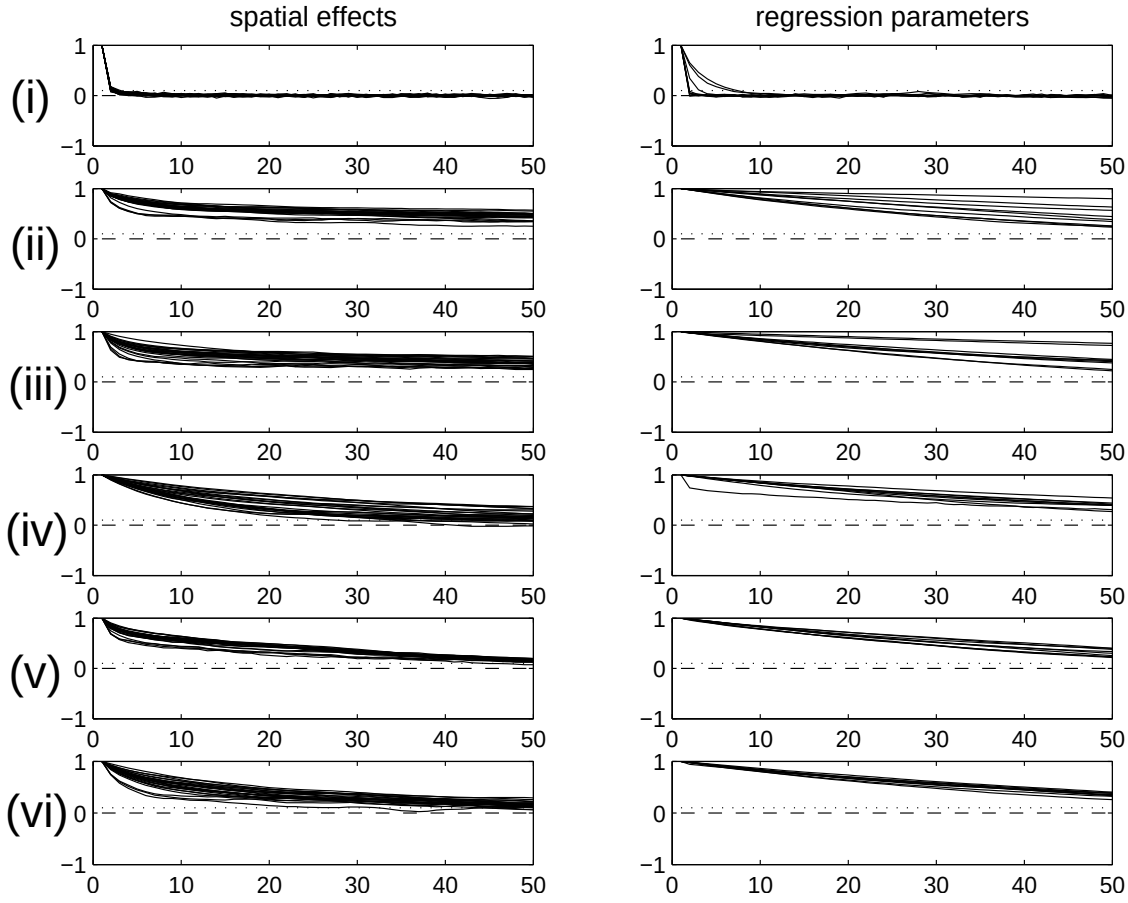

Figure 6.4: Estimated empirical autocorrelations for the spatial effects (left panel) and the regression parameters $\beta_1, .., \beta_8$ (right panel) for the independence MH sampler (i), the *block* (ii), *centered* (iii), *coll1* (iv), *coll2* (v) and *coll3* (vi) Gibbs samplers for data set $\mathbf{y}_4$.

errors for the regression parameters are rather high for all Gibbs samplers and in particular considerably higher than for the high heterogeneity data $\mathbf{y}_3$. The MH independence sampler in contrast exhibits a high precision for all parameters again. Compared to data set $\mathbf{y}_3$, the standard errors for all parameters resulting from the MH sampler are about twice as large for data set $\mathbf{y}_4$, this loss of precision however is much smaller than for the Gibbs samplers. According to the performance measure $R_{rel} \cdot m_{rel}$ for the regression parameters, the MH sampler outperforms the Gibbs samplers considerably. For example, although the Gibbs sampler *coll2* sampler only requires 31 % of the computation time of the MH sampler for one iteration (see Table 6.3), 30.33 ($R_{rel} \cdot m_{rel}$ for β_2) times the effort of the MH sampler for data set $\mathbf{y}_4$ would be needed in order to obtain for all parameters a precison comparable to the MH sampler.

Note that, compared to the collapsed algorithm *coll2*, the collapsed algorithm *coll3* does

not display significantly lower standard errors, neither in Study 1 nor in Study 2. The additional computational effort required for *coll3* which is more than twice as large as for *coll2*, see Table 6.3, does not pay off.

data	sampler	spatial effects	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
$\mathbf{y}_3$	ind	0.0020	0.0346	0.0327	0.0003	0.0003	0.0008	0.0010	0.0003	0.0002
		1	1	1	1	1	1	1	1	1
	block	0.0031	0.1771	0.1627	0.0013	0.0022	0.0042	0.0039	0.0015	0.0009
		0.65	7.07	6.68	5.07	14.54	7.44	4.11	6.75	5.47
	centered	0.0036	0.1955	0.1768	0.0015	0.0020	0.0038	0.0045	0.0013	0.0010
		0.84	8.30	7.60	6.50	11.56	5.87	5.27	4.88	6.50
	coll1	0.0040	0.1487	0.1635	0.0011	0.0018	0.0032	0.0032	0.0015	0.0009
		1.20	5.54	7.50	4.03	10.80	4.80	3.07	7.50	6.08
	coll2	0.0022	0.1561	0.1736	0.0014	0.0021	0.0030	0.0031	0.0012	0.0010
		0.38	6.31	8.74	6.75	15.19	4.36	2.98	4.96	7.75
	coll3	0.0024	0.1899	0.1505	0.0014	0.0022	0.0031	0.0028	0.0016	0.0011
		0.82	17.17	12.07	12.41	30.65	8.56	4.47	16.21	17.24
$\mathbf{y}_4$	ind	0.0048	0.0673	0.0611	0.0006	0.0006	0.0017	0.0021	0.0006	0.0005
		1	1	1	1	1	1	1	1	1
	block	0.0214	0.5199	0.3323	0.0038	0.0041	0.0217	0.0211	0.0076	0.0030
		5.37	16.11	7.99	10.83	12.61	43.99	27.26	43.32	9.72
	centered	0.0114	0.5906	0.4910	0.0040	0.0052	0.0150	0.0209	0.0060	0.0055
		1.47	20.02	16.79	11.56	19.53	20.24	25.75	26.00	31.46
	coll1	0.0189	0.6749	0.6181	0.0052	0.0057	0.0129	0.0133	0.0049	0.0049
		4.65	30.17	30.70	22.53	27.08	17.27	12.03	20.01	28.81
	coll2	0.0058	0.5505	0.6044	0.0038	0.0048	0.0076	0.0070	0.0050	0.0041
		0.45	20.74	30.33	12.43	19.84	6.20	3.44	21.53	20.84
	coll3	0.0091	0.5301	0.4789	0.0052	0.0056	0.0097	0.0096	0.0048	0.0044
		2.05	35.36	35.02	42.81	49.65	18.56	11.91	36.48	44.14

Table 6.5: Estimated $\hat{\sigma}^2_{MC}$ (upper row) for the regression parameters $\beta_1, .., \beta_8$ and estimated average $\hat{\sigma}^2_{MC}$ for the spatial effects $\boldsymbol{\gamma} + \beta_0$ in the MH independence, *block*, *coll1* sampler, $\boldsymbol{\gamma}$ in the *centered*, $\beta_0 + \sigma\boldsymbol{\gamma}$ in the *coll2* and $\beta_0 + \sigma L\boldsymbol{\gamma}$ in the *coll3* sampler, as well as $R_{rel} \cdot m_{rel}$ (lower row) for data set $\mathbf{y}_3$ and $\mathbf{y}_4$ using different update strategies in Study 2.

6.3.4 Application to car insurance data

Finally we apply the discussed Gibbs samplers as well as the independence MH sampler on the car insurance data set described at the beginning of this section. The Monte Carlo errors for the posterior means of the regression and the spatial effects as well as the corresponding values of $R_{rel} \cdot m_{rel}$ are reported in Table 6.6. Similar results as for data set $\mathbf{y}_4$ which is very close to our real data, are observed. In particular for the regression parameters, the performance of all Gibbs samplers is considerably worse than the performance of the MH independence sampler. When using the non-centered scale and variance parameterisations at least for the spatial effects reasonable low errors are obtained, however, according to the relative effort $R_{rel} \cdot m_{rel}$ the MH sampler is still superior.

sampler	spatial effects	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
independence	0.0046	0.0673	0.0628	0.0006	0.0007	0.0017	0.0020	0.0006	0.0005
	1	1	1	1	1	1	1	1	1
block	0.0192	0.5201	0.3823	0.0051	0.0083	0.0116	0.0203	0.0037	0.0047
	4.70	16.13	10.01	19.51	37.96	12.57	27.82	10.27	23.86
centered	0.0138	0.5465	0.5847	0.0045	0.0068	0.0126	0.0145	0.0048	0.0028
	2.34	17.14	22.54	14.63	24.54	14.28	13.67	16.64	8.15
coll1	0.0207	0.5967	0.5753	0.0040	0.0073	0.0155	0.0082	0.0043	0.0032
	6.08	23.58	25.28	13.33	32.63	24.94	5.04	15.41	12.29
coll2	0.0116	0.4359	0.6172	0.0046	0.0063	0.0122	0.0096	0.0057	0.0044
	1.97	13.00	29.94	18.22	25.11	15.97	7.14	27.98	24.01
coll3	0.0100	0.5167	0.5945	0.0056	0.0060	0.0115	0.0110	0.0054	0.0036
	2.69	33.60	51.08	49.65	41.88	26.08	17.24	46.17	29.55

Table 6.6: Estimated $\hat{\sigma}_{MC}$ (upper row) for the regression parameters $\beta_1, .., \beta_8$ and average estimated $\hat{\sigma}_{MC}$ for the spatial effects $\boldsymbol{\gamma} + \beta_0$ in the MH independence, *block*, *coll1* sampler, $\boldsymbol{\gamma}$ in the *centered*, $\beta_0 + \sigma\boldsymbol{\gamma}$ in the *coll2* and $\beta_0 + \sigma L\boldsymbol{\gamma}$ in the *coll3* sampler, as well as $R_{rel} \cdot m_{rel}$ (lower row) for the car insurance data using different update strategies.

6.4 Summary and conclusions

We have presented a new MCMC methodology for spatial Poisson regression models, extending the approach by Frühwirth-Schnatter and Wagner (2004b). Using data augmentation we have shown that a straightforward Gibbs sampler for spatial Poisson models is available. Several update schemes like a joint block update of the intercept and the spatial effects as well as collapsed algorithms have been discussed. Further we have addressed the issue of model parameterisation, centered as well as non-centered model parameterisations in the mean, the scale and the variance have been considered. The performance of the Gibbs sampler based on different model parameterisations and update schemes has been compared to a single site MH independence sampler on simulated and real data. Performance is measured in terms of the computational costs required in order to obtain the same precision of the posterior means of the parameters.

For data which are not too homogeneous, the Gibbs samplers display good mixing and reasonable small Monte Carlo errors. In particular for data with small spatial random effects, the performance is improved when collapsed Gibbs samplers are used, while the centered parameterisation is not very efficient any more in this case. The MH independence sampler however exhibits the smallest Monte Carlo errors for all parameters for data with both small and large spatial effects. Taking additionally the required computation times of the samplers into account, the MH sampler gives the best performance.

For data with low heterogeneity the Monte Carlo errors increase significantly for all Gibbs samplers, mixing of the samplers is much worse. The MH sampler in contrast also mixes well for low heterogeneity data, the precision of the posterior means of the parameters is considerably higher than for the Gibbs samplers. Considering the computation times of the samplers and the required MCMC iterations in order to obtain the same precision for all parameters, the MH sampler clearly outperforms the Gibbs samplers for low heterogeneity data. Similar results are observed for the real data which also display low heterogeneity.

In the literature various approaches for MCMC estimation in spatial Poisson models are provided. Knorr-Held and Rue (2002) discuss efficient block sampling MH algorithms for Markov random field models in disease mapping, based on the methodology developed in Rue (2001). Haran et al. (2003) study MH algorithms with proposal distributions based on Structured MCMC, introduced by Sargent et al. (2000), for spatial Poisson models, while Christensen et al. (2005) discuss Langevin-Hastings updates in spatial GLMM's. Rue et al. (2004) present non-Gaussian approximations to hidden Markov random fields

and give applications in disease mapping and geostatistical models. These methods have been found to be superior to a conventional MH sampler only performing individual updates of the parameters. Therefore, since a single site MH sampler clearly outperformed the Gibbs samplers developed in this chapter, a comparison of the Gibbs samplers to these methods seems to be unnecessary. However, the performance of the Gibbs samplers might be improved by applying the reparameterisation techniques presented in Christensen et al. (2005).

Chapter 7

Spatial regression models for claim size

Premiums in insurance are based on the expected total claim size which is determined both by the number of claims as well as by the individual or average claim sizes. Spatial regression models for the number of claims have been discussed in detail in the previous chapters. Now, we will consider spatial regression models for claim size. One important contribution of this thesis is that we allow for dependencies between the number of claims and claim size. This is in contrast to the classical compound Poisson model going back to Lundberg (1903), where independence of claim frequency and claim size is assumed. In the classical Poisson-Gamma model the number of claims is assumed to follow a Poisson distribution and to be independent of the claim sizes which are modelled Gamma distributed. The use of GLMs in actuarial science has been discussed by Haberman and Renshaw (1996) who give several applications, including premium rating in non-life insurance based on models for claim frequency and average claim size. A more detailed study of GLMs for claim frequency and average claim sizes taking covariate information into account is given in Renshaw (1994). Taylor (1989) and Boskov and Verrall (1994) analyse household contents insurance data incorporating geographic information. Whereas Taylor (1989) uses spline functions, Boskov and Verrall (1994) assume a spatial Bayesian model based on Besag et al. (1991). However, instead of performing a separate modelling of claim frequency and claim size, in both papers adjusted loss ratios are fitted.

Another approach, which also does not perform a separate analysis of claim size and frequency is given by Jørgensen and de Souza (1994) and Smyth and Jørgensen (2002). They use a compound Poisson model, which they call Tweedie's compound Poisson model due to its association to exponential dispersion models, for analysing car insurance data.

Based on the joint distribution of the number of claims and the total claim sizes, they model the claim rate, defined by the total costs per exposure unit, directly. The problem of premium rating taking latent structures into account and using separate models for claim frequency and claim size has been addressed by Dimakos and Frigessi (2002). Based on a spatial Poisson regression model and a spatial Gamma regression model for the average claim size, they determine premiums by the product of the expected claim frequency and the expected claim size. This approach relies on the independence assumption of claim frequency and claim size. Here the spatial structure is modelled via an improper Markov Random Field following Besag et al. (1991).
This thesis extends the approach by Dimakos and Frigessi (2002) in several ways. Whereas Dimakos and Frigessi (2002) consider only models for the average claim size per policyholder, we alternatively assume models for the individual claim sizes of each policyholder and investigate whether this leads to improved predictions of the total claim sizes. Further we allow for dependencies between the number of claims and claim size. In particular, claim size is modelled conditionally on the number of claims which allows us to include the observed number of claims as covariate. The models for individual and average claim sizes are presented in Sections 7.1 and 7.2, respectively.
Based on the MCMC output of the models for claim frequency and the average and individual claim sizes, respectively, the posterior predictive distribution of the total claim sizes can be approximated using simulation. This is described in Section 7.3. We would like to emphasize again, that independence of claim size and claim frequency is not necessary here.

