Retrospective sampling in MCMC with an application to COM-Poisson regression

Charalampos Chanialidis[*], Ludger Evers[*], Tereza Neocleous [*], and Agostino Nobile[†]

[*]School of Mathematics and Statistics, University of Glasgow
[†]Department of Mathematics, University of York

This refers to the pre-print version for the paper accepted in STAT.
The post-print version can be found here.

Abstract

The normalisation constant in the distribution of a discrete random variable may not be available in closed form; in such cases the calculation of the likelihood can be computationally expensive. Approximations of the likelihood or approximate Bayesian computation (ABC) methods can be used; but the resulting MCMC algorithm may not sample from the target of interest. In certain situations one can efficiently compute lower and upper bounds on the likelihood. As a result, the target density and the acceptance probability of the Metropolis-Hastings algorithm can be bounded. We propose an efficient and exact MCMC algorithm based on the idea of retrospective sampling. This procedure can be applied to a number of discrete distributions, one of which is the COM-Poisson distribution. In practice the bounds on the acceptance probability do not need to be particularly tight in order to accept or reject a move. We demonstrate this method using data on the emergency hospital admissions in Scotland in 2010, where the main interest lies in the estimation of the variability of admissions, since it is considered as a proxy for health inequalities.

Keywords: Bayesian methods; likelihood; Markov chain Monte Carlo; regression

1 Introduction

Bayesian and likelihood inference require the repeated evaluation of the likelihood function, the probability (density) function $\pi(y|\theta)$ of the observed data y, regarded as a function of the model parameter θ. In some sampling models the function $\pi(y|\theta)$ is completely specified, in others it is only available up to a constant of proportionality $Z(\theta)$, which must be taken into account when making inference about θ. We propose an MCMC algorithm which only makes use of cheaply computed, arbitrarily precise, upper and lower bounds on $Z(\theta)$. We consider in detail the case of the COM-Poisson distribution.

Approximating the normalisation constant with respect to the observed data can make MCMC methods inaccurate. Instead, we propose an MCMC scheme based on the idea of retrospective sampling: first draw the uniform random variable U which is used to decide on the outcome of the Metropolis-Hastings acceptance/rejection move and then perform any calculations needed on the acceptance ratio. Depending on the value of U, the acceptance ratio (which involves Z both in the numerator and denominator) may not be needed to be known exactly.

The article is organised as follows. Sections 2 and 3 present an overview of the COM-Poisson distribution and the basic idea behind the proposed retrospective sampling scheme. In Section 4 we describe a general approach to the construction of the bounds on $Z(\theta)$ for the case of discrete random variables. In Section 5 we apply the method to data on hospital emergency admissions in Scotland, where identifying geographical areas with high variance is the main interest since variation in admissions is a proxy for health inequalities. Concluding remarks can be found in Section 6.

2 COM-Poisson distribution

The COM-Poisson distribution (Conway & Maxwell, 1962) is a two-parameter generalisation of the Poisson distribution that allows for different levels of dispersion. The probability mass function of the COM-Poisson(λ, ν) distribution is

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)} \quad y = 0, 1, 2, \ldots \tag{1}$$

for $\lambda > 0$ and $\nu \geq 0$, where $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$. The additional parameter ν allows the distribution to model under- dispersed ($\nu > 1$) or over-dispersed

($\nu < 1$) data. The Poisson distribution is a special case ($\nu = 1$).

The normalisation constant $Z(\lambda, \nu)$ does not have a closed form (for $\nu \neq 1$) and has to be approximated, but can be upper bounded. An asymptotic approximation exists, which is reasonably accurate for $\lambda > 10^{\nu}$ (Minka et al., 2003).

Shmueli et al. (2005) describe methods for estimating the parameters of the COM-Poisson and show its flexibility in fitting count data compared to other distributions. They show that

$$\mathbb{E}[Y] \approx \lambda^{\frac{1}{\nu}} + \frac{1}{2\nu} - \frac{1}{2}, \mathbb{V}[Y] \approx \frac{\lambda^{\frac{1}{\nu}}}{\nu}. \tag{2}$$

This parametrisation of the COM-Poisson does not have a clear centering parameter, so we will use the reparametrisation $\mu = \lambda^{\frac{1}{\nu}}$ as proposed by Guikema & Coffelt (2008). The probability mass function is then

$$P(Y = y|\mu, \nu) = \left(\frac{\mu^y}{y!}\right)^{\nu} \frac{1}{Z(\mu, \nu)} \quad y = 0, 1, 2, \ldots \tag{3}$$

with $Z(\mu, \nu) = \sum_{j=0}^{\infty} \left(\frac{\mu^j}{j!}\right)^{\nu}$. The mean and variance can be approximated by

$$\mathbb{E}[Y] \approx \mu + \frac{1}{2\nu} - \frac{1}{2}, \mathbb{V}[Y] \approx \frac{\mu}{\nu}. \tag{4}$$

Thus, in the new parametrisation μ closely approximates the mean, unless both μ and ν are small. The mode of the distribution is $\lfloor \mu \rfloor$, as this formulation is just a tempered Poisson distribution.