7.1 Modelling individual claim sizes

For policyholder $i = 1, .., n$ let $S_{ik}, k = 1, .., N_i$, denote the individual claim sizes for the N_i observed claims. It is natural for the analysis of claim size to take only observations with positive claim size into account. In this thesis we are interested in modelling claim sizes resulting from traffic accidents in car insurance, not including IBNR (incurred but not reported) losses. These kind of data typically have a skewed distribution, but do no contain extremely high claims which would require the use of heavy tailed distributions like for example the Pareto distribution, see for example Mikosch (2004). Therefore we assume a Gamma model for claim size. In particular, we assume the individual claim sizes

to be independent Gamma distributed conditionally on N_i

$$S_{ik}|N_i \sim Gamma(\mu_i^{SI}, v),\ k = 1, .., N_i, i = 1, .., n \tag{7.1}$$

with mean and variance given by

$$E(S_{ik}|N_i) = \mu_i^{SI} \qquad \text{and} \qquad Var(S_{ik}|N_i) = \frac{(\mu_i^{SI})^2}{v}.$$

For the parameterisation used here, the density of the Gamma distribution is given by

$$f(s_{ik}|\mu_i^S, v) = \frac{v}{\mu_i^{SI}\Gamma(v)}\Big(\frac{vs_{ik}}{\mu_i^{SI}}\Big)^{v-1}\exp\Big(-\frac{vs_{ik}}{\mu_i^{SI}}\Big).$$

We consider a regression on the mean μ_i^{SI} including covariates $\mathbf{w}_i$ and spatial effects $\boldsymbol{\zeta}^I$ for the J geographical regions. By choosing a log link we obtain

$$\mu_i^{SI} = \exp(\mathbf{w}_i'\boldsymbol{\alpha}^I + \zeta^I_{R_{(i)}}), k = 1, .., N_i, i = 1, .., n. \tag{7.2}$$

Here $\mathbf{w}_i = (1, w_{i1}, .., w_{ip})'$ denotes the vector of covariates for the i-th observation including an intercept, $\boldsymbol{\alpha}^I = (\alpha_0^I, \alpha_1^I, .., \alpha_p^I)'$ the vector of unknown regression parameters and $\boldsymbol{\zeta}^I = (\zeta_1^I, .., \zeta_J^I)$ the vector of spatial effects which are modelled by the CAR prior specified in Chapter 4, like in the Poisson model for claim frequency. Since we consider a model for the individual claim sizes conditionally on the number of claims, the observed number of claims N_i may be introduced as a covariate as well. The number of claims per policyholder observed in car insurance data typically is very low, therefore we include N_i as a factor covariate with reference level $N_i = 1$. Hence, including number of claims effects denoted by $\alpha^I_{N_i=k}, k = 2, .., \max_i N_i$, the mean μ_i^{SI} reads as follows

$$\begin{aligned} \mu_i^{SI} &= \exp(\mathbf{w}_i'\boldsymbol{\alpha}^I + \zeta^I_{R_{(i)}}) \\ &= \exp(\sum_{k=0}^{p} w_{ik}\alpha_k^I + \sum_{k=2}^{\max_i N_i} n_{ki}\alpha^I_{N_i=k} + \zeta^I_{R_{(i)}}), k = 1, .., N_i, i = 1, .., n, \end{aligned} \tag{7.3}$$

where $n_{ki} = \begin{cases} 1, & N_i = k \\ 0, & \text{otherwise} \end{cases}$.

Additional variability between policyholders might be modelled by replacing, the number of claims effects $\alpha^I_{N_i=k}, k = 1, .., \max_i N_i$ in (7.3) with normal distributed policyholder specific random effects $c_i|N_i \sim N(\alpha^I_{N_i=k}, \sigma^2_{N_i=k})$ centered at $\alpha^I_{N_i=k}$ and variance $\sigma^2_{N_i=k}$ depending on the number of observed claims. However, this model extension did not improve the model fit for the data considered in the application presented in Chapter 8 and therefore will not be further persecuted.

7.2 Modelling average claim sizes

Alternatively, the average claim sizes $S_i, i = 1, .., n$ can be modelled. The average claim size S_i for policyholder i is given by

$$S_i := \sum_{k=1}^{N_i} \frac{S_{ik}}{N_i}.$$

Since we assume the $S_{ik}|N_i, k = 1, .., N_i$ to be independent and identically distributed, the average claim size S_i given the observed number of claims N_i is again Gamma distributed with mean $E(S_i|N_i) = \mu_i^{SA} = \mu_i^{SI}$ and variance $Var(S_i|N_i) = \frac{(\mu_i^{SA})^2}{N_i v}$, that is we have

$$S_i|N_i \sim Gamma(\mu_i^{SA}, N_i v). \tag{7.4}$$

Again a regression is performed on the mean μ_i^{SA}, both covariates and spatial effects are taken into account. Note, that the covariates $\tilde{\mathbf{w}} = (1, \tilde{w}_1, .., \tilde{w}_{\tilde{p}})$ do not necessarily have to be the same ones as for the individual claim sizes and that the estimated regression parameters and spatial effects for the average claim size will differ from the ones for the individual claim sizes. To point this out, we use the index A for the mean μ_i^{SA}, the regression parameters $\boldsymbol{\alpha}^A$ and the spatial effects $\boldsymbol{\zeta}^A$ in the model for average claim sizes. With a log link we obtain the following model

$$\mu_i^{SA} = \exp(\tilde{\mathbf{w}}_i' \boldsymbol{\alpha}^A + \zeta_{R(i)}^A).$$

Here again, number of claims effects $\alpha_{N_i=k}^A, k = 1, .., \max_i N_i$, can be included leading to

$$\begin{aligned} \mu_i^{SA} &= \exp(\tilde{\mathbf{w}}_i' \boldsymbol{\alpha}^A + \zeta_{R_{(i)}}^A) \\ &= \exp(\sum_{k=0}^{\tilde{p}} \tilde{w}_{ik} \alpha_k^A + \sum_{k=2}^{\max_i N_i} n_{ki} \alpha_{N_i=k}^A + \zeta_{R_{(i)}}^A), k = 1, .., N_i, i = 1, .., n, \end{aligned} \tag{7.5}$$

with n_{ki} defined as in the previous section.

Prior distributions for claim size models

Similar to the models for the number of claims (see Section 5.2) we have little prior knowledge on the regression parameters $\boldsymbol{\alpha}^I$ and $\boldsymbol{\alpha}^A$ in the models for individual and average claim sizes, respectively. Therefore we assume a normal prior with large standard deviation, in particular,

$$\boldsymbol{\alpha}^I \sim N_{p+\max_i N_i}(0, \tau^2 I_{p+\max_i N_i}),\ \boldsymbol{\alpha}^A \sim N_{\tilde{p}+\max_i N_i}(0, \tau^2 I_{\tilde{p}+\max_i N_i})$$

with $\tau^2 = 100$ which is a common choice. For the scale parameter v the gamma prior $v|a,b \sim Gamma(a,b)$, i.e. $\pi(v|a,b) = \frac{b^a}{\Gamma(a)} v^{a-1} \exp(-vb)$ with $a = 1$ is assumed, resulting in the conditional mean and variance given by $E(v|a=1,b) = \frac{1}{b}$ and $Var(v|a=1,b) = \frac{1}{b^2}$. Following a fully Bayesian approach we also assign a noninformative gamma prior to the hyperparameter b, in particular $b|c,d \sim Gamma(c,d)$, i.e. $\pi(b|c,d) = \frac{d^c}{\Gamma(c)} b^{c-1} \exp(-bd)$ with $c = 1$ and $d = 0.005$, resulting in $E(b|c,d) = 200$ and $Var(b|c,d) = 40000$. However, the models turn out to be very robust with respect to the prior on b, a very similar estimated posterior mean of v is obtained when b is fixed to 0.005, which is a popular choice for a flat gamma prior.
The spatial effects are modeled using the CAR prior described in Chapter 4, i.e.

$$\boldsymbol{\zeta}^l|\sigma^2,\psi \sim N_J(\mathbf{0}, \sigma^2 Q^{-1}),\ l = I, A,$$

with the elements of Q defined as in (4.3). For the hyperparameters σ^2 and ψ the same prior distributions as in Section 5.2 are assumed. The hyperparameter b can be sampled directly from a Gamma distribution. For the regression parameters, the spatial hyperparameters and the scale parameter v a single component MH algorithm with symmetric random walk is used. The spatial effects are updated component by component using an independence MH sampler with a t-proposal distribution with the same mode and inverse curvature at the mode as the target distribution, see Section 2.2.2.

7.3 Posterior predictive distribution of the total claim size

The distribution of the total claim sizes is not available analytically, but can be determined numerically using recursion formulas going back to Panjer (1981) when independence of claim size and claim frequency is assumed. In our approach the independence assumption is violated, however, based on the MCMC output of the models for the number of claims and claim size the posterior predictive distribution of the total claim size can be approximated. For this independence of claim size and the number of claims is not required. In the following we describe how the total claim size $S_i^T = \sum_{k=1}^{N_i} S_{ik}$ for policyholder $i = 1,..,n$ can be simulated based on the MCMC output. Let $\hat{\boldsymbol{\beta}}^{Nj}$, $\hat{\boldsymbol{\gamma}}^{Nj}$,$\hat{\boldsymbol{\alpha}}^{Aj}$, $\hat{\boldsymbol{\zeta}}^{Aj}$, $\hat{\boldsymbol{\alpha}}^{Ij}$, $\hat{\boldsymbol{\zeta}}^{Ij}$, $j = 1,..,R$ denote the MCMC draws after burnin for the regression parameters and spatial effects of the claim frequency and claim size models, respectively. The quantities $\hat{v}^{Aj}$ and $\hat{v}^{Ij}$ denote the MCMC draws of v in the Gamma model for average and individual

claim sizes, respectively. Then, for models based on the average claim sizes, proceed as follows. For $j = 1, .., R$

- simulate $N_i^j \sim Poisson(\hat{\mu_i}^{Nj})$ where $\hat{\mu_i}^{Nj} := t_i \exp(\mathbf{x}_i^{N\prime}\hat{\boldsymbol{\beta}}^{Nj} + \hat{\gamma}_{R(i)}^{Nj})$
- if $N_i^j = 0$ set $S_i^{Tj} = 0$
- otherwise simulate:
  $S_i^j \sim Gamma(\hat{\mu}_i^{SAj}, \hat{v}^{Aj} N_i^j)$ where $\hat{\mu_i}^{SAj} := \exp(\tilde{\mathbf{w}}_i'\hat{\boldsymbol{\alpha}}^{Aj} + \hat{\zeta}_{R(i)}^{Aj})$ and set $S_i^{Tj} = N_i^j \cdot S_i^j$

Based on the models for individual claim sizes, the simulation of the total claim size changes to: for $j = 1, .., R$

- simulate $N_i^j \sim Poisson(\hat{\mu_i}^{Nj})$ where $\hat{\mu_i}^{Nj} := t_i \exp(\mathbf{x}_i^{N\prime}\hat{\boldsymbol{\beta}}^{Nj} + \hat{\gamma}_{R(i)}^{Nj})$
- if $N_i^j = 0$ set $S_i^{Tj} = 0$
- otherwise simulate for $k = 1, .., N_i^j$:
  $S_{ik}^j \sim Gamma(\hat{\mu}_i^{SIj}, \hat{v}^{Sj})$ where $\hat{\mu_i}^{SIj} := \exp(\mathbf{w}_i'\hat{\boldsymbol{\alpha}}^{Ij} + \hat{\zeta}_{R(i)}^{Ij})$ and set $S_i^{Tj} = \sum_{k=1}^{N_i^j} S_{ik}^j$

Thus, a sample $S_i^{Tj}, j = 1, .., R$ of the total claim size S_i^T is obtained by which the posterior predictive distribution of S_i^T can be approximated.

Chapter 8

Application to German car insurance data

The models described in the previous chapter will now be used to analyse individual and average claim sizes in a data set from a German car insurance company. Since we aim to obtain the posterior predictive distribution of the total claim size, we also analyse the number of claims in this data set. This will be done using some of the models discussed in Chapter 5.

Our main questions of interest for this application are the followings: does the inclusion of spatial effects improve the models and can we observe a spatial pattern for the expected number of claims and the expected claim sizes? Do the individual and average claim sizes of a policyholder depend on the number of observed claims, i.e. are there significant number of claims effects for the models for claim size? Based on the models for the number of claims and claim size, we will finally approximate the posterior predictive distribution of the total claim sizes. Here again, we are interested, to what extent the inclusion of spatial and claim number effects influences the total claim sizes. Should we prefer models for the individual or the average claim sizes?

Models are compared using several criteria like the deviance information criterion (DIC), the predictive model choice criterion (PMCC) and several scoring rules discussed in Chapter 3. The inclusion of spatial effects leads to a significantly improved model fit both for claim frequency and claim size, more accurate predictions of the total claim sizes are obtained. When spatial effects are neglected the posterior predictive means of the total claim sizes in some regions with particular high (low) observed total claims are estimated considerably lower (higher) than based on the spatial models. Further, effects for the number of claims turn out to be significant for the claim size models, for an increasing

number of claims, both average and individual claim sizes tend to decrease. The results for the average and individual claim sizes are very similar, therefore modelling of average claim sizes seems to be sufficient.

8.1 Data description

The data set contains information about policyholders in Germany with full comprehensive car insurance within the year 2000. Not all policyholders were insured during the whole year, however the exposure time t_i of each policyholder is known. Several covariates like age and gender of the policyholder, kilometers driven per year, type of car and age of car are given in the data. The deductible which differs between policyholders will also be included as covariate. Germany is divided into 440 regions, for each policyholder the region he is living in is known. We analyse a subset of these data, in particular we only consider traffic accident data for policyholders with three types of midsized cars. The resulting data set contains about 350000 observations. There is a very large amount of observations with no claim in the data set and the maximum number of observed claims is only 4, see Table 8.1.

number of claims	percentage of observations
0	0.960
1	0.039
2	$6.9 \cdot 10^{-4}$
3	$1.7 \cdot 10^{-5}$
4	$2.8 \cdot 10^{-6}$

Table 8.1: Summary of the observed claim frequencies in the data.

The histogram of the observed positive individual claim sizes in DM given in Figure 8.1 reveals that the distribution of the claim sizes is highly skewed. The average individual claim size is given by DM 5371.0, the largest observed claim size takes the value DM 49339.1 which is less than 0.01 % of the sum of all individual claim sizes. Therefore, the data do not contain extreme values giving rise for the use of heavy tailed distributions and the Gamma model seems to be justified.
For an increasing number of observed claims, the average individual claim sizes decrease, see Table 8.2, indicating a negative correlation between claim size and number of claims.

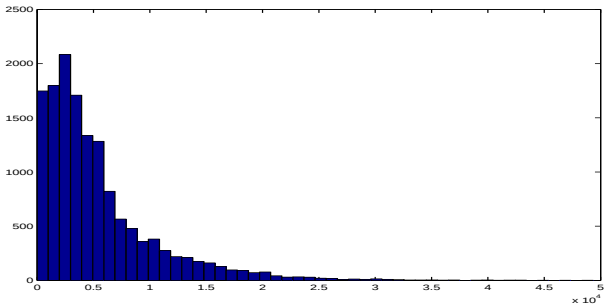

Figure 8.1: Histogram of the observed positive individual claim sizes.

	all observations	$N_i = 1$	$N_i = 2$	$N_i = 3$	$N_i = 4$
mean	5371.0	5389.9	4403.8	3204.2	330.5

Table 8.2: Mean of the observed individual claim sizes taken over all observations and over observations with $N_i = k, k = 1, 2, 3, 4$ observed claims separately.

8.2 Modelling claim frequency

We first address the modelling of claim frequency. In order to identify significant covariates and interactions, the data set is analysed in Splus first using a Poisson model without spatial effects. The obtained covariates specification is then used as a starting specification for our MCMC algorithms. An intercept, seventeen covariates like age, gender of the policyholders or mileage and interactions were found to be significant for explaining claim frequency. However, for reasons of confidentiality no details about these covariates and their effects will be reported. In order to obtain low correlations between covariates, we use centered and standardized covariates throughout the whole application. Since Germany is divided into 440 irregular spaced regions, 440 spatial effects are introduced for the MCMC analysis. We are interested in spatial effects after adjusting for population effects, therefore the population density in each of these regions is included as covariate as well. In particular the population density is considered on a logarithmic scale which turned out to give the best fit in the initial Splus analysis.

We assume the Poisson, generalized Poisson (GP) and zero inflated Poisson (ZIP) regression models presented in Chapter 5. All models include the same covariates and interactions. The same prior distributions (see Section 5.2) and MCMC algorithms as in Chapter 5 are used. The MCMC algorithms are run for 10000 iterations, a burnin of 1000 iterations is found to be sufficient after investigation of the MCMC trace plots. Both models including and without spatial effects, all containing the same covariates, are

assumed. The estimated posterior means of the model specific parameters λ in the GP and p in the ZIP models as well as summary statistics on the estimated dispersion factors $\hat{\varphi} := \frac{1}{R}\sum_{j=1}^{R}\frac{1}{(1-\hat{\lambda}^j)}$ in the GP and $\hat{\varphi} := \frac{1}{R}\sum_{j=1}^{R}(1+\hat{p}^j\hat{\mu}_i^j)$ in the ZIP model are reported in Table 8.3. Here $\hat{\lambda}^j$, $\hat{p}^j$ and $\hat{\mu}_i^j = t_i \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}^j)$ denote the j-th MCMC iterates, R gives the number of recorded MCMC iterations after burnin. The overdispersion parameter λ in the GP model without and including spatial effects is estimated very close to zero, the estimated posterior mean of the dispersion factor takes the value 1.0004, indicating that no overdispersion is present in these data. The proportion of extras zeros p in the ZIP models is estimated by only 2.61 % and 3.05 % in the models including and without spatial effects. Therefore, despite the large amount of zero observations in the data, most of them are explained as zeros arising from the underlying Poisson model. However, the dispersion factors in the ZIP models are estimated notedly higher than in the GP model, i.e. the variability of the data is estimated higher in the ZIP models.

model	γ	model specific parameters	$\hat{\varphi}_i$ mean	min	25 %	50 %	75 %	max
GP	no	$\hat{\lambda}: 1.80 \cdot 10^{-4}$ $(4.40 \cdot 10^{-6}, 6.70 \cdot 10^{-4})$	1.0004					
GP	yes	$\hat{\lambda}: 1.97 \cdot 10^{-4}$ $(7.74 \cdot 10^{-6}, 7.19 \cdot 10^{-4})$	1.0004					
ZIP	no	$\hat{p}$: 0.0305 $(0.0010, 0.0886)$	1.0639	1.0626	1.0634	1.0639	1.0643	1.0782
ZIP	yes	$\hat{p}$: 0.0261 $(0.0007, 0.0902)$	1.0548	1.0537	1.0543	1.0547	1.0551	1.0663

Table 8.3: Estimated posterior means of the model specific parameters λ in the GP model and p in the ZIP model with 95% credible intervals given in brackets and summary statistics for the estimated model specific dispersion parameters $\hat{\varphi}_i$.

model	$\min_i \hat{\gamma}_i$	$\max_i \hat{\gamma}_i$
Poisson	-0.441	0.285
GP	-0.430	0.282
ZIP	-0.429	0.279

Table 8.4: Range of the estimated posterior means of the spatial effects in the Poisson, GP and ZIP model.

The range of the estimated posterior means of the spatial effects $\hat{\gamma}_i = \frac{1}{R}\sum_{j=1}^{R}\hat{\gamma}_i^j, i =$

$1, .., J$, given in Table 8.4, is almost the same in all models and only slightly decreases in the GP and ZIP model compared to the Poisson model.

We compare models using the DIC and the PMCC, see Table 8.5. The DIC is about the same for all spatial models, giving no preference to the GP and ZIP models. However, when spatial effects are neglected, the DIC values increase significantly for all models, indicating that significant spatial effects are present. This shows that, after taking the information given by the covariates into account, there is still some unexplained spatial heterogeneity present in the data which is captured by the spatial effects. Although the inclusion of the population density in each region allows for geographic differences already, spatial random effects still have a significant influence on explaining the expected number of claims. The values of the PMCC also decrease for the spatial models compared to the non spatial ones, the lowest value is obtained for the spatial Poisson model. Note, that the first term of the PMCC representing the model fit is about the same for all spatial and non spatial models, respectively. The second term of the PMCC in contrast, is considerably higher for the ZIP models, indicating that the variability estimated in the ZIP model is rather high. This concurs with the estimated dispersion factors which are high in the ZIP model, compared to the GP model.

Hence, the spatial Poisson regression model seems to be sufficient for these data, apart from the unobserved spatial heterogeneity captured by the spatial effects no overdispersion could be detected in the data. Therefore, in the following only the Poisson model will be considered.

model	$\boldsymbol{\gamma}$	DIC	$E[D(\boldsymbol{\theta}\|\mathbf{y})]$	p_D	PMCC	$\sum_{i=1}^{n}(\mu_i - y_i)^2$	$\sum_{i=1}^{n}\sigma_i^2$
Poisson	no	122372	122356	16.0	28624	14297	14328
Poisson	yes	122143	122045	98.2	28613	14280	14334
GP	no	122374	122358	16.1	28629	14297	14332
GP	yes	122144	122048	96.4	28617	14280	14337
ZIP	no	122374	122358	15.9	29547	14297	15250
ZIP	yes	122142	122052	90.8	29399	14281	15117

Table 8.5: DIC, posterior mean of the deviance $D(\boldsymbol{\theta}|\mathbf{y})$, effective number of parameters p_D and PMCC, split in its two components, for the considered models for claim frequency with and without spatial effects $\boldsymbol{\gamma}$.

The left panel in the top row in Figure 8.2 shows a map of the estimated posterior means of the spatial effects in the Poisson model. The corresponding posterior means of the risk

factors $\exp(\gamma_i)$ for the minimum and maximum spatial effects are given by 0.65 and 1.34. A trend from the east to the west of Germany is visible, the risk for claims tends to be lower in the east and increases towards the south western regions. A map of the 80 % credible intervals for the spatial effects is given by the right panel. For the eastern and south western regions significant spatial effects are present, whereas the spatial effects for the regions in the middle of Germany do not significantly differ from zero.

8.3 Modelling average claim sizes

In this section the average claim sizes $S_i := \sum_{k=1}^{N_i} S_{ik}$ are analysed using the spatial Gamma regression Model (7.4), i.e. $S_i|N_i \sim Gamma(\mu_i^{SA}, N_i v)$ with mean specification (7.5). Considering only observations with a positive number of claims, altogether 14066 observations are obtained. Again, significant covariates and interactions are identified by analysing the data in Splus first, assuming a Gamma model without spatial effects. An intercept and fourteen covariates including gender, type and age of car as well as the population density in each region, modelled as polynomial of order four, have been found to have significant influence. Further, the observed number of claims N_i is included as covariate. Since the highest number of claims is four, the number of claims is treated as a factor with three levels where $N_i = 1$ is taken as reference level. These covariates will be taken into account when analysing the data set using MCMC. Therefore, including spatial and number of claims effects the mean μ_i^{SA} is specified by

$$\mu_i^{SA} = \exp(\tilde{\mathbf{w}}_i'\boldsymbol{\alpha}^A + \zeta_{R(i)}^A) = \exp(\sum_{k=1}^{15} \tilde{w}_{ik}\alpha_k^A + n_{2i}\alpha_{N_i=2}^A + n_{3i}\alpha_{N_i=3}^A + n_{4i}\alpha_{N_i=4}^A + \zeta_{R(i)}^A)$$

where $n_{ki} = \begin{cases} 1, & N_i = k \\ 0, & \text{otherwise} \end{cases}$. We consider both models with and without number of claims effects $\alpha_{N_i=k}^A, k = 2, 3, 4$ in the following, to quantify the influence of these effects. The MCMC sampler for Model (7.4) including and without spatial effects and the observed number of claims is run for 10000 iterations. Again a burnin of 1000 is found to be sufficiently large. Models are compared using the DIC, the PMCC and some of the scoring rules given in Section 3. The lowest value of the DIC is obtained for the model both including N_i and spatial effects (see Table 8.6). Although the increase of the estimated effective number of parameters is very small when the number of claims is included as covariate, the values of the posterior mean of the deviance decreases by 28 and 40 in the model with and without spatial effects respectively, indicating a significant improvement.

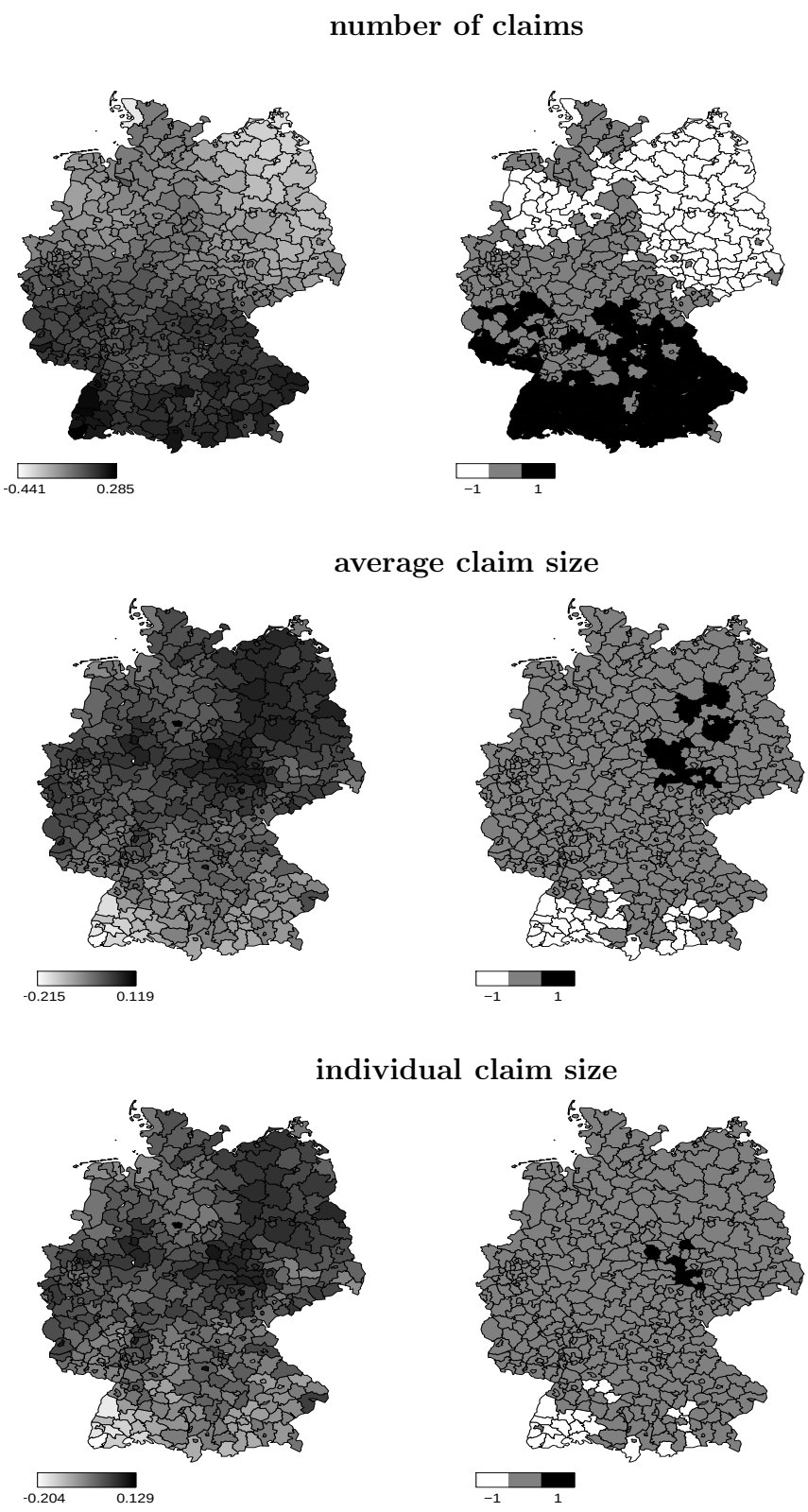

Figure 8.2: Map of the estimated posterior means (left) together with map of the 80 % credible intervals (right) for the spatial effects in the Poisson (top row), average (middle row) and individual (bottom row) claim size regression model. For grey regions, zero is included in the credible interval, black regions indicate strictly positive, white regions strictly negative credible intervals.

average claim size models

model with $\alpha^A_{N_i=k}$	model with $\boldsymbol{\zeta}^A$	DIC	$E[D(\boldsymbol{\theta}\|\mathbf{y})]$	p_D
yes	yes	269092	269020	72.7
yes	no	269136	269119	16.5
no	yes	269122	269048	73.9
no	no	269175	269159	15.9

$\alpha^A_{N_i=k}$	$\boldsymbol{\zeta}^A$	LS	$IS_{\alpha=0.5}$	$IS_{\alpha=0.1}$	CRPS	PMCC	$\sum_{i=1}^n(\mu_i - y_i)^2$	$\sum_{i=1}^n \sigma_i^2$
yes	yes	-9.5642	-11152	-4082.4	-2471.9	$7.332 \cdot 10^{11}$	$3.540 \cdot 10^{11}$	$3.792 \cdot 10^{11}$
yes	no	-9.5699	-11225	-4105.2	-2481.8	$7.363 \cdot 10^{11}$	$3.575 \cdot 10^{11}$	$3.788 \cdot 10^{11}$
no	yes	-9.5669	-11161	-4086.7	-2474.2	$7.290 \cdot 10^{11}$	$3.538 \cdot 10^{11}$	$3.752 \cdot 10^{11}$
no	no	-9.5734	-11234	-4114.9	-2484.3	$7.321 \cdot 10^{11}$	$3.580 \cdot 10^{11}$	$3.741 \cdot 10^{11}$

individual claim size models

model with $\alpha^I_{N_i=k}$	model with $\boldsymbol{\zeta}^I$	DIC	$E[D(\boldsymbol{\theta}\|\mathbf{y})]$	p_D
yes	yes	273888	273803	85.1
yes	no	273935	273914	21.2
no	yes	273961	273883	78.1
no	no	274009	273991	17.7

$\alpha^I_{N_i=k}$	$\boldsymbol{\zeta}^I$	LS	$IS_{\alpha=0.5}$	$IS_{\alpha=0.1}$	CRPS	PMCC	$\sum_{i=1}^n(\mu_i - y_i)^2$	$\sum_{i=1}^n \sigma_i^2$
yes	yes	-9.5539	-11129	-4072.3	-2467.6	$7.428 \cdot 10^{11}$	$3.596 \cdot 10^{11}$	$3.832 \cdot 10^{11}$
yes	no	-9.5600	-11206	-4101.6	-2478.5	$7.452 \cdot 10^{11}$	$3.636 \cdot 10^{11}$	$3.816 \cdot 10^{11}$
no	yes	-9.5570	-11157	-4087.2	-2473.8	$7.432 \cdot 10^{11}$	$3.611 \cdot 10^{11}$	$3.822 \cdot 10^{11}$
no	no	-9.5628	-11232	-4106.7	-2483.6	$7.468 \cdot 10^{11}$	$3.650 \cdot 10^{11}$	$3.819 \cdot 10^{11}$

Table 8.6: DIC, posterior mean of the deviance $D(\boldsymbol{\theta}|\mathbf{y})$, effective number of parameters p_D and mean score $S_n(\boldsymbol{\theta})$ for scoring rules LS, IS_α ($\alpha = 0.5, 0.1$), CRPS and PMCC, split in its two components, for the models including and without spatial and claim number effects. In the top tables results for the average claim size models, in the lower tables results for the individual claim size models are given.

The results for the scoring rules and the PMCC, divided into its two components, are reported in Table 8.6 as well. The computation of the DIC, the PMCC and the scores is based on 5000 iterations of the MCMC output, the first 5000 iterations are neglected. Note, that the computation of the interval score IS_α and the continuous ranked probability score CRPS is based on simulated data, whereas the logarithmic score LS and the PMCC are calculated directly using the MCMC output. For the logarithmic score LS, the interval score IS_α and the CRPS the highest mean score $S_n(\boldsymbol{\theta})$ is obtained for the model including spatial effects and number of claims effects. This confirms the significance of spatial and number of claims effects. According to the negatively oriented PMCC the spatial models also are to be preferred to the non-spatial ones. However, lower values of the PMCC are obtained for the models without number of claims effects which is mainly caused by the second term of the PMCC. Here it should be kept in mind, that the PMCC is not a proper scoring rule as noted in Section 3.2.3.