The fact that $Z(\mu, \nu)$, and thus the probability mass function $P(Y = y|\mu, \nu)$, is very expensive to compute, has been a key limiting factor for the use of the COM-Poisson distribution. In particular, in a Bayesian approach using the Metropolis-Hastings algorithm, each move requires an evaluation of $Z(\mu, \nu)$ in order to compute the acceptance ratio. In Section 3 we will present an MCMC algorithm that does not need $Z(\mu, \nu)$ to be computed exactly. In Section 4 we will then derive arbitrarily precise lower and upper bounds for $Z(\mu, \nu)$.

3 Retrospective sampling

In simulation-based Bayesian inference we are interested in drawing samples from the posterior distribution of the parameters $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$,

where $\pi(\theta)$ denotes the prior distibution of θ. In the Metropolis-Hastings algorithm a Markov chain is constructed in which, when the current state is θ, a candidate state θ^* is drawn from a proposal distribution $q(\theta^*|\theta)$ and then accepted with probability $\min\{1, p\}$ with

$$p = \frac{\pi(y|\theta^*)\pi(\theta^*)}{\pi(y|\theta)\pi(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}. \tag{5}$$

If θ^* is rejected, the chain remains at the current state θ. In order to accept the candidate θ^* with probability $\min\{1, p\}$, the acceptance ratio p is compared to a random $U \sim \text{Unif}(0, 1)$ and θ^* is accepted if $U < p$. The key idea of the proposed algorithm is that p needs to be known exactly only if U and p are very close. This requires exchanging the order of simulation. This idea, known as retrospective sampling, was first proposed by Papaspiliopoulos & Roberts (2008) as a way of sampling from a Dirichlet process. It has been used, amongst other things, for simulation of diffusion sample paths by Beskos et al. (2006) and Sermaidis et al. (2013).

Suppose we have a sequence of increasingly and arbitrarily precise lower and upper bounds for $\pi(y|\theta)$, denoted by $\check{\pi}(y|\theta)$ and $\hat{\pi}(y|\theta)$, respectively. Plugging these bounds into (5) yields lower and upper bounds for the acceptance probability

$$\check{p}_n = \frac{\check{\pi}_n(y|\theta^*)\pi(\theta^*)}{\hat{\pi}_n(y|\theta)\pi(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \qquad \hat{p}_n = \frac{\hat{\pi}_n(y|\theta^*)\pi(\theta^*)}{\check{\pi}_n(y|\theta)\pi(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \tag{6}$$

By construction $\check{p}_n \leq p \leq \hat{p}_n$ as well as $\check{p}_n \to p$ and $\hat{p}_n \to p$ as $n \to \infty$. The proposed algorithm for deciding on the acceptance of θ^* then proceeds as follows.

1. Draw $U \sim \text{Unif}(0, 1)$ and set the number of refinements $n = 0$.
2. Compute $\check{p}_n$ and $\hat{p}_n$ and compare them to U.
   - If $U \leq \check{p}_n$, accept the candidate value.
   - If $U > \hat{p}_n$, reject the candidate value.
   - If $\check{p}_n < U < \hat{p}_n$, refine the bounds, i.e increase n and return to step 2.

Figure 1 illustrates this idea. Panel 1a shows the Metropolis-Hastings strategy where p is the acceptance ratio and u_1, u_2 are two realizations of the $\text{Unif}(0, 1)$ distribution. In the first case we accept the candidate value θ^* since $u_1 < p$ whereas in the second case we reject it since $u_2 > p$. The

acceptance and rejection regions can also be seen in the figure. Panel 1b shows the retrospective sampling strategy along with the refinement region. When a realization of the Unif$(0,1)$ falls into the refinement region (e.g. u_3), then in order to make a decision we have to refine the bounds. Panel 1c shows the new bounds, where the refined lower bound is above u_3 and as a result we accept the candidate value θ^*.

Because the bounds $\check{p}_n$ and $\hat{p}_n$ are arbitrarily tight, the algorithm will eventually accept or reject a candidate value θ^*.

4 Piecewise geometric bounds

This section explains how the bounds required in the previous algorithm can be constructed for discrete distributions with probability mass function given by

$$\pi(y|\theta) = Z(\theta)^{-1} p_\theta(y), \tag{7}$$

where the normalisation constant $Z(\theta) = \sum_y p_\theta(y)$ is not available in closed form. The COM-Poisson distribution is an example of such a distribution.

For ease of presentation we will assume that only computing the right tail is computationally challenging. At the end of the section we will explain how the method can be applied to bounding the left tail as well.

A simple way of reducing the computational burden is to compute the normalisation constant $Z(\theta)$ up to a kth term and use this as a lower bound for $Z(\theta)$. An upper bound can be obtained by also considering an upper bound for the remaining terms. For the approach to be computationally efficient, k should be chosen to be not too large, which in turn implies that the upper bound for the remaining terms will be rather loose.