A map of the posterior means and the 80 % credible intervals of the spatial effects for the model both including spatial effects and N_i is given in the middle row in Figure 8.2. Similar results are obtained for the model without N_i. The estimated posterior means of the risk factors $\exp(\gamma_i)$ for the minimum and maximum spatial effects range from 0.81 to 1.13. Contrary to the estimated spatial effects for claim frequency, the average claim size tends to be higher for some regions in the east of Germany, whereas for regions in the south western part lower claim sizes are to be expected. Again, according to the 80 % credible intervals, the spatial effects are only significant for some regions in the east and the south west of Germany.

The estimated posterior means together with 95 % credible intervals of the number of claims effects $\alpha^A_{N_i=k}, k = 2, 3, 4$ are reported in Table 8.7. For an easier interpretation of the results we also give the estimated posterior means of the factors $\exp(\alpha^A_{N_i=k}), k = 2, 3, 4$ which quantify the relative risk in contrast to observations with the same covariates but only one observed claim. Compared to a policyholder with one observed claim, the expected average claim size for an observation with two observed claims decreases by about 25 %. If three or four claims have been reported, the expected average claim size even decreases by about 75 % and 92 %, respectively. This is illustrated graphically in Figure 8.3 where scatterplots of the estimated posterior means $\hat{\mu_i}^{SA}$ in the model without N_i (x-axis) to the ones in the model including N_i (y-axis) are given. In the left panel the scatterplot for all observations with one claim is given, while in the middle and right panel the corresponding plots for all observations with two and three observed claims, respectively, are shown. Since only one observation with four claims is given in the data,

average claim size models

parameter	posterior mean of $\alpha^A_{N_i=k}$	posterior mean of $\exp(\alpha^A_{N_i=k})$
$\alpha^A_{N(i)=2}$	-0.295	0.745
	$(-0.382, -0.203)$	$(0.683, 0.817)$
$\alpha^A_{N(i)=3}$	-1.951	0.146
	$(-2.376, -1.462)$	$(0.093, 0.232)$
$\alpha^A_{N(i)=4}$	-2.642	0.082
	$(-3.473, -1.544)$	$(0.031, 0.214)$

individual claim size models

parameter	posterior mean of $\alpha^I_{N(i)=k}$	posterior mean of $\exp(\alpha^I_{N_i=k})$
$\alpha^I_{N(i)=2}$	-0.237	0.789
	$(-0.305, -0.167)$	$(0.737, 0.846)$
$\alpha^I_{N(i)=3}$	-0.559	0.575
	$(-0.756, -0.349)$	$(0.470, 0.705)$
$\alpha^I_{N(i)=4}$	-2.636	0.0842
	$(-3.507, -1.501)$	$(0.030, 0.223)$

Table 8.7: Estimated posterior means of the number of claims effects and the risk factors $\exp(\alpha^l_{N(i)=k}), l = I, A$ in the Gamma model for average (top) and individual (bottom) claim sizes including spatial effects, with the 95 % credible intervals given in brackets.

the scatterplot for $N_i = 4$ is omitted. The estimation of the expected average claim sizes for the observations with one claim is very close in the models with and without number of claims effects, which is to be expected since $N_i = 1$ is the baseline. The means μ_i^{SA} for observations with two or three claims in contrast are estimated notedly smaller when number of claims effects are taken into account.

8.4 Modelling individual claim sizes

Now, the individual claim sizes $S_{ik}, k = 1, .., N_i, i = 1, .., n$ will be analysed directly using Model (7.1), i.e. $S_{ik}|N_i \sim Gamma(\mu_i^{SI}, v)$.This time we do not consider the data aggregated for each policyholder like in Section 8.3, but deal with N_i positive claims for every person, resulting in 14325 observations altogether. The data have again been

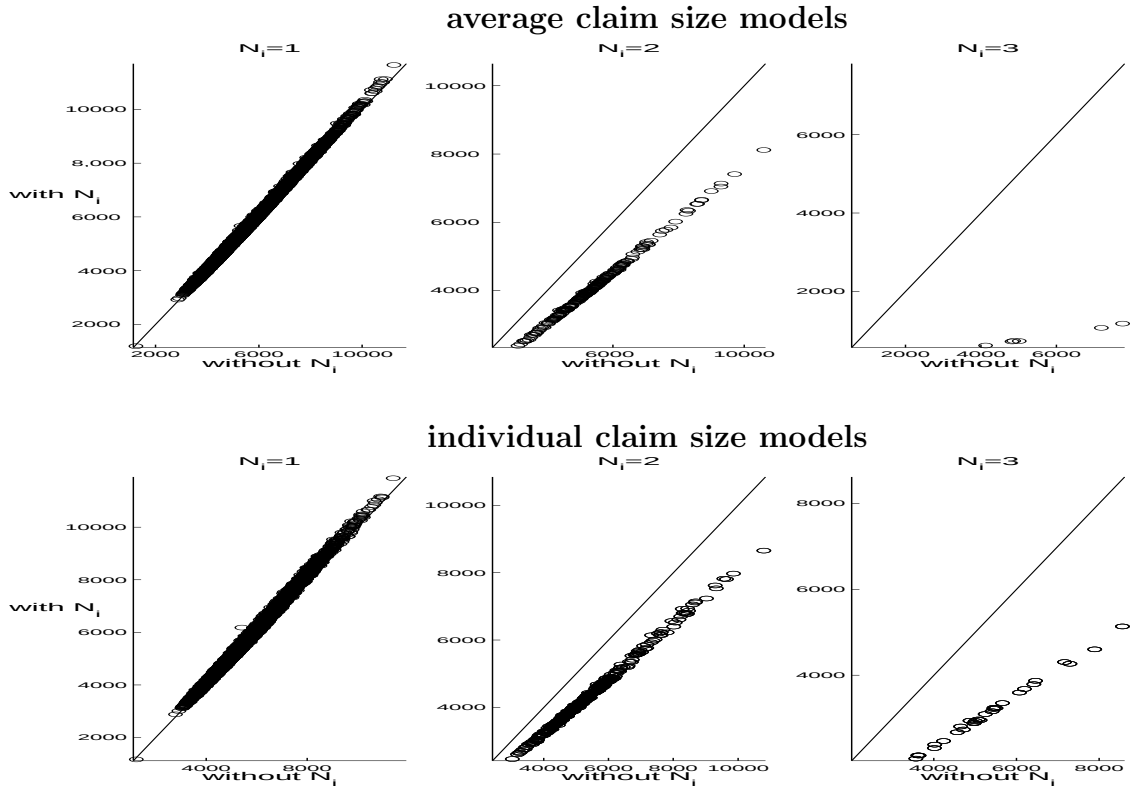

Figure 8.3: Scatterplots of the estimated posterior means $\hat{\mu}_i^{SA}$ in the spatial Gamma models for average (top) and individual (bottom) claim sizes with and without number of claims effects, the solid line gives the 45 degree line.

analysed in Splus first using a Gamma model without taking spatial effects into account. Altogether an intercept and sixteen covariates, the same ones as for the average claim size and two additional interaction terms, turned out to be significant for explaining the individual claim sizes. Again, we assume both models with and without the observed number of claims N_i as a factor covariate and including and without spatial effects. When both spatial and claim number effects are included, the mean μ_i^{SI} reads

$$\mu_i^{SI} = \exp(\mathbf{w}_i'\boldsymbol{\alpha}^I + \zeta_{R(i)}^I) = \exp(\sum_{k=1}^{17} w_{ik}\alpha_k^I + n_{2i}\alpha_{N_i=2}^I + n_{3i}\alpha_{N_i=3}^I + n_{4i}\alpha_{N_i=4}^I + \zeta_{R(i)}^I).$$

10000 iterations of the MCMC algorithm are performed and a burnin of 1000 is taken. According to the DIC values for these models given in Table 8.6, the model fit is significantly improved by including the number of claims as covariate. Taking spatial effects into account additionally gives a better model fit. The results for the scoring rules and the PMCC are also given in Table 8.6. Confirming the results of the DIC, the highest

mean score of scoring rules LS, IS_α and CRPS and the lowest value of the PMCC is obtained for the model both including spatial effects and number of claims effects. The estimated posterior means of the spatial effects for the model including N_i together with the 80 % credible intervals are given in the bottom row in Figure 8.2. The results are very similar to the findings for the average claim size in the previous section, here the minimum and maximum posterior means of the risk factors $\exp(\gamma_i)$ are given by 0.82 and 1.14, respectively.

In Table 8.7 the estimated posterior means together with 95% credible intervals are reported for the number of claims effects $\alpha^I_{N_i=k}$ and the relative risk factors $\exp(\alpha^I_{N_i=k}), k = 2, 3, 4$, in the model including spatial effects. The effects are significant for all levels, indicating a decrease of the expected individual claim sizes for an increasing number of claims. The results are also illustrated in Figure 8.3 using scatterplots. Again, the estimated posterior means $\hat{\mu_i}^{SI}$ of the model means μ_i^{SI} based on the spatial model without claim number effects for observations with two or three claims are clearly higher than the ones based on the spatial model including the number of claims as covariate. Note, that in comparison to the results for the average claim size where the factor $\exp(\alpha^A_{N_i=3})$ for observations with three claims was estimated by 0.146, the factor $\exp(\alpha^I_{N_i=3})$ for the individual claim sizes is estimated by 0.575, resulting in a considerably lower decrease of the expected individual claim sizes. A direct comparison of the models for average and individual claim sizes based on the DIC, the PMCC or the scoring rules is not possible, since the average claim size models are based on aggregated data resulting in less observations. However, qualitatively the findings are very similar, in both approaches the best model fit is obtained when both spatial and claim number effects are taken into account.

8.5 Posterior predictive distribution of the total claim size

Based on the MCMC output of the models for the number of claims and claim size the posterior predictive distribution of the total claim size can be approximated, as described in Section 7.3. In order to compare the simulated total claim sizes S_i^T based on the different models for claim size and claim frequency, we compute the continuous ranked probability score CRPS and the predictive model choice criterion PMCC. The DIC and the logarithmic score cannot be computed here, since they are based on the explicit form of the total claim size distribution which is not available. The interval score will also

be omitted out of the following reasons: Due to the large amount of observations with zero claims in our data set, the percentage of simulations with total claim size equal to zero is also very high. Zero will be included in the $(1-\alpha)$ 100 % posterior predictive intervals of the total claim sizes for $\alpha = 0.5, 0.1$ for almost all observations. Therefore only observations falling above the upper quantiles of the prediction intervals would be considered as outliers and be penalized. Hence, the use of the interval score will not be appropriate any more. Instead we consider one-sided quantities here like the quantiles r_α at level $\alpha = 0.95, 0.99$ and the number of observations falling above these quantiles and compute the quantile score QS_α described in Section 3.2.3 for $\alpha = 0.95, 0.99$. Both the scores as well as the PMCC are computed using 5000 simulations of the total claim sizes S_i^T.

The results for the simulations based on models for the average and the individual claim sizes, reported in Tables 8.8 and 8.9, are qualitatively the same.

The PMCC favours the simulations based on the models including spatial effects for the number of claims only, further better results are achieved when number of claims effects are taken into account. This is caused especially by the second term of the PMCC, representing the model variances, which are considerably lower when the number of claims is included as covariate in the claim size models.

The mean scores for the CRPS and the quantile scores $QS_\alpha, \alpha = 0.95, 0.99$, are very close for all models, in general slightly higher scores are obtained for simulations based on a spatial Poisson model for the number of claims. Further, the simulations based on a spatial model for both claim frequency and claim size and including number of claims effects tend to achieve the highest score. The size of the quantiles seems to be mainly determined by the inclusion or neglect of spatial effects in the Poisson model for the number of claims. The quantiles at level $\alpha = 0.95$ are higher when spatial effects are included on the number of claims, reflecting a higher model complexity. The percentage of observations falling above the 95 % quantile ranges from 3.60 % to 3.65 %, lying below the expected 5 %. This might be caused by the fact noted already above. Since for some observations even the 95 % quantile will be only zero, a zero observation will not be regarded as an outlier. This might be overcome by randomizing zero observations, i.e. considering zero observations as outliers with a certain probability when the 95 % quantile takes the value zero. The 99 % quantiles in contrast, are slightly higher when no spatial Poisson models are assumed, the percentage of outliers is close to the expected 1 %. Comparing the results in Tables 8.8 and 8.9, no significant difference is observed for the CRPS and the quantile scores between the simulations based on average and individual claim sizes, respectively. When number

freq γ	size ζ^A	PMCC	$\sum_{i=1}^n(\mu_i - y_i)^2$	$\sum_{i=1}^n \sigma_i^2$
with $\alpha^A_{N_i=k}$				
yes	yes	$1.5757 \cdot 10^{12}$	$7.8032 \cdot 10^{11}$	$7.9540 \cdot 10^{11}$
no	no	$1.5735 \cdot 10^{12}$	$7.8069 \cdot 10^{11}$	$7.9279 \cdot 10^{11}$
yes	no	$1.5710 \cdot 10^{12}$	$7.8046 \cdot 10^{11}$	$7.9089 \cdot 10^{11}$
no	yes	$1.5856 \cdot 10^{12}$	$7.8088 \cdot 10^{11}$	$8.0477 \cdot 10^{11}$
without $\alpha^A_{N_i=k}$				
yes	yes	$1.5960 \cdot 10^{12}$	$7.8055 \cdot 10^{11}$	$8.1541 \cdot 10^{11}$
no	no	$1.5909 \cdot 10^{12}$	$7.8088 \cdot 10^{11}$	$8.1000 \cdot 10^{11}$
yes	no	$1.5894 \cdot 10^{12}$	$7.8072 \cdot 10^{11}$	$8.0871 \cdot 10^{11}$
no	yes	$1.6052 \cdot 10^{12}$	$7.8108 \cdot 10^{11}$	$8.2414 \cdot 10^{11}$

freq	size		95 %			99 %		
γ	ζ^A	CRPS	quantile	outliers	$QS_{0.95}$	quantile	outliers	$QS_{0.99}$
with $\alpha^A_{N_i=k}$								
yes	yes	-212.26	476.3	3.60 %	-205.2	6400.6	1.15 %	-134.9
no	no	-212.35	456.0	3.65 %	-205.7	6426.9	1.16 %	-135.5
yes	no	-212.27	480.7	3.60 %	-205.3	6393.5	1.17 %	-135.4
no	yes	-212.36	460.0	3.65 %	-205.8	6454.0	1.16 %	-135.5
without $\alpha^A_{N_i=k}$								
yes	yes	-212.30	473.6	3.60 %	-205.2	6433.6	1.16 %	-135.3
no	no	-212.30	453.8	3.65 %	-205.7	6455.1	1.16 %	-135.5
yes	no	-212.28	477.8	3.60 %	-205.3	6422.6	1.17 %	-135.7
no	yes	-212.34	456.2	3.65 %	-205.9	6482.1	1.16 %	-135.7

Table 8.8: In the upper table the PMCC, split in its two components is given for several models for the simulated total claim sizes S_i^T, based on models for average claim sizes. In the lower table the mean score $S_n(\boldsymbol{\theta})$ for the CRPS, the 95 % and 99 % quantiles, the percentage of observations lying above these quantiles and the corresponding quantile mean scores $QS_\alpha, \alpha = 0.95, 0.99$, are given.

of claims effects are included, better values of the PMCC are obtained when models for the average claim sizes are assumed. Without number of claims effects in contrast, the PMCC gives a slight preference to simulations based on individual claim sizes. Hence,

freq γ	size ζ^I	PMCC	$\sum_{i=1}^n(\mu_i - y_i)^2$	$\sum_{i=1}^n \sigma_i^2$
with $\alpha^I_{N_i=k}$				
yes	yes	$1.5869 \cdot 10^{12}$	$7.8078 \cdot 10^{11}$	$8.0607 \cdot 10^{11}$
no	no	$1.5830 \cdot 10^{12}$	$7.8103 \cdot 10^{11}$	$8.0196 \cdot 10^{11}$
yes	no	$1.5807 \cdot 10^{12}$	$7.8091 \cdot 10^{11}$	$7.9976 \cdot 10^{11}$
no	yes	$1.5960 \cdot 10^{12}$	$7.8132 \cdot 10^{11}$	$8.1473 \cdot 10^{11}$
without $\alpha^I_{N_i=k}$				
yes	yes	$1.5911 \cdot 10^{12}$	$7.8089 \cdot 10^{11}$	$8.1017 \cdot 10^{11}$
no	no	$1.5878 \cdot 10^{12}$	$7.8120 \cdot 10^{11}$	$8.0664 \cdot 10^{11}$
yes	no	$1.5865 \cdot 10^{12}$	$7.8101 \cdot 10^{11}$	$8.0550 \cdot 10^{11}$
no	yes	$1.6002 \cdot 10^{12}$	$7.8141 \cdot 10^{11}$	$8.1882 \cdot 10^{11}$

freq	size		95 %			99 %		
γ	ζ^I	CRPS	quantile	outliers	$QS_{0.95}$	quantile	outliers	$QS_{0.99}$
with $\alpha^A_{N_i=k}$								
yes	yes	-212.28	484.3	3.60 %	-205.5	6442.8	1.15 %	-135.0
no	no	-212.38	464.4	3.64 %	-206.0	6469.6	1.15 %	-135.3
yes	no	-212.35	489.3	3.60 %	-205.6	6434.0	1.16 %	-135.5
no	yes	-212.35	467.0	3.65 %	-206.2	6498.1	1.16 %	-135.5
without $\alpha^A_{N_i=k}$								
yes	yes	-212.27	477.2	3.61 %	-205.5	6427.2	1.16 %	-135.0
no	no	-212.33	457.2	3.65 %	-206.0	6453.8	1.16 %	-135.7
yes	no	-212.32	481.3	3.61 %	-205.6	6422.5	1.16 %	-135.5
no	yes	-212.36	459.6	3.65 %	-206.1	6481.6	1.16 %	-135.6

Table 8.9: In the upper table the PMCC, split in its two components is given for several models for the simulated total claim sizes S_i^T, based on models for individual claim sizes. In the lower table the mean score $S_n(\boldsymbol{\theta})$ for the CRPS, the 95 % and 99 % quantiles, the percentage of observations lying above these quantiles and the corresponding quantile mean scores $QS_\alpha, \alpha = 0.95, 0.99$, are given.

although more detailed data information is available for the individual claim size models, this additional knowledge has no significant influence on the prediction of the total claim sizes, at least when number of claims effects are taken into account.

In Figure 8.4 map plots of the observed total claim sizes and the posterior predictive means $\frac{1}{R}\sum_{j=1}^{R} S_i^{Tj}$ of the simulated total claim sizes, averaged over each region, are given. Since we only consider the posterior predictive mean of the simulated total claim sizes, it is natural that the map displaying the true total claim sizes shows more extreme values. Hence, for a better visual comparison of the maps, we have built six classes for the total claim size in these plots, assuming equal length for the four middle classes, but summarizing extremely small or high values in broader classes. The simulations are based on the models for average claim sizes with the number of claims included as covariate. The plots look very similar for the simulations based on the models for the individual claim sizes. When spatial effects are included in the Poisson model (middle row), an increasing trend from the east to the west is observable for the simulated total claim sizes. The additional inclusion of spatial effects for the average claim size leads to small changes in the very eastern and south western parts of Germany. The rough spatial structure of the observed total claim sizes (top) is represented reasonable well for these two models. However, if spatial effects are only included for the average claim sizes, the regions with high observed total claim sizes in the middle and south western parts of Germany are not detected. The same holds for the simulations based on the models without any spatial effects. For regions in the east of Germany with rather low true total claim sizes for example, the mean of the total claim sizes is estimated up to 1.27 times as high when no spatial effects at all are taken into account compared to a spatial modelling of claim frequency and claim size. For one south western region with large observed total claims in contrast, the posterior mean of the simulated total claim size based on non spatial models is only estimated 0.69 times as large compared to the simulations based on spatial models for claim frequency and claim size.

The estimated probabilities for the total claim sizes being equal to zero as well as density estimates of the positive simulated total claim sizes of the policyholders in the two regions Hannover and Lörrbach are given in Figure 8.5. For Hannover the largest posterior mean of the spatial effect in the average claim size model was estimated ($\hat{\zeta}^A = 0.12$), while in Lörrbach the smallest effect $\hat{\zeta}^A = -0.22$ was observed. The estimated posterior means of the spatial effects in the Poisson model for the number of claims are given by $\hat{\gamma} = -0.10$ in Hannover and $\hat{\gamma} = 0.29$ in Lörrbach. Figure 8.5 shows that the estimated probability for zero total claim sizes and the density estimates of the positive total claim sizes notedly change when spatial effects are included to the models for claim frequency and average claim size. In Hannover, the inclusion of spatial effects to the models for claim frequency and the average claim size leads to a higher estimated probability of zero total claim sizes

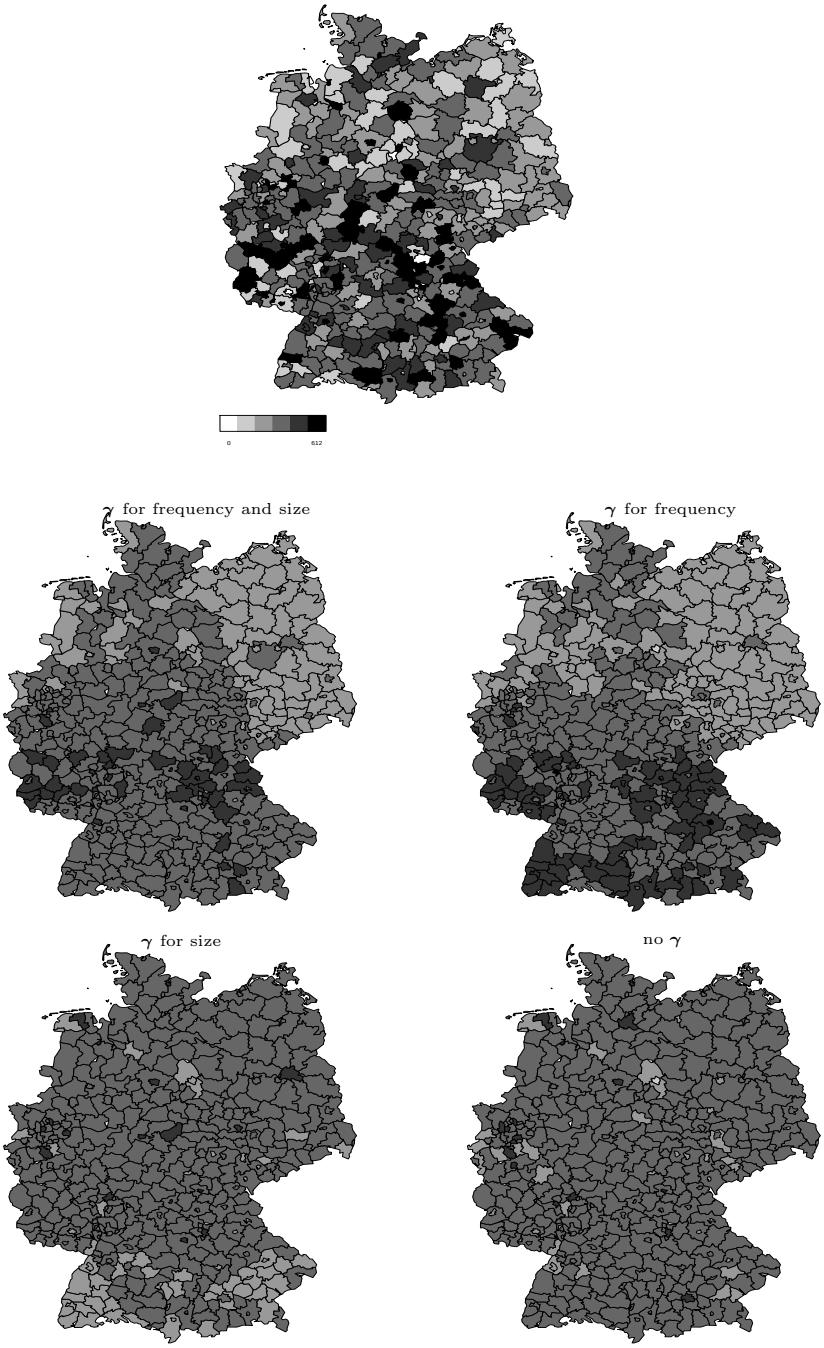

Figure 8.4: Observed total claim sizes (top) and simulated total claim sizes based on Poisson and Gamma models for average claim sizes with and without spatial effects. grey level classification: $[0, 100), [100, 150), [150, 200), [200, 250), [200, 300), [300, \infty)$

and heavier tails for the estimated density of the positive total claim sizes. In Loerrbach in contrast, the estimated probability for zero total claim sizes decreases and more mass is given to small positive total claim sizes when spatial effects are added.

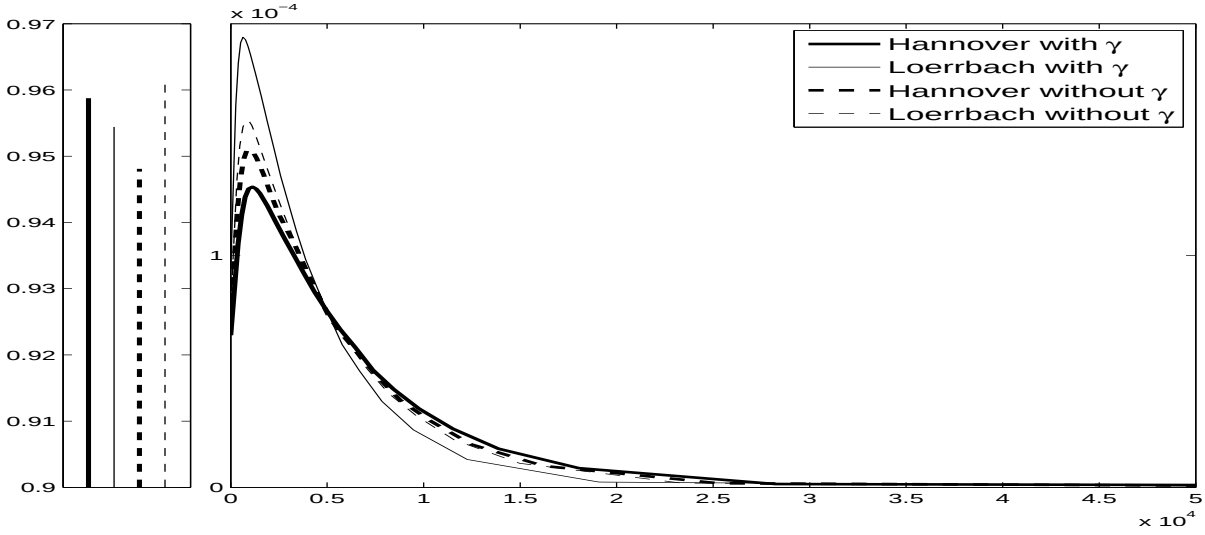

Figure 8.5: Estimated probabilities for zero total claim sizes (left panel) and density estimates of the positive total claim sizes (right panel) of the policyholders in the regions Hannover and Loerrbach based on spatial (solid lines) and non spatial (dashed lines) models for both claim frequency and average claim size.

Ideally, when the predictive quality of models is of interest, the data should not be used twice, i.e. parameter estimation should be based on part of the data only and predictions should be done for the remaining data. Since in this section model comparison rather than prediction was focussed, all data were used for parameter estimation and simulation of the total claim sizes. However, for the sake of completeness, we also fitted the Poisson models for claim frequency and the Gamma models for the average claim size based on 75 % of the data only and simulated the total claim sizes for the remaining 25 % of the data. Results for the PMCC, the CRPS and the quantile scores, reported in Table 8.10, are qualitatively the same as observed before. The mean scores are very close for all models, the quantile scores give a slight preference to simulations based on a spatial Poisson model. Note, that the mean scores take lower values now compared to the simulations based on all data. Further, about 9 % of the observations exceed the 95 % quantile, about 2.9 % fall above the 99 % quantile. This shows, that prediction of the true total claim sizes is worse here. However, this is to be expected, since the information given in these 25 % of the data has not been accounted for in parameter estimation.

freq γ	size ζ^A	PMCC	$\sum_{i=1}^n(\mu_i - y_i)^2$	$\sum_{i=1}^n \sigma_i^2$
with $\alpha^A_{N_i=k}$				
yes	yes	$4.0026 \cdot 10^{11}$	$2.0460 \cdot 10^{11}$	$1.9567 \cdot 10^{11}$
no	no	$4.0013 \cdot 10^{11}$	$2.0465 \cdot 10^{11}$	$1.9548 \cdot 10^{11}$
yes	no	$3.9932 \cdot 10^{11}$	$2.0460 \cdot 10^{11}$	$1.9470 \cdot 10^{11}$
no	yes	$4.0193 \cdot 10^{11}$	$2.0475 \cdot 10^{11}$	$1.9718 \cdot 10^{11}$
without $\alpha^A_{N_i=k}$				
yes	yes	$4.0472 \cdot 10^{11}$	$2.0467 \cdot 10^{11}$	$2.0005 \cdot 10^{11}$
no	no	$4.0447 \cdot 10^{11}$	$2.0468 \cdot 10^{11}$	$1.9979 \cdot 10^{11}$
yes	no	$4.0384 \cdot 10^{11}$	$2.0467 \cdot 10^{11}$	$1.9917 \cdot 10^{11}$
no	yes	$4.0627 \cdot 10^{11}$	$2.0482 \cdot 10^{11}$	$2.0145 \cdot 10^{11}$

freq	size		95 %			99 %		
γ	ζ^A	CRPS	quantile	outliers	$QS_{0.95}$	quantile	outliers	$QS_{0.99}$
with $\alpha^A_{N_i=k}$								
yes	yes	-213.73	470.3	8.9 %	-207.2	6376.2	2.9 %	-139.4
no	no	-213.75	451.4	9.0 %	-207.5	6411.0	2.9 %	-139.3
yes	no	-213.76	472.3	8.9 %	-207.3	6370.3	2.9 %	-139.3
no	yes	-213.82	452.0	9.0 %	-207.7	6247.5	2.9 %	-139.8
without $\alpha^A_{N_i=k}$								
yes	yes	-213.77	467.2	8.9 %	-207.3	6412.0	3.0 %	-139.3
no	no	-213.80	448.3	9.0 %	-207.4	6444.1	2.9 %	-139.4
yes	no	-213.72	469.3	8.9 %	-207.4	6403.0	2.9 %	-139.5
no	yes	-213.75	448.6	9.0 %	-207.8	6459.9	2.9 %	-140.6

Table 8.10: PMCC, split in its two components, mean score $S_n(\boldsymbol{\theta})$ for the CRPS, the 95 % and 99 % quantiles, the percentage of observations lying above these quantiles and the corresponding quantile mean scores $QS_\alpha, \alpha = 0.95, 0.99$, for several models for the simulated total claim sizes S_i^T, based on models for average claim sizes. Parameter estimation is based on 75 % of the data, total claim sizes are simulated for remaining 25 % of the data.

8.6 Summary and conclusions

We have presented a Bayesian approach for modelling claim frequency and claim size taking both covariates as well as spatial effects into account. For the number of claims

a Poisson model turned out to give a sufficient fit. For these data no significant overdispersion or the presence of extra zeros was detected. In contrast to the common approach where independence of the number of claims and claim size is assumed, we do not need this assumption. Instead, we have shown, that by including the observed number of claims as covariate for claim size, models for claim sizes are significantly improved. If for example a policyholder caused two or three claims, the expected average claim size decreases by about 25 and 75 %, respectively, compared to a policyholder with only one claim.

We have considered models for both individual and average claim sizes, both models gave very similar results in our application to car insurance data. Based on the models for claim frequency and individual and average claim sizes, respectively, we finally approximated the posterior predictive distribution of the total claim sizes using simulation. According to several scoring rules and the PMCC, especially the inclusion of spatial effects in the model for claim frequency leads to improved predictions of the total claim sizes. However, the inclusion of number of claims effects in the claim size models hardly affects the total claim sizes according to the scoring rules. This can be explained by the fact, that very rarely more than one claim is simulated and therefore number of claims effects have almost no impact.

Further, the additional information available when the non-aggregated data for individual claim sizes are analysed, does not lead to better predictions of the total claim size. The modelling of average claim sizes turned out to be sufficient.