On the other hand, if the ratio of consecutive probabilities is bounded by constants over a certain range of y

$$\check{a}_{y_0,y_1} \leq \frac{p_\theta(y+1)}{p_\theta(y)} \leq \hat{a}_{y_0,y_1}, \quad y \in \{y_0, y_0+1, \ldots, y_1-1\}, \tag{8}$$

then tighter bounds can be obtained, at little excess computationally cost.

We will now construct bounds based on the constants $\check{a}_{y_0,y_1}, \hat{a}_{y_0,y_1}$. These tighter bounds are based on including piecewise bounds on a sequence of increasingly large blocks of probabilities in the tails. This corresponds to using the following lower bound and upper bound for $Z(\theta)$:

$$\check{Z}(\theta) = E(\theta) + \check{B}(\theta), \qquad \hat{Z}(\theta) = E(\theta) + \hat{B}(\theta) + \hat{R}(\theta),$$

where $E(\theta) = \sum_{j=0}^{k_1} p_\theta(j)$ is obtained by computing the sum of the first k_1 terms exactly. $\check{B}(\theta)$ and $\hat{B}(\theta)$ are piecewise bounds on blocks of probabilities, computed as set out below. $\hat{R}(\theta)$ is an upper bound on the remaining terms.

If (8) holds, then for all $j \in \{0, \ldots, r\}$ with $r = y_1 - 1 - y_0$

$$(\check{a}_{y_0,y_1})^j \leq \frac{p_\theta(y_0+j)}{p_\theta(y_0)} = \frac{p_\theta(y_0+1)}{p_\theta(y_0)} \cdots \frac{p_\theta(y_0+j)}{p_\theta(y_0+j-1)} \leq (\hat{a}_{y_0,y_1})^j. \tag{9}$$

We can rewrite the sum of the block of $r+1$ probabilities as

$$\sum_{j=0}^{r} p_\theta(y_0+j) = p_\theta(y_0) \sum_{j=0}^{r} \frac{p_\theta(y_0+j)}{p_\theta(y_0)}. \tag{10}$$

Taking advantage of (9) and (10) we obtain the bounds:

$$\begin{aligned}
\check{b}_{y_0,y_1}(\theta) &= p_\theta(y_0) \sum_{j=0}^{r} (\check{a}_{y_0,y_1})^j = p_\theta(y_0) \frac{1-(\check{a}_{y_0,y_1})^{r+1}}{1-\check{a}_{y_0,y_1}} \leq \sum_{j=0}^{r} p_\theta(y_0+j) \\
\hat{b}_{y_0,y_1}(\theta) &= p_\theta(y_0) \sum_{j=0}^{r} (\hat{a}_{y_0,y_1})^j = p_\theta(y_0) \frac{1-(\hat{a}_{y_0,y_1})^{r+1}}{1-\hat{a}_{y_0,y_1}} \geq \sum_{j=0}^{r} p_\theta(y_0+j).
\end{aligned} \tag{11}$$

These bounds are computed in blocks of probabilities. Denote by $s = (k_1, \ldots, k_{l_n})$ the sequence of end-points of the piecewise bounds, we then define

$$\check{B}(\theta) = \sum_{i=1}^{l_n-1} \check{b}_{k_i,k_{i+1}-1}(\theta), \quad \hat{B}(\theta) = \sum_{i=1}^{l_n-1} \hat{b}_{k_i,k_{i+1}-1}(\theta), \tag{12}$$

which satisfies $\check{B}(\theta) \leq \sum_{j=k_1}^{k_{l_n}-1} p_\theta(j) \leq \hat{B}(\theta)$. Finally, using the tail bound

$$\hat{R}(\theta) = \hat{b}_{k_{l_n},\infty}(\theta) \geq \sum_{j=k_{l_n}}^{\infty} p_\theta(j), \tag{13}$$

we obtain the desired result that

$$\check{Z}(\theta) = E(\theta) + \check{B}(\theta) \leq \sum_{j=0}^{\infty} p_\theta(j) \leq E(\theta) + \hat{B}(\theta) + \hat{R}(\theta) = \hat{Z}(\theta) \tag{14}$$

The previous bounds and the number of terms k_1 should have n, the number of refinements, as an extra index since everytime there is a need for refinement we have to choose a larger k_1 and the bounds will be different. For

the rest of the section we will assume that n is fixed. The bounds are increasingly tight as long as $k_1 \to \infty$. In practice the values of k_1 and k_{l_n} are chosen depending on θ and the magnitude the previous contribution to the sum made. In our experience, choosing k_s such that $k_{s+1} - k_s = d^s$ for $d \approx 2$ works well in practice.

So far we have set out how to compute bounds for the right tail of the distribution. If the mode of the distribution is large, then it is advisable to use the same strategy as above for the left tail too. The approach is essentially the same, with the main difference being that the bounds are computed right to left and that the summation will stop at 0.