Appendix A

Proof of result (4.5)

Proof of (4.5): Let Q and V be partitioned as follows

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \quad \text{with} \quad Q_{11} \in \mathbb{R}^{J-1,J-1}, Q_{22} \in \mathbb{R}$$

$$Q^{-1} := V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \quad \text{with} \quad V_{11} \in \mathbb{R}^{J-1,J-1}, V_{22} \in \mathbb{R}$$

For the proof results from the intrinsic conditional autoregressive Gaussian process (see Besag and Kooperberg (1995)) will be used. The precision matrix of the intrinsic CAR model is denoted by

$$Q^0 = \begin{pmatrix} Q^0_{11} & Q^0_{12} \\ Q^0_{21} & Q^0_{22} \end{pmatrix} \text{with} \quad Q^0_{ij} = \begin{cases} N_i & \text{for } i = j \\ -1 & \text{for } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

which is connected to the model used in this paper by

$$Q = \psi \cdot Q^0 + I \tag{A.1}$$

for $\psi > 0$. The following results (see Czado and Prokopenko (2004)) will be used:

$$(Q^0_{11})^{-1} Q^0_{12} = -\mathbf{1}_{J-1} \tag{A.2}$$

$$Q^0_{22} = Q^0_{21}[(Q^0_{11})^{-1} Q^0_{12}] \tag{A.3}$$

Using Rao (1973), p.33, yields to

$$\begin{aligned} V_{22} &= (Q_{22} - Q_{21}(Q_{11})^{-1}Q_{12})^{-1} \\ &\overset{(A.1)}{=} (\psi Q_{22}^0 + 1 - \psi Q_{21}^0(\psi Q_{11}^0 + I_{J-1})^{-1}\psi Q_{12}^0)^{-1} \\ &= [\psi(Q_{22}^0 + \frac{1}{\psi} - Q_{21}^0(Q_{11}^0 + \frac{1}{\psi}I_{J-1})^{-1}Q_{12}^0)]^{-1} \end{aligned}$$

and

$$\begin{aligned} \lim_{\psi\to\infty} V_{22} &= \lim_{\psi\to\infty} \frac{\frac{1}{\psi}}{Q_{22}^0 + \frac{1}{\psi} - Q_{21}^0(Q_{11}^0 + \frac{1}{\psi}I_{J-1})^{-1}Q_{12}^0} \\ &\overset{\frac{1}{\psi}=x}{=} \lim_{x\to 0} \frac{x}{Q_{22}^0 + x - Q_{21}^0(Q_{11}^0 + xI_{J-1})^{-1}Q_{12}^0} \\ &\overset{\text{l'Hospital}}{=} \lim_{x\to 0} \frac{1}{1 - Q_{21}^0(Q_{11}^0 + xI_{J-1})^{-2}Q_{12}^0} = \frac{1}{1 + Q_{21}^0(Q_{11}^0)^{-2}Q_{12}^0} \\ &= \frac{1}{1 + [Q_{21}^0(Q_{11}^0)^{-1}][(Q_{11}^0)^{-1}Q_{12}^0]} \overset{(A.2)}{=} \frac{1}{1 + (J-1)} = \frac{1}{J} \end{aligned}$$

The theorem of l'Hospital can be applied here since the denominator in the second equality takes the value 0 for $x = 0$ using (A.3). By Anderson (1958) and Rao (1973), p.33,

$$\begin{aligned} E(\boldsymbol{\gamma}_{-\mathbf{J}}|\gamma_J) &= \mu_J = V_{12}(V_{22})^{-1}\gamma_J = -(Q_{11})^{-1}Q_{12}\gamma_J \\ &= -(Q_{11}^0\psi + I_{J-1})^{-1}(\psi Q_{12}^0)\gamma_J = -[\psi(Q_{11}^0 + \frac{1}{\psi}I_{J-1})]^{-1}\psi Q_{12}^0\gamma_J \\ &= -\frac{1}{\psi}(Q_{11}^0 + \frac{1}{\psi}I_{J-1})^{-1}\psi Q_{12}^0\gamma_J = -(Q_{11}^0 + \frac{1}{\psi}I_{J-1})^{-1}Q_{12}^0\gamma_J \end{aligned}$$

and

$$\lim_{\psi\to\infty} \mu_J = -(Q_{11}^0)^{-1}Q_{12}^0\gamma_J \overset{(A.2)}{=} \gamma_J \mathbf{1}_{J-1}$$

Further, by Anderson (1958) the covariance matrix of $\boldsymbol{\gamma}_{-\mathbf{J}}|\gamma_J$ is given by

$$Cov(\boldsymbol{\gamma}_{-\mathbf{J}}|\gamma_J) = \sigma^2(V_{11} - V_{12}V_{22}^{-1}V_{21}) = \sigma^2 Q_{11}^{-1}.$$

Here again Rao (1973), p33, was used. Taking the limit yields to

$$\lim_{\psi\to\infty} Q_{11}^{-1} = \lim_{\psi\to\infty} \frac{1}{\psi}(Q_{11}^0 + \frac{1}{\psi}I_{J-1})^{-1} = 0.$$

Appendix B

Details on the MCMC algorithms for the count data models

For the count data models presented in Chapter 5 Metropolis within-Gibbs samplers are implemented in Matlab. The regression parameters $\boldsymbol{\beta}$, the spatial random effects $\boldsymbol{\gamma}$, the spatial hyperparameters ψ, σ^2 and the model specific dispersion parameters are updated component by component. When the full conditional distribution can be sampled from directly, a Gibbs step is implemented otherwise a MH step is used. Therefore in the following the full conditionals for all model parameters in the considered models are derived.

B.1 GP regression model

The joint posterior distribution of the regression parameters $\boldsymbol{\beta}$, the overdispersion parameter λ and the spatial effects $\boldsymbol{\gamma}$ in the GP regression model is proportional to

$$\begin{aligned}
p(\boldsymbol{\beta}, \lambda, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{x}, \psi, \sigma^2) &\propto \Big[\prod_{i=1}^{n} p(y_i|\boldsymbol{x_i}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda)\Big] \cdot \pi(\boldsymbol{\gamma}|\psi, \sigma^2) \cdot \pi(\lambda) \cdot \pi(\boldsymbol{\beta}) \\
&\propto \Big[\prod_{i=1}^{n} \exp\{\log \mu_i + (y_i - 1)\log(\mu_i + y_i \frac{\lambda}{1-\lambda}) + y_i \log(1-\lambda) \\
&- \mu_i(1-\lambda) - \lambda y_i\}\Big] \cdot \pi(\boldsymbol{\gamma}|\psi, \sigma^2) \cdot \pi(\lambda) \cdot \pi(\boldsymbol{\beta}) \\
&= \Big[\prod_{i=1}^{n} \exp\{\boldsymbol{x_i}'\boldsymbol{\beta} + \gamma_{R(i)} + (y_i - 1)\log\Big(\exp(\boldsymbol{x_i}'\boldsymbol{\beta} + \gamma_{R(i)}) + y_i \frac{\lambda}{1-\lambda}\Big) \\
&+ y_i \log(1-\lambda) - \exp(\boldsymbol{x_i}'\boldsymbol{\beta} + \gamma_{R(i)})(1-\lambda) - \lambda y_i\}\Big] \\
&\cdot \pi(\boldsymbol{\gamma}|\psi, \sigma^2) \cdot \pi(\lambda) \cdot \pi(\boldsymbol{\beta}).
\end{aligned}$$

The prior distributions are chosen as in Section 5.2. From this joint posterior distribution we derive the full conditional distributions for the regression parameters $\boldsymbol{\beta}$

$$\begin{aligned} p(\boldsymbol{\beta}|\lambda,\boldsymbol{\gamma},\mathbf{y},\mathbf{x}) &\propto \prod_{i=1}^{n}\exp\{\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)}+(y_i-1)\log\Big(\exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})+y_i\frac{\lambda}{1-\lambda}\Big) \\ &- \exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})(1-\lambda)-\frac{1}{2}\boldsymbol{\beta}'\tau^{-2}I_{k+1}\boldsymbol{\beta}\}, \end{aligned}$$

for the overdispersion parameter

$$\begin{aligned} p(\lambda|\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{y},\mathbf{x}) &\propto \prod_{i=1}^{n}\exp\{(y_i-1)\log\Big(\exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})+y_i\frac{\lambda}{1-\lambda}\Big)-y_i\log\frac{1}{1-\lambda} \\ &- \exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})(1-\lambda)-\lambda y_i\} \end{aligned}$$

and for the spatial effects

$$\begin{aligned} p(\boldsymbol{\gamma}|\lambda,\boldsymbol{\beta},\mathbf{y},\mathbf{x}) &\propto \prod_{i=1}^{n}\exp\{\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)}+(y_i-1)\log\Big(\exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})+y_i\frac{\lambda}{1-\lambda}\Big) \\ &- \exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})(1-\lambda)-\frac{1}{2\sigma^2}\boldsymbol{\gamma}'Q\boldsymbol{\gamma}\}. \end{aligned}$$

In particular we have for the l-th component of $\boldsymbol{\gamma}$ given the remaining spatial components $\boldsymbol{\gamma}_{-l}$, the data and all other parameters

$$\begin{aligned} p(\gamma_l|\lambda,\boldsymbol{\beta},\mathbf{y},\mathbf{x},\boldsymbol{\gamma}_{-l}) &\propto \Big[\prod_{i\,with\,R(i)=l}\exp\{\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)}+(y_i-1)\log\Big(\exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})+y_i\frac{\lambda}{1-\lambda}\Big) \\ &- \exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})(1-\lambda)\}\Big]\exp\{-\frac{1}{2\sigma^2}(\gamma_l^2Q_{l,l}+2\gamma_l\sum_{j\neq l}\gamma_jQ_{j,l})\} \\ &= \Big[\prod_{i\,with\,R(i)=l}\exp\{\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)}+(y_i-1)\log\Big(\exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})+y_i\frac{\lambda}{1-\lambda}\Big) \\ &- \exp(\boldsymbol{x_i}'\boldsymbol{\beta}+\gamma_{R(i)})(1-\lambda)\}\Big]\exp\{-\frac{1}{2\sigma^2}[\gamma_l(\gamma_lQ_{l,l}+2\sum_{j\sim l}\gamma_jQ_{j,l})]\}. \end{aligned}$$

The full conditional distributions for the spatial hyperparameters are given by

$$\begin{aligned} p(\psi|\sigma^2,\boldsymbol{\gamma},\boldsymbol{\beta},\lambda,\mathbf{y}) &\propto p(\psi|\sigma^2,\boldsymbol{\gamma}) \\ &\propto |Q|^{\frac{1}{2}}\exp\Big[-\frac{1}{2\sigma^2}\boldsymbol{\gamma}'Q\boldsymbol{\gamma}\Big]\cdot\pi(\psi) \\ &= \exp\Big[\frac{1}{2}\Big(\log|Q|-\frac{1}{\sigma^2}\boldsymbol{\gamma}'Q\boldsymbol{\gamma}\Big)\Big]\cdot\pi(\psi) \end{aligned}$$

and

$$\begin{aligned} p\Big(\frac{1}{\sigma^2}|\boldsymbol{\gamma},\psi\Big) &\propto \Big(\frac{1}{\sigma^2}\Big)^{\frac{J}{2}}\cdot\exp\Big[-\frac{1}{2}\boldsymbol{\gamma}'Q\boldsymbol{\gamma}\cdot\frac{1}{\sigma^2}\Big]\cdot\Big(\frac{1}{\sigma^2}\Big)^{a-1}\cdot\exp[-b\frac{1}{\sigma^2}] \\ &= \Big(\frac{1}{\sigma^2}\Big)^{\frac{J}{2}+a-1}\cdot\exp\Big[-(\frac{1}{2}\boldsymbol{\gamma}'Q\boldsymbol{\gamma}+b)\cdot\frac{1}{\sigma^2}\Big]\sim Gamma\Big(\frac{J}{2}+a,\frac{1}{2}\boldsymbol{\gamma}'Q\boldsymbol{\gamma}+b\Big). \end{aligned}$$

Since the full conditional of σ^2 is again Inverse Gamma, σ^2 can be sampled directly. For the remaining parameters a MH Algorithm with either a symmetric random walk or an independence proposal described as in Section 2.2.2 is used. In independence samplers a t-proposal with the same mode and inverse curvature at the mode as the target distribution is used, see Section 2.2.2 for details. The mode is determined using the bisection algorithm, hence the first and the second derivatives of the full conditionals have to be calculated. For the Poisson regression model the same algorithm with λ fixed to 0 can be used. The update of the spatial hyperparameters ψ and σ^2 is the same for all models and is therefore not mentioned any more in the following.

B.2 NB regression model

The joint posterior distribution of $\boldsymbol{\beta}, r$ and $\boldsymbol{\gamma}$ in the NB regression model is given by

$$p(\boldsymbol{\beta}, r, \boldsymbol{\gamma}|\mathbf{y}, \mathbf{x}) \propto \Big[\prod_{i=1}^{n} p(y_i|r, \boldsymbol{\beta}, \boldsymbol{\gamma})\Big] \cdot \pi(\boldsymbol{\beta}) \cdot \pi(\boldsymbol{\gamma}|\psi, \sigma^2) \cdot \pi(r|a, b) \cdot \pi(b|c, d).$$

We obtain the following full conditionals:

- $p(\boldsymbol{\beta}|r, \boldsymbol{\gamma}, \mathbf{y}, \mathbf{x}) \propto \Big[\prod_{i=1}^{n} \Big(\frac{r}{\mu_i+r}\Big)^r \Big(\frac{\mu_i}{\mu_i+r}\Big)^{y_i}\Big] \cdot \pi(\boldsymbol{\beta})$
- $p(\boldsymbol{\gamma}|r, \boldsymbol{\beta}, \mathbf{y}, \mathbf{x}) \propto \Big[\prod_{i=1}^{n} \Big(\frac{r}{\mu_i+r}\Big)^r \Big(\frac{\mu_i}{\mu_i+r}\Big)^{y_i}\Big] \cdot \pi(\boldsymbol{\gamma}|\psi, \sigma^2)$
- $p(r|\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{x}, a, b) \propto \Big[\prod_{i=1}^{n} \frac{\Gamma(y_i+r)}{\Gamma(r)} \Big(\frac{r}{\mu_i+r}\Big)^r \Big(\frac{\mu_i}{\mu_i+r}\Big)^{y_i}\Big] r^{a-1} e^{-rb}$
- $p(b|r, c, d) \propto b^{a+c-1} \cdot e^{-b(r+d)} \sim Gamma(a + c, r + d)$

The prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the same as in the GP model. Only b can be sampled directly using a Gibbs step, the remaining parameters are updated using a single component MH algorithm. For the regression parameters $\boldsymbol{\beta}$ and the spatial effects $\boldsymbol{\gamma}$ independence proposal distributions, as described in the previous section, are chosen, whereas for the parameter r a symmetric random walk proposal is assumed.

B.3 ZI models

In this section details about the MCMC algorithms for the ZI models are given.

B.3.1 Gamerman Proposal for regression parameters and spatial effects in the ZIP model

The ZIP distribution is not a distribution of the exponential family, however for all observations with $z_i = 0$, i.e. for all observations arising from the Poisson distribution with mean μ_i, we are still in a GLM setting. Therefore Gamermans proposal distribution , see Section 2.2.3 can be used. We use the following notation:

$$Y_i \sim ZIP(p_i, \mu_i) \text{ with } \mu_i = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_{R(i)}) := t_i \exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})$$

where $\tilde{\mathbf{x}}_i' = (\mathbf{x}_i', \mathbf{v}_i')$ with $\mathbf{v}_i = (v_{i1}, .., v_{iJ}), v_{ij} = \begin{cases} 1, & \text{if } R(i) = j \\ 0 & \text{otherwise} \end{cases}$ and $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma})'$. According to Section 5.2 we assume a Normal prior for $\boldsymbol{\alpha}$ in the general notation $\pi(\boldsymbol{\alpha}) \sim N(\boldsymbol{\mu_0}, V_0)$. For all observations y_i with $z_i = 0$ the likelihood can be expressed as

$$f(y_i|\mu_i, p_i) = \exp[y_i \log(\mu_i) - \mu_i - \log y_i! + \log(1 - p_i)],$$

i.e. we have $b(\theta_i) = \mu_i = \exp(\theta_i)$ and the canonical link is given by $g(\mu_i) = \log(\mu_i) = \eta_i = \tilde{\mathbf{x}}_i'\boldsymbol{\alpha}$. According to Section 2.2.3 the weights are then given by $w_{ii}^{-1} = b''(\theta_i)(g'(\mu_i))^2 = \mu_i \frac{1}{\mu_i^2} = \mu_i^{-1}$. For the ZIP model the matrix of weights is therefore given by

$$W(\boldsymbol{\alpha}) = diag\Big([(1 - z_i)\mu_i]^{-1}\Big) = diag\Big([(1 - z_i)\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})]^{-1}\Big).$$

With this choice of $W(\boldsymbol{\alpha})$ the Gamerman algorithm can then be applied for the update of $\boldsymbol{\alpha}$ given $\mathbf{y}, \mathbf{z}$ and $\mathbf{p}$ as described in Section 2.2.3. Note, that for computational efficiency we do not update the whole vector $\boldsymbol{\alpha}$ in one block but divide it into several components. In our model, $\boldsymbol{\beta}$ is updated in one block, whereas the spatial effects $\gamma_j, j = 1, .., J$ are updated component by component. As described in Section 2.2.3 the fixed components are then treated as offsets.

B.3.2 Independence Sampler for ZIGP and ZIP models

Since the GP distribution does not belong to the exponential family, the implementation of Gamerman's proposal is not straightforward anymore. Therefore we use an independence sampler with a t- proposal distribution with the same mode and inverse curvature at the mode for a separate update of $\beta_j, j = 1, .., k$ and $\gamma_j, j = 1, .., J$ as described in Section 2.2.2. Again the bisection method is used to compute the mode of the target distributions. Then, it remains to evaluate the inverse curvature of the target at the mode.

Therefore in the following, the full conditionals as well as their first and second derivative of all model parameters are reported. Of course the same algorithms can be used for the ZIP model with λ fixed to 0.

Independence proposals for regression parameters

Recall that $\pi(\beta_j) \sim N(0, \tau^2)$. The first and second derivatives of the full conditional of β_j

$$\begin{aligned} l(\beta_j) &:= \log p(\beta_j|\lambda, p, \boldsymbol{\beta_{-j}}, \boldsymbol{\gamma}, \mathbf{y}, \mathbf{z}) \propto \sum_{i=1}^{n}(1-z_i)\Big[\log \mu_i + (y_i - 1)\log(\mu_i + \frac{\lambda}{1-\lambda}y_i) \\ &- \mu_i(1-\lambda)\Big] + \log \pi(\beta_j) \\ &= \sum_{i=1}^{n}(1-z_i)\Big[\log \mu_i + (y_i - 1)\log(\mu_i + \frac{\lambda}{1-\lambda}y_i) - \mu_i(1-\lambda)\Big] - \frac{1}{2\tau^2}\beta_j^2 \\ &= \sum_{i:z_i=0}\Big[\log \mu_i + (y_i - 1)\log(\mu_i + \frac{\lambda}{1-\lambda}y_i) - \mu_i(1-\lambda)\Big] - \frac{1}{2\tau^2}\beta_j^2 \end{aligned}$$

are given by

$$\frac{\partial l(\beta_j)}{\partial \beta_j} = \sum_{i:z_i=0}\Big[x_{ij}(1 + (y_i - 1)\frac{\mu_i}{\mu_i + \frac{\lambda}{1-\lambda}y_i} - \mu_i(1-\lambda))\Big] - \frac{\beta_j^2}{\tau^2}$$

$$\frac{\partial^2 l(\beta_j)}{\partial \beta_j^2} = \sum_{i:z_i=0}\Big\{x_{ij}x_{ik}\mu_i\Big[(y_i - 1)\frac{\frac{\lambda}{1-\lambda}y_i}{(\mu_i + \frac{\lambda}{1-\lambda}y_i)^2} - (1-\lambda)\Big]\Big\} - \frac{1}{\tau^2}$$

Independence proposals for spatial effects

For the spatial effects γ_j, $j = 1, ..J$ we obtain the following full conditional and first and second derivatives:

$$\begin{aligned} l(\gamma_j) &:= \log p(\gamma_j|\lambda, p, \boldsymbol{\beta}, \boldsymbol{\gamma_{-j}}, \mathbf{y}, \mathbf{z}) \propto \sum_{i:R(i)=j}(1-z_i)\Big[\log \mu_i + (y_i - 1)\log(\mu_i + \frac{\lambda}{1-\lambda}y_i) \\ &- \mu_i(1-\lambda)\Big] - \frac{1+|\psi|N_j}{2\sigma^2}\Big(\gamma_j - \frac{\psi}{1+|\psi|N_j}\sum_{i\sim j}\gamma_i\Big)^2 \\ &= \sum_{i:R(i)=j\&z_i=0}\Big[\log \mu_i + (y_i - 1)\log(\mu_i + \frac{\lambda}{1-\lambda}y_i) - \mu_i(1-\lambda)\Big] \\ &- \frac{1+|\psi|N_j}{2\sigma^2}\Big(\gamma_j - \frac{\psi}{1+|\psi|N_j}\sum_{i\sim j}\gamma_i\Big)^2 \end{aligned}$$

$$\frac{\partial l(\gamma_j)}{\partial \gamma_j} = \sum_{i:R(i)=j\&z(i)=0} \Big[1 + (y_i - 1)\frac{\mu_i}{\mu_i + \frac{\lambda}{1-\lambda} y_i} - \mu_i(1-\lambda)\Big] - \frac{1+|\psi|N_j}{\sigma^2}\Big(\gamma_j - \frac{\psi}{1+|\psi|N_j}\sum_{i\sim j}\gamma_i\Big)$$

$$\frac{\partial^2 l(\gamma_j)}{\partial \gamma_j \partial \gamma_k} = \sum_{i:R(i)=j\&z_i=0} \mu_i \Big[(y_i - 1)\frac{\frac{\lambda}{1-\lambda} y_i}{(\mu_i + \frac{\lambda}{1-\lambda} y_i)^2} - (1-\lambda)\Big] - \frac{1+|\psi|N_j}{\sigma^2}$$

B.3.3 Update of p

Without regression for p

We consider the case of a constant p first. Based on the likelihood given in (5.9) including the latent variables $\mathbf{z}$, p can be updated directly using a Gibbs step. Assuming a uniform prior for p, i.e. $\pi(p) \sim U([0,1])$, the full conditional for p is given by

$$\begin{aligned} p(p|\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, \mathbf{y}, \mathbf{x}, \mathbf{z}) &\propto \Big[\prod_{i:y_i=0} p^{z_i}(1-p)^{1-z_i} \cdot \prod_{i:y_i>0}(1-p)\Big] \\ &= p^{\sharp i:y_i=0\,\&\,z_i=1}(1-p)^{\sharp i:y_i>0 + \sharp i:y_i=0\,\&\,z_i=0} \\ &:= p^a(1-p)^b, \end{aligned}$$

i.e. $p \sim Beta(a+1, b+1)$.

Regression for p

Now we assume that p may depend on some covariates $\tilde{\mathbf{x}}_i$ and unknown regression parameters $\boldsymbol{\alpha} = (\alpha_1, .., \alpha_m)$ via a logit-link, i.e. $p_i = \dfrac{\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}{1+\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}, \quad i = 1, .., n$. For the parameters $\alpha_j, j = 1, .., m$ the Normal prior $\pi(\alpha_j) \sim N(0, \tau_\alpha^2)$ is assumed.

Gamerman's Algorithm

If $\boldsymbol{\alpha}$ is updated given the latent variables $\mathbf{z}$, Gamerman's proposal can be used. The full conditional for $\boldsymbol{\alpha}$ is given by

$$\begin{aligned} p(\boldsymbol{\alpha}|\mathbf{z}) &\propto \Big[\prod_{i:y_i=0} p_i^{z_i}(1-p_i)^{1-z_i} \cdot \prod_{i:y_i>0}(1-p_i)\Big] \cdot \pi(\boldsymbol{\alpha}) \\ &= \Big[\prod_{i:y_i=0} \Big(\frac{\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}{1+\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}\Big)^{z_i}\Big(\frac{1}{1+\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}\Big)^{1-z_i}\Big]\Big[\prod_{i:y_i>0}\Big(\frac{1}{1+\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}\Big)\Big] \cdot \pi(\boldsymbol{\alpha}) \\ &= \Big[\prod_{i=1}^{n} \frac{\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha} z_i)}{1+\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}\Big] \cdot \pi(\boldsymbol{\alpha}). \end{aligned}$$

Since marginally $z_i \sim Bernoulli(p_i)$ the mean of z_i is given by

$$E(z_i) = \mu_i = p_i = \frac{\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}{1 + \exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}.$$

Further

$$\eta_i = \tilde{\mathbf{x}}_i'\boldsymbol{\alpha} = \log\frac{p_i}{1-p_i} = \log\frac{\mu_i}{1-\mu_i} = g(\mu_i)$$

and the derivative of $g(\mu_i)$ is given by

$$g'(\mu_i) = \frac{1}{\mu_i(1-\mu_i)}.$$

Define

$$\tilde{z}_i(\boldsymbol{\alpha}) = \eta_i + (z_i - \mu_i)g'(\mu_i) = \tilde{\mathbf{x}}_i'\boldsymbol{\alpha} + (z_i - p_i)\frac{1}{p_i(1-p_i)}$$

and $W(\boldsymbol{\alpha}) = diag(W_i(\boldsymbol{\alpha}))$ where $W_i(\boldsymbol{\alpha}) = \frac{1}{g'(\mu_i)} = \mu_i(1-\mu_i)$. Based on results from the IWLS algorithm, see Section 2.2.3, in Gamerman's Algorithm a multivariate normal proposal with mean

$$f(\boldsymbol{\alpha}) = (V_0^{-1} + X'W(\boldsymbol{\alpha})X)^{-1}(V_0^{-1}\mu_0 + X'W(\boldsymbol{\alpha})\tilde{z}(\boldsymbol{\alpha}))$$

and variance

$$cov = (V_0^{-1} + X'W(\boldsymbol{\alpha})X)^{-1})$$

is chosen for $\boldsymbol{\alpha}$, i.e. $q(\tilde{\boldsymbol{\alpha}}|\boldsymbol{\alpha}) \sim N(f(\boldsymbol{\alpha}), cov(\boldsymbol{\alpha}))$. The resulting acceptance probability is then given by

$$\min\Big\{1, \frac{p(\tilde{\boldsymbol{\alpha}}|\mathbf{z})}{p(\boldsymbol{\alpha}|\mathbf{z})}\frac{q(\boldsymbol{\alpha}|\tilde{\boldsymbol{\alpha}})}{q(\tilde{\boldsymbol{\alpha}}|\boldsymbol{\alpha})}\Big\}.$$

B.3.4 Collapsed Algorithms

In simulated data, due to identifiability problems between the intercept β_0 and p no stationarity of the chains was reached for β_0 and p, mixing was rather bad. Since the full conditionals of these two parameters are only linked via the latent variables $\mathbf{z}$ we used collapsed algorithms where we update β_0 and p with $\mathbf{z}$ integrated out, that is based on the likelihood given in (5.6). Therefor convergence of the chains was improved a lot. Again an independence sampler described as in Section 2.2.2 is used. The necessary first and second derivatives of the full conditionals are given in the following sections.

Update of the intercept with z integrated out

The likelihood in the ZIGP-Model with the latent variables integrated out is given by

$$\begin{aligned}\log l(y) \quad \propto \quad & \sum_{i:y_i=0} \log\Big[p + (1-p)\exp(-\mu_i(1-\lambda))\Big] \\ + \quad & \sum_{i:y_i>0} \Big[\log(1-p) + \log\mu_i + (y_i-1)\log\Big(\mu_i + \frac{\lambda}{1-\lambda}y_i\Big) + y_i\log(1-\lambda) \\ - \quad & \mu_i(1-\lambda) - \lambda y_i\Big].\end{aligned}$$

Therefore we have

$$\begin{aligned}l(\beta_0) := \log p(\beta_0|\mathbf{y}, p, \lambda, \boldsymbol{\beta}_{-0}, \boldsymbol{\gamma}) \quad \propto \quad & \sum_{i:y_i=0} \log\Big[p + (1-p)\exp(-\mu_i(1-\lambda))\Big] \\ + \quad & \sum_{i:y_i>0} \Big[\log\mu_i + (y_i-1)\log(\mu_i + \frac{\lambda}{1-\lambda}y_i) - \mu_i(1-\lambda)\Big] \\ - \quad & \frac{\beta_0^2}{2\tau^2}.\end{aligned}$$

The first and second derivative are then given by

$$\begin{aligned}l'(\beta_0) \quad = \quad & \sum_{i:y_i=0} \frac{-(1-p)\exp(-\mu_i(1-\lambda))\mu_i(1-\lambda)}{p + (1-p)\exp(-\mu_i(1-\lambda))} + \sum_{i:y_i>0}\Big[1 + (y_i-1)\frac{\mu_i}{\mu_i + \frac{\lambda}{1-\lambda}y_i} - \mu_i(1-\lambda)\Big] \\ - \quad & \frac{\beta_0}{\tau^2}\end{aligned}$$

and

$$\begin{aligned}l''(\beta_0) \quad = \quad & \sum_{i:y_i=0} \frac{\mu_i(1-\lambda)(1-p)\exp(-\mu_i(1-\lambda))\Big[p(\mu_i(1-\lambda)-1) - (1-p)\exp(-\mu_i(1-\lambda))\Big]}{[p + (1-p)\exp(-\mu_i(1-\lambda))]^2} \\ + \quad & \sum_{i:y_i>0}\Big[(y_i-1)\frac{\frac{\lambda}{1-\lambda}y_i\mu_i}{(\mu_i + \frac{\lambda}{1-\lambda}y_i)^2} - \mu_i(1-\lambda)\Big] - \frac{1}{\tau^2}.\end{aligned}$$

Update of constant p with z integrated out

The logarithm of the full conditional for p with the latent variables $\mathbf{z}$ integrated out is given by

$$l(p) := \log p(p|\mathbf{y}, \lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}) \quad \propto \quad \sum_{i:y_i=0} \log\Big[p + (1-p)\exp(-\mu_i(1-\lambda))\Big] + \sum_{i:y_i>0} \log(1-p)$$

The first and the second derivatives are therefore given by

$$l'(p) = \sum_{i:y_i=0} \frac{1 - \exp(-\mu_i(1-\lambda))}{p + (1-p)\exp(-\mu_i(1-\lambda))} - \sum_{i:y_i>0} \frac{1}{1-p}$$

$$l''(p) = \sum_{i:y_i=0} \frac{-[1-\exp(-\mu_i(1-\lambda))]^2}{[p+(1-p)\exp(-\mu_i(1-\lambda))]^2} - \sum_{i:y_i>0} \frac{1}{(1-p)^2}.$$

Note that $l''(p) < 0$, i.e. the full conditional of p is log concave. Therefore, adaptive rejection sampling (ARS) can be applied as well.