A graphical explanation of the procedure can be seen in Figure 2 where the first two blocks are comprised of 3 and 5 probabilities respectively.

4.1 Bounds on the COM-Poisson normalisation constant

In the COM-Poisson case, the bounds in (8) and (11) are

$$\check{a}_{y_0,y_1} = \left(\frac{\mu}{y_1}\right)^\nu, \quad \hat{a}_{y_0,y_1} = \left(\frac{\mu}{y_0+1}\right)^\nu, \tag{15}$$

for $y_0, y_1 > \lfloor \mu \rfloor$ and

$$\check{b}_{y_0,y_1}(\theta) = p(y_0)\left(\frac{\mu}{y_1}\right)^\nu \frac{1-\left(\frac{\mu^r}{y_1^r}\right)^\nu}{1-\left(\frac{\mu}{y_1}\right)^\nu}, \; \hat{b}_{y_0,y_1}(\theta) = p(y_0)\left(\frac{\mu}{y_0+1}\right)^\nu \frac{1-\left(\frac{\mu^r}{(y_0+1)^r}\right)^\nu}{1-\left(\frac{\mu}{y_0+1}\right)^\nu}, \tag{16}$$

where $r = y_1 - 1 - y_0$ and $p(y_0) = \left(\frac{\mu^{y_0}}{y_0!}\right)^\nu$ is the unnormalised density of the COM-Poisson. Bounds for the left tail of the distribution are computed in a similar way.

4.2 Weighted Poisson distributions

Del Castillo & Pérez-Casany (1998) developed a family of distributions, known as weighted Poisson distributions, that can handle both under- and overdispersion. A random variable Y is defined to have a weighted Poisson distribution if its probability mass function can be written as

$$P(Y = y) = e^{-\lambda}\frac{\lambda^y w_y}{W y!} \quad y = 0, 1, 2, \ldots \tag{17}$$

where $W = e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j w_j}{j!}$. The weight function is defined as $w_y = (y+a)^r$ with $a \geq 0, r \in \mathbb{R}$. The Poisson distribution is a special case when $r = 0$.

These distributions are used for modelling data with partial recording: when the event $Y = y$ occurs, a Poisson variable is recorded with probability proportional to w_y.

The bounds, found in (8), for this distribution for $r > 0$ are

$$\check{a}_{y_0,y_1} = \frac{\lambda}{y_1}\frac{w_{y_1}}{w_{y_1-1}},\ \hat{a}_{y_0,y_1} = \frac{\lambda}{y_0+1}\frac{w_{y_0+1}}{w_{y_0}}. \tag{18}$$

Similar bounds can be constructed for $r < 0$.

5 Application on emergency hospital admissions in Scotland

As an illustration of the method, we consider data on the hospital emergency admissions for each intermediate geography (1235 in total) in Scotland for the year 2010. Scotland is divided into 6505 small areas, called datazones, each containing around 350 households. An intermediate geography is comprised of neighbouring datazones. The Scottish Index of Multiple Deprivation (SIMD) is the Scottish government's official tool for identifying datazones suffering from deprivation. This index provides a relative ranking for each datazone, from 1 (most deprived) to 6505 (least deprived). The Scottish government's cut-off for a datazone to be considered deprived is to belong in the 15% most deprived datazones in Scotland. Using the SIMD ranks for areas larger than datazones (such as intermediate geographies) one can consider the percentage of datazones within that intermediate geography that are in the 15% most deprived e.g. if an intermediate geography is comprised of 20 datazones and 10 of them are in the 15% most deprived then its local share is 50%. This can also be applied in larger areas such as local authorities.

Tables 1 and 2 refer to the local share of deprived datazones for each local authority in Scotland. Local authorities in the west of Scotland such as Glasgow City, Inverclyde, North Ayrshire, North Lanarkshire, and West Dunbartonshire have a high local share of the most deprived datazones while at the same time their local share of least deprived datazones is small. Parts of the east of Scotland (Edinburgh City, East Lothian) show the opposite trend. It is important to note that the local share percentages of each local authority are not always showing which areas are deprived. Local authorities such as Eilean Siar, Orkney Islands, and Shetland Islands do not have any datazones in the 15% most deprived in the 2009 SIMD. This happens due

to the small number of intermediate geographies they are comprised of and not because they are considered to be affluent areas.

This approach, of using a cut-off point for the datazones, has its drawbacks since datazones that just miss the 15% cut-off point are treated the same as the ones that are far away from it. A better approach, and the one followed in this paper, would be to weight every datazone and average over all the datazones that belong to the same intermediate geography. Datazones with a small SIMD rank (most deprived) will have a higher weight and each datazone's SIMD rank contributes for the deprivation of the intermediate geography they belong to.