Regression for p with z integrated out

Due to the correlation between $\boldsymbol{\alpha}$ and the intercept β_0 induced by $\mathbf{z}$ no stationarity was reached. Therefore we update $\boldsymbol{\alpha}$ with $\mathbf{z}$ integrated out as well using and independence sampler. The full conditional of α_j is then given by

$$\begin{aligned} l(\alpha_j) := \log p(\alpha_j|\mathbf{y},\lambda,\boldsymbol{\alpha_{-j}},\boldsymbol{\beta},\boldsymbol{\gamma}) \quad &\propto \quad \sum_{i:y_i=0} \log\Big[\frac{1}{1+\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}(\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha}) + \exp(-\mu_i(1-\lambda)))\Big] \\ &- \quad \sum_{i:y_i>0} \log(1+\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})) - \frac{\alpha_j^2}{2\tau_\alpha^2} \end{aligned}$$

and the first and second derivatives by

$$\frac{\partial l(\alpha_j)}{\partial \alpha_j} = \sum_{i:y_i=0} \frac{\tilde{x}_{ij}\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}{\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha}) + \exp(-\mu_i(1-\lambda))} - \sum_{i=1}^{n} \frac{\tilde{x}_{ij}\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}{\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})+1} - \frac{\alpha_j}{\tau_\alpha^2}$$

and

$$\frac{\partial^2 l(\alpha_j)}{\partial \alpha_j^2} = \sum_{i:y_i=0} \frac{\tilde{x}_{ij}^2\exp(-\mu_i(1-\lambda))\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}{[\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha}) + \exp(-\mu_i(1-\lambda))]^2} - \sum_{i=1}^{n} \frac{\tilde{x}_{ij}^2\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})}{[\exp(\tilde{\mathbf{x}}_i'\boldsymbol{\alpha})+1]^2} - \frac{1}{\tau_\alpha^2}.$$

λ-update without z's

In the ZIGP Model convergence was further complicated by additional correlation between λ and β_0 and p. To overcome these problems also λ was updated with the latent variables $\mathbf{z}$ integrated out. The full conditional of λ is then given by

$$\begin{aligned} l(\lambda) := \log p(\lambda|\mathbf{y},p,\boldsymbol{\beta},\boldsymbol{\gamma}) \quad &\propto \quad \sum_{i:y_i=0} \log\Big[p+(1-p)\exp(-\mu_i(1-\lambda))\Big] \\ &+ \quad \sum_{i:y_i>0} \Big[(y_i-1)\log\Big(\mu_i + \frac{\lambda}{1-\lambda}y_i\Big) + y_i\log(1-\lambda) - \lambda(y_i-\mu_i)\Big] \end{aligned}$$

and we obtain the following first and second derivatives:

$$\begin{aligned} l'(\lambda) \quad &= \quad \sum_{i:y_i=0} \frac{(1-p)\exp(-\mu_i(1-\lambda))\mu_i}{p+(1-p)\exp(-\mu_i(1-\lambda))} + \sum_{i:y_i>0} \Big[(y_i-1)y_i\frac{1}{(1-\lambda)^2(\mu_i+\frac{\lambda}{1-\lambda}y_i)} \\ &- \quad y_i\frac{1}{1-\lambda} - y_i + \mu_i\Big] \end{aligned}$$

$$
\begin{aligned}
l''(\lambda) &= \sum_{i:y_i=0} \frac{p(1-p)\mu_i^2 \exp(-\mu_i(1-\lambda))}{[p+(1-p)\exp(-\mu_i(1-\lambda))]^2} \\
&+ \sum_{i:y_i>0} \Big[\frac{y_i(y_i-1)}{(1-\lambda)^3(\mu_i+\frac{\lambda}{1-\lambda}y_i)} \cdot \Big(2 - \frac{y_i}{(1-\lambda)(\mu_i+\frac{\lambda}{1-\lambda}y_i)}\Big) - y_i \frac{1}{(1-\lambda)^2} \Big]
\end{aligned}
$$

Appendix C

Collapsed algorithm in Section 6.2.4

For the collapsed algorithm in Section 6.2.4 we consider $p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R}) = \int p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}) d\boldsymbol{\gamma}$. We have

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}) &\propto \exp\Big\{-\frac{1}{2}\Big[\sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2}(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta} + \tilde{\mathbf{v}}_i'\boldsymbol{\gamma})^2 + \boldsymbol{\gamma}'\sigma^{-2}Q\boldsymbol{\gamma} + \boldsymbol{\beta}'\tau^{-2}I\boldsymbol{\beta}\Big]\Big\} \\ &= \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\beta}'\tau^{-2}I\boldsymbol{\beta} + \sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2}(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta})^2\Big]\Big\} \\ &\times \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\gamma}'\Big(\sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i' + \sigma^{-2}Q\Big)\boldsymbol{\gamma} + 2\boldsymbol{\gamma}'\sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i(\tilde{y}_i + \tilde{\mathbf{x}}_i'\boldsymbol{\beta})\Big]\Big\} \\ &:= c(\boldsymbol{\beta}) \times \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\gamma}'A\boldsymbol{\gamma} + 2\boldsymbol{\gamma}'a\Big]\Big\} \end{aligned} \tag{C.1}$$

where $A := \sum_{i=1}^{N} \frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i' + \sigma^{-2}Q$. Further

$$\begin{aligned} \exp &\Big\{-\frac{1}{2}\Big[\boldsymbol{\gamma}'A\boldsymbol{\gamma} + 2\boldsymbol{\gamma}'a\Big]\Big\} \\ &\propto \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\gamma}'A\boldsymbol{\gamma} + 2\boldsymbol{\gamma}'A(A^{-1}a) + (A^{-1}a)'A(A^{-1}a) - (A^{-1}a)'A(A^{-1}a)\Big]\Big\} \\ &\propto \exp\Big\{-\frac{1}{2}\Big[(\boldsymbol{\gamma} + A^{-1}a)'A(\boldsymbol{\gamma} + A^{-1}a) - (A^{-1}a)'A(A^{-1}a)\Big]\Big\} \end{aligned}$$

and therefore

$$\begin{aligned} \int \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\gamma}'A\boldsymbol{\gamma} + 2\boldsymbol{\gamma}'a\Big]\Big\} d\boldsymbol{\gamma} &\propto (2\pi)^{\frac{J}{2}}|A|^{-\frac{1}{2}} \exp\Big\{\frac{1}{2}(A^{-1}a)'A(A^{-1}a)\Big\} \\ &\propto \exp\Big\{\frac{1}{2}(A^{-1}a)'A(A^{-1}a)\Big\} \end{aligned} \tag{C.2}$$

From (C.1) and (C.2) it then follows that

$$\int \quad p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{R}) d\boldsymbol{\gamma}$$

$$\begin{aligned} &\propto c(\boldsymbol{\beta})\exp\Big\{\frac{1}{2}(A^{-1}a)'A(A^{-1}a)\Big\} \\ &\propto \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\beta}'(\tau^{-2}I+\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i')\boldsymbol{\beta}+2\boldsymbol{\beta}'\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{x}_i\tilde{y}_i-a'A^{-1}a\Big]\Big\} \end{aligned}$$

Finally, with

$$\begin{aligned} a'A^{-1}a &= \Big(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{y}_i+\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i'\boldsymbol{\beta}\Big)'A^{-1}\Big(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{y}_i+\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i'\boldsymbol{\beta}\Big) \\ &\propto \boldsymbol{\beta}'\Big(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i'\Big)'A^{-1}\Big(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i'\Big)\boldsymbol{\beta}+2\boldsymbol{\beta}'\Big(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i'\Big)'A^{-1}\Big(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{y}_i\Big) \end{aligned}$$

it follows that

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{\tau},\mathbf{R}) \;\propto\; &\exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\beta}'\Big(\tau^{-2}I+\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'-(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')'A^{-1}(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')\Big)\boldsymbol{\beta} \\ &-\; 2\boldsymbol{\beta}'\Big((\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')'A^{-1}(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{y}_i)-\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{y}_i\Big)\Big]\Big\}, \end{aligned}$$

i.e.

$$\boldsymbol{\beta}|\boldsymbol{\tau},\mathbf{R}\sim N(\Sigma_{col}^{-1}\boldsymbol{\mu}_{col},\Sigma_{col}^{-1})$$

with

$$\Sigma_{col}:=\tau^{-2}I+\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'-(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')'A^{-1}(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')$$

and

$$\boldsymbol{\mu}_{col}:=(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{\mathbf{x}}_i')'A^{-1}(\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{v}}_i\tilde{y}_i)-\sum_{i=1}^{N}\frac{1}{\tilde{s}_i^2}\tilde{\mathbf{x}}_i\tilde{y}_i.$$

Bibliography

Agarwal, D. K., A. E. Gelfand, and S. Citron-Pousty (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics 9*, 341–355.

Anderson, T. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Angers, J.-F. and A. Biswas (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Computational Statistics & Data Analysis 42*, 37–46.

Banerjee, S., B. Carlin, and A. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data*. New York: Chapman & Hall/CRC.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems. *Statistical Science 10*, 3–41.

Besag, J. and C. Kooperberg (1995). On conditional and intrinsic autoregressions. *Biometrika 82*, 733–746.

Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math. 43*, 1–59. With discussion.

Boskov, M. and R. Verrall (1994). Premium rating by geographic area using spatial models. *ASTIN Bullletin 24 (1)*, 131–143.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78 (1)*, 1–3.

Christensen, O., G. Roberts, and M. Sköld (2005). Robust MCMC methods for spatial GLMM's. *to appear in Journal of Computational and Graphical Statistics*.

Christensen, O. and R. Waagepetersen (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics 58*, 280–286.

Consul, P. (1989). *Generalized Poisson Distributions. Properties and Applications*. New York: Marcel Dekker, Inc.

Consul, P. and G. Jain (1973). A generalization of the Poisson distribution. *Technometrics 15*, 791–799.

Czado, C. and A. Min (2005). Consistency and asymptotic normality of the maximum likelihood estimator in a zero-inflated generalized Poisson regression. *submitted*.

Czado, C. and S. Prokopenko (2004). Modeling transport mode decisions using hierarchical binary spatial regression models with cluster effects. *Discussion paper 406, SFB 386 Statistische Analyse diskreter Strukturen*. http://www.stat.uni-muenchen.de/sfb386/.

Dellaportas, P. and G. O. Roberts (2003). An introduction to MCMC. In *Spatial statistics and computational methods (Aalborg, 2001)*, Volume 173 of *Lecture Notes in Statist.*, pp. 1–41. New York: Springer.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998). Model-based geostatistics. *J. Roy. Statist. Soc. Ser. C 47*(3), 299–350. With discussion and a reply by the authors.

Dimakos, X. and A. Frigessi (2002). Bayesian premium rating with latent structure. *Scandinavian Actuarial Journal* (3), 162–184.

Famoye, F. and K. Singh (2003a). On inflated generalized Poisson regression models. *Advances and Applications in Statistics 3 (2)*, 145–158.

Famoye, F. and K. Singh (2003b). Zero inflated generalized Poisson regression model. *submitted*.

Frühwirth-Schnatter, S. (2004). Efficient Bayesian parameter estimation. In S. K. A.C. Harvey and N. Shephard (Eds.), *State space and unobserved component models*, pp. 123–151. Cambridge: Cambridge Univ. Press.

Frühwirth-Schnatter, S. and H. Wagner (2004a). Data augmentation and Gibbs sampling for regression models of small counts. *IFAS Research Paper Series 2004-04*.

Frühwirth-Schnatter, S. and H. Wagner (2004b). Gibbs sampling for parameter-driven models of time series of small counts with applications to state space modelling. *IFAS Research Paper Series 2004-01*.

Frühwirth-Schnatter, S. and H. Waldl (2004). Data augmentation and Gibbs sampling for logistic models. *IFAS Research Paper Series 2004-06*.

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

Gelfand, A. and S. Ghosh (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika 85 (1)*, 1–11.

Gelfand, A., S. Sahu, and B.P.Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika 82 (3)*, 479–488.

Gelfand, A. and A. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Gelfand, A. E. and P. Vounatsou (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics 4 (1)*, 11–25.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian Data Analysis, Second Edition*. Boca Raton: Chapman & Hall/CRC.

Gelman, A., X. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica 6*, 733–807.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transaction of Pattern Analysis and Machine Intelligence 6*, 721–741.

Geyer, C. (1992). Practical Markov Chain Monte Carlo. *Statistical Science 7*(4), 473–511.

Gilks, W., S. Richardson, and D. Spiegelhalter (1996a). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, pp. 1–19. Boca Raton: Chapman & Hall/CRC.

Gilks, W., S. Richardson, and D. Spiegelhalter (1996b). *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman & Hall/CRC.

Gilks, W. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics 41 (2)*, 337–348.

Gneiting, T. and A. E. Raftery (2004). Strictly proper scoring rules, prediction and estimation. Technical report no. 463, Department of Statistics, University of Washington.

Gschlößl, S. and C. Czado (2005a). Does a Gibbs sampler approach to spatial Poisson regression models outperform a single site MH sampler? *submitted*.

Gschlößl, S. and C. Czado (2005b). Modelling count data with overdispersion and spatial effects. *Discussion paper 412, SFB 386 Statistische Analyse diskreter Strukturen*. http://www.stat.uni-muenchen.de/sfb386/.

Gschlößl, S. and C. Czado (2005c). Spatial modelling of claim frequency and claim size in insurance. *submitted*.

Haberman, S. and A. Renshaw (1996). Generalized linear models and actuarial science. *The Statistician 45*, 407–436.

Han, C. and B. Carlin (2001). Markov chain Monte Carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association 96*, 1122–1132.

Haran, M., J. Hodges, and B. Carlin (2003). Accelerating computation in Markov random field models for spatial data via structured MCMC. *Journal of Computational and Graphical Statistics 12*(2), 249–264.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Hjort, N., F. Dahl, and G. Steinbakk (2005). Post-Processiong Posterior Predictive P-values. *submitted for publication in JASA*.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14 (4)*, 382–417.

Jin, X., B. Carlin, and S. Banerjee (2004). Generalized hierarchical multivariate CAR models for areal data. *submitted to Biometrics*.

Joe, H. and R. Zhu (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with Negative Binomial distribution. *Biometrical Journal 47*, 219–229.

Jørgensen, B. and M. C. P. de Souza (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal 1*, 69–93.

Kass, R. and A. Raftery (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association 90*, 773–795.

Knorr-Held, L. and H. Rue (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics. Theory and Applications 29*, 325–338.

Lambert, D. (1992). Zero-inflated Poisson regression with and application to defects in manufacturing. *Technometrics 34 (1)*, 1–14.

Laud, P. and J. Ibrahim (1995). Predictive model selection. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 57*(1), 247–262.

Liu, J., W. Wong, and A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika 81 (1)*, 27–40.

Lundberg, F. (1903). Approximerad framställning av sannolikhetsfunktionen. aterförsäkring av kollektivrisker. Akad. afhandling. almqvist och wicksell, uppsala, Almqvist och Wicksell, Uppsala.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. 2nd edn, London: Chapman & Hall.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machine. *J. of Chemical Physics 21*, 1087–1091.

Mikosch, T. (2004). *Non-Life Insurance Mathematics. An Introduction with Stochastic Processes*. New York: Springer.

Panjer, H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin 11*, 22–26.

Papaspiliopoulos, O., G. O. Roberts, and M. Sköld (2003). Non-centered parameterizations for hierarchical models and data augmentation. In *Bayesian statistics, 7 (Tenerife, 2002)*, pp. 307–326. New York: Oxford Univ. Press. With a discussion by Alan E. Gelfand, Ole F. Christensen and Darren J. Wilkinson, and a reply by the authors.

Pettitt, A., I. Weir, and A. Hart (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing 12 (4)*, 353–367.

Rao, C. (1973). *Linear Statistical Inference and Its Applications, 2nd Ed.* New York: Wiley.

Renshaw, A. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin 24*, 265–285.

Roberts, G. and S. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, B 59 (2)*, 291–317.

Rodrigues, J. (2003). Bayesian analysis of zero-inflated distributions. *Communications in Statistics 32 (2)*, 281–289.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 63*(4), 597–614.

Rue, H., I. Steinsland, and S. Erland (2004). Approximating hidden Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 66*(4), 877–892.

Sargent, D., J. Hodges, and B. Carlin (2000). Structured Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics 9*, 217–234.

Smyth, G. K. and B. Jørgensen (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin 32 (1)*, 143–157.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B 64 (4)*, 583–640.

Stern, H. and N. Cressie (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine 19*, 2377–2397.

Sun, D., R. K. Tsutakawa, and P.L.Speckman (1999). Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika 86*, 341–350.

Székely, G. (2003). ϵ-statistics: The energy of statistical samples. Technical report no. 2003-16, Department of Mathematics and Statistics, Bowling Green State University, Ohio.

Taylor, G. (1989). Use of spline functions for premium rating by geographic area. *ASTIN Bullletin 19 (1)*, 89–122.

van der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica 59 (1)*, 45–56.

West, M. (1985). Generalized linear models: scale parameters, outlier accommodation and prior distributions. In *Bayesian statistics, 2 (Valencia, 1983)*, pp. 531–557. Amsterdam: North-Holland. With discussion and a reply by the author.

Winkelmann, R. (2003). *Econometric Analysis of Count Data. 4th Edition*. Berlin: Springer-Verlag.