The Scottish Government classifies urban and rural areas across Scotland based on two criteria: population and accessibility to areas of contiguous high population density postcodes (that make up what is known as a settlement). The joint classification can be seen in table 3. Using this classification as an ordinal covariate is not appropriate due to how it is coded. For example the 6th class (accessible rural area) is closer to an urban area than the previous two. Instead, we will use as covariates the percentages of those classes within each intermediate geography, e.g. if an intermediate geography is comprised of 6 datazones where 3 of them are coded as large urban areas and the other 3 as accessible small towns, the percentages of the first and the third class will be 50% and 0% for all the other classes. For ease of interpretation we center all covariates.

5.1 Regression model and results

Sellers & Shmueli (2010) propose a COM-Poisson regression model based on the (λ, ν) formulation whereas Guikema & Coffelt (2008) prefer to work with the (μ, ν) reformulation. Modifying the latter model we take into account the population and age structure of each intermediate geography and include expected counts (E_i) of hospital emergency admissions for each intermediate geography. The expected counts are computed using the age structure of each intermediate geography's population, together with estimates of the probabilities of hospitalisation in each age group. To account for the spatial autocorrelation of the data we use a conditional autoregressive model, which

is specified as follows

$$
\begin{aligned}
P(Y_i = y_i|\mu_i, \nu_i) &= \left(\frac{\mu_i^{y_i}}{y_i!}\right)^{\nu_i} \frac{1}{Z(\mu_i, \nu_i)}, \\
Z(\mu_i, \nu_i) &= \sum_{j=0}^{\infty} \left(\frac{\mu_i^{j}}{j!}\right)^{\nu_i}, \\
\log \frac{\mu_i}{E_i} &= \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \phi_i \Rightarrow \mathbb{E}[Y_i] \approx E_i \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \phi_i\}, \\
\log \nu_i &= -\boldsymbol{x}_i^\intercal \boldsymbol{c} \qquad \Rightarrow \mathbb{V}[Y_i] \approx E_i \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \phi_i + \boldsymbol{x}_i^\intercal \boldsymbol{c}\}.
\end{aligned} \tag{19}
$$

Y is the dependent random variable being modelled (emergency hospital admissions), E_i is the expected emergency hospital admissions for the i intermediate geography, ϕ_i are the random effects for the parameter μ, while $\boldsymbol{\beta}$ and $\boldsymbol{c}$ are the regression coefficients for the centering link function and the shape link function. Finally, the covariates $\boldsymbol{x}_i$ are comprised of: the deprivation weight of the intermediate geography i, the percentages of each urban/rural class within the intermediate geography i (using large urban areas as the baseline model), and 32 dummy variables that relate the intermediate geography i to its local authority.

The CAR prior being used for the random effects ϕ_i in this model is given by

$$
\phi_k|\phi_{-k} \sim \mathrm{N}\left(\frac{\rho \sum_{i=1}^{n} w_{ki}\phi_i}{\rho \sum_{i=1}^{n} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{i=1}^{n} w_{ki} + 1 - \rho}\right) \tag{20}
$$

and was proposed by Leroux et al. (2000) for modelling varying strengths of spatial autocorrelation. It can be seen as a generalisation of Besag et al. (1991) CAR prior where the first model can only represent strong spatial autocorrelation and produces smooth random effects. The random effects for non-neighbouring areas are conditionally independent given the values of the random effects of all the other areas. The parameter ρ can be seen as a spatial autocorrelation parameter, with $\rho = 0$ corresponding to independence, while $\rho = 1$ corresponds to strong spatial autocorrelation. In the first case there is an absence of spatial correlation in the data and the overdispersion is not caused by a spatial heterogeneity while in the second case all the overdispersion is due to the spatial autocorrelation. When $0 < \rho < 1$, the random effects are correlated and the data present a combination of spatial structured and unstructured components. Lee (2011) compares four of the most common CAR models and concludes that the model by Leroux et al. (2000) is the most appealing from both theoretical and practical standpoints.

In this formulation the coefficients have a direct link to either the mean or the variance, providing insight into the behavior of the dependent variable.

Larger values of $\boldsymbol{\beta}$ and $\boldsymbol{c}$ can be translated to higher mean and higher variance for the response variable. We implement a Bayesian approach for the previous model, and propose an efficient and exact MCMC algorithm based on the piecewise geometric bounds and the retrospective sampling algorithm. We use noninformative multivariate normal priors for the regression coefficients with a mean of zero and a variance of 10^6. A uniform prior on the unit interval is specified for ρ, and a uniform prior on the interval $(0, 1000)$ is adopted for τ^2. In addition, the proposal distribution q is chosen to be a multivariate normal centered at the current value. Thus, the second ratio in (5) cancels out due to the symmetry of the multivariate normal distribution. This algorithm is known as a random walk Metropolis-Hastings.

Table 4 shows the non-model-based regression coefficients for each local authority (32 local authorities in total). These coefficients refer to the intercepts of the 32 regression models (one for each local authority) where the offset (e.g. $\log E_i$) is the only covariate. It must be noted that some of the local authorities were comprised of a small number of data points, for example Orkney Islands, Shetland Islands, and Eilean Siar include less than 10 intermediate geographies. Table 5 shows the regression coefficients for the model in (19). The COM-Poisson coefficients for ν of most covariates are positive which is a sign of overdispersion. Table 5 shows that there is a wide range of values for the coefficients $\boldsymbol{c}$. They can take negative values (Orkney Islands) and up to greater than 2 (Dumfries & Galloway, Scottish Borders). The regression coefficients b_1, c_1 for the deprivation weights have positive posterior median estimates, 0.87 and 0.05 respectively, with $(0.84, 0.90)$ and $(-0.36, 0.52)$ as their 95% credible intervals. This translates to higher emergency hospital admissions for intermediate geographies with high deprivation. This is not true for the variance, since the credible interval includes negative values. The data have a strong spatial autocorrelation as can be seen, in Table 6, from the credible intervals of the autocorrelation parameter ρ.

Figure 3 shows the medians (plotted as diamonds) and the 95% credible intervals (plotted as lines) for the regression coefficients for both models. The black lines refer to the non-model-based coefficients whereas the red lines refer to the regression model including all covariates. In the top panel it can be seen that adjusting for the covariates (deprivation, urban/rural classification, and local authorities) shifts the regression coefficients towards zero. As we mentioned earlier, modelling the variance is the main interest in this application since it helps us identify areas with health inequalities. Comparing the panels in Figure 3 reveals a different pattern for the mean effects and variance effects of the local authorities. Local authorities with

large μ coefficients (corresponding to poor health) do not necessarily have large ν coefficients (corresponding to large health inequalities). This can be seen in local authorities such as North Ayrshire, North Lanarkshire and South Ayrshire.

The coefficients for the percentages of each class can be seen in Figure 4 where large urban areas are considered to be the baseline model. The remaining classes are plotted with regards to their distance from an urban area. The black circle represents the large urban area class whereas the blue, brown and violet lines represent the urban area, small town and rural area classes respectively. It can be seen that very remote small towns have higher (on average) emergency hospital admissions (see Panel 4a) and higher variance (see Panel 4b) compared to large urban areas.

Finally, Figure 5 shows the standardised incidence ratio of the average emergency hospital admissions using the non-model-based coefficients (on the left) and the coefficients of the full model (on the right).

R (R Core Team, 2014) was used for all the computations in this paper. Traceplots, density plots, autocorrelation plots (for every regression coefficient) and results for the Gelman and Rubin diagnostic, (Gelman & Rubin, 1992), were employed to assess convergence of the MCMC sampler to the posterior distribution, using the coda package (Plummer et al., 2006).

5.2 Range of bounds for the acceptance probability

The computational speed of the proposed technique depends on which strategy is chosen to refine the bounds. We chose to increase the number of terms that are computed exactly for the estimation of the normalisation constant and use the piecewise geometric bounds for the remaining terms. We start by computing exactly 240 terms for every $Z(\mu_i, \nu_i)$ and every time the bounds needed to be refined we computed 100 more terms for the observations that have a large difference between the upper and lower bound. One can increase the precision by a specified amount every time a refinement is necessary. Tables 7 and 8 show the percentages of every possible outcome (acceptance, rejection, or further refinement) for different number of refinements. Almost half of the time, for both parameters, there is no need to refine the bounds and we can make an accept/reject decision on just computing 240 terms. Tables 9 and 10 show the mean values for the difference of the log bounds. One can see that, even when a large number of refinements is needed in order to make a decision, the difference of the log bounds is still large. This shows that there is no need to be very precise in our approximations. The weighted averages for the log differences and the computed terms are also shown. The

weighted average for the log difference of the bounds when the MCMC rejects the candidate value for the first parameter is 2.51 which means that a rejection decision can often be reached with very loose upper and lower bounds.

6 Conclusion

In this paper we have focused on the problem of computing the normalising constant in the probability mass function of a discrete random variable, when it is not available in closed form and its computation is demanding. We have shown that if the ratios of consecutive probabilities can be bounded over ranges of the random variable, lower and upper bounds on the normalising constant can be derived, leading to an exact and fast MCMC algorithm in a Bayesian inferential setting. Furthermore, the results show that in order for the MCMC algorithm to make a decision between accepting or rejecting a candidate move, the approximation of the bounds of the acceptance probability does not need to be precise. We applied this method to emergency hospital admissions data in Scotland, using the COM-Poisson distribution. We concentrated on the COM-Poisson distribution, a very flexible generalisation of the Poisson distribution, since it allows modelling the mean and the variance explicitly. As a result, we were able to identify areas with a high level of health inequalities.

7 Acknowledgements

We would like to thank Peter Craigmile, Chris Holmes, Dirk Husmeier, and Duncan Lee for their helpful comments.

References

Besag, J, York, J & Mollié, A (1991), 'Bayesian image restoration, with two applications in spatial statistics,' *Annals of the Institute of Statistical Mathematics*, **43**(1), pp. 1–20, doi:10.1007/BF00116466.

Beskos, A, Papaspiliopoulos, O, Roberts, GO & Fearnhead, P (2006), 'Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion),' *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(3), pp. 333–382, doi:10.1111/j.1467-9868.2006.00552.x.

Bivand, R (2014), *spdep: Spatial dependence: weighting schemes, statistics and models*, r package version 0.5-71.

Bivand, R & Lewin-Koh, N (2013), *maptools: Tools for reading and handling spatial objects*, r package version 0.8-27.

Conway, RW & Maxwell, WL (1962), 'A queuing model with state dependent service rate,' *Journal of Industrial Engineering*, **12**, pp. 132–136.

Del Castillo, J & Pérez-Casany, M (1998), 'Weighted poisson distributions for overdispersion and underdispersion situations,' *Annals of the Institute of Statistical Mathematics*, **50**(3), pp. 567–585, doi:10.1023/A:1003585714207.

Eddelbuettel, D & François, R (2011), 'Rcpp: Seamless R and C++ integration,' *Journal of Statistical Software*, **40**(8), pp. 1–18.

Gelman, A & Rubin, DB (1992), 'Inference from iterative simulation using multiple sequences.' *Statistical Science*, **7**(4), pp. 457–472.

Guikema, SD & Coffelt, JP (2008), 'A flexible count data regression model for risk analysis.' *Risk analysis: an official publication of the Society for Risk Analysis*, **28**, pp. 213–223, doi:10.1111/j.1539-6924.2008.01014.x.

Lee, D (2011), 'A comparison of conditional autoregressive models used in Bayesian disease mapping.' *Spatial and Spatio-temporal Epidemiology*, **2**(2), pp. 79–89.

Lee, D (2013), 'CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors,' *Journal of Statistical Software*, **55**(13), pp. 1–24.

Leroux, B, Lei, X & Breslow, N (2000), 'Estimation of disease rates in small areas: A new mixed model for spatial dependence,' in Halloran, M & Berry, D (eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Springer New York, vol. 116 of *The IMA Volumes in Mathematics and its Applications*, pp. 179–191, doi:10.1007/978-1-4612-1284-3_4.

Minka, TP, Shmueli, G, Kadane, JB, Borle, S & Boatwright, P (2003), 'Computing with the COM-Poisson distribution,' Tech. rep., CMU Statistics Department.

Papaspiliopoulos, O & Roberts, GO (2008), 'Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models,' *Biometrika*, **95**, pp. 169–186, doi:10.1093/biomet/asm086.

Plummer, M, Best, N, Cowles, K & Vines, K (2006), 'Coda: Convergence diagnostics and output analysis for mcmc,' *R News*, **6**(1), pp. 7–11.

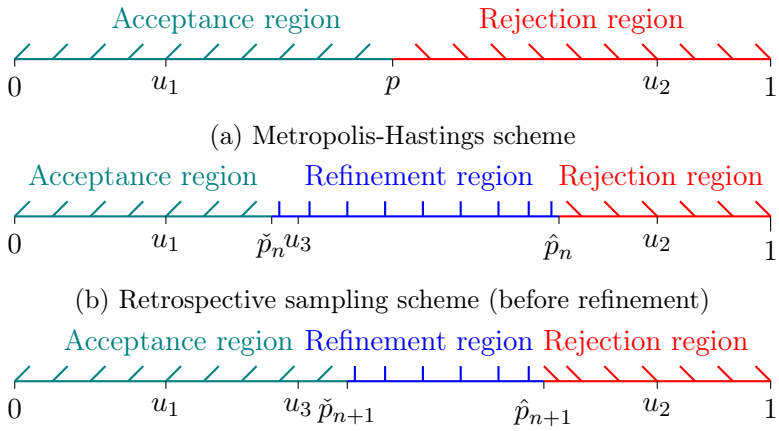
R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Scottish Government (2009), 'Scottish index of multiple deprivation,' doi: http://dx.doi.org/10.5255/UKDA-SN-6871-1.

Sellers, KF & Shmueli, G (2010), 'A flexible regression model for count data,' *Annals of Applied Statistics*, **4**(2), pp. 943–961, doi:10.1214/09-aoas306.

Sermaidis, G, Papaspiliopoulos, O, Roberts, GO, Beskos, A & Fearnhead, P (2013), 'Markov chain Monte Carlo for exact inference for diffusions,' *Scandinavian Journal of Statistics*, **40**(2), pp. 294–321, doi: 10.1111/j.1467-9469.2012.00812.x.

Shmueli, G, Minka, TP, Kadane, JB, Borle, S & Boatwright, P (2005), 'A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution,' *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, pp. 127–142, doi:10.1111/j.1467-9876.2005.00474.x.



(a) Metropolis-Hastings scheme

(b) Retrospective sampling scheme (before refinement)

(c) Retrospective sampling scheme (after refinement)

Figure 1: Illustration of the retrospective sampling algorithm (panels b and c) in contrast to the standard Metropolis-Hastings algorithm (panel a).

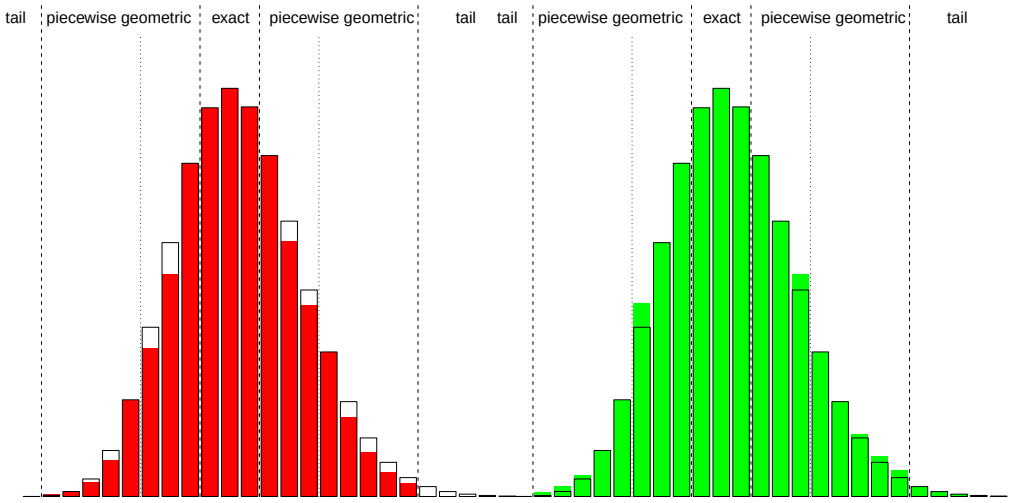


Figure 2: Computing the lower and upper bounds of the normalisation constant in blocks of probabilities. We first compute exactly the probabilities close to the mode of the distribution, and then compute the lower and upper bounds in blocks of probabilities where we only have to compute exactly the first probability of each block.

Table 1: The local share considers the percentage of a local authority's datazones that are amongst the 15% most deprived in Scotland.

Local Authorities	Local Share	Local Authorities	Local Share
Aberdeen City	10.49	Highland	5.48
Aberdeenshire	1.33	Inverclyde	38.18
Angus	4.23	Midlothian	3.57
Argyll & Bute	8.20	Moray	0.86
Clackmannanshire	18.75	North Ayrshire	24.02
Dumfries & Galloway	5.70	North Lanarkshire	21.29
Dundee City	30.17	Orkney Islands	0.00
East Ayrshire	17.53	Perth & Kinross	3.43
East Dunbartonshire	3.15	Renfrewshire	20.09
East Lothian	2.50	Scottish Borders	3.85
East Renfrewshire	4.17	Shetland Islands	0.00
Edinburgh, City of	10.93	South Ayrshire	12.24
Eilean Siar	0.00	South Lanarkshire	14.57
Falkirk	8.63	Stirling	6.36
Fife	11.26	West Dunbartonshire	26.27
Glasgow City	43.52	West Lothian	9.00

Table 2: The local share considers the percentage of a local authority's datazones that are amongst the 15% least deprived in Scotland.

Local Authorities	Local Share	Local Authorities	Local Share
Aberdeen City	35.58	Highland	4.79
Aberdeenshire	21.93	Inverclyde	2.73
Angus	9.86	Midlothian	14.29
Argyll & Bute	4.10	Moray	8.62
Clackmannanshire	9.38	North Ayrshire	2.79
Dumfries & Galloway	4.15	North Lanarkshire	6.22
Dundee City	11.17	Orkney Islands	0.00
East Ayrshire	6.49	Perth & Kinross	12.57
East Dunbartonshire	47.24	Renfrewshire	17.76
East Lothian	18.33	Scottish Borders	4.62
East Renfrewshire	57.50	Shetland Islands	0.00
Edinburgh, City of	39.53	South Ayrshire	16.33
Eilean Siar	0.00	South Lanarkshire	12.31
Falkirk	13.20	Stirling	18.18
Fife	12.36	West Dunbartonshire	2.54
Glasgow City	4.76	West Lothian	16.11

Table 3: Scottish Government joint Urban/Rural Classification.

Class	Class name	Description
1	Large Urban Areas	settlements of over 125.000 people.
2	Other Urban Areas	settlements of 10.000 to 125.000 people.
3	Accessible Small Towns	settlements of between 3.000 and 10.000 people, and within a 30 minute drive time to a settlement of 10.000 or more.
4	Remote Small Towns	settlements of between 3.000 and 10.000 people, and with a drive time between 30 and 60 minutes to a settlement of 10.000 or more.
5	Very Remote Small Towns	settlements of between 3.000 and 10.000 people, and with a drive time of over 60 minutes to a settlement of 10.000 or more.
6	Accessible Rural Areas	Areas with a population of less than 3.000 people, and within a 30 minute drive time to a settlement of 10.000 or more.
7	Remote Rural Areas	Areas with a population of less than 3.000 people, and with a drive time between 30 and 60 minutes to a settlement of 10.000 or more.
8	Very Remote Rural Areas	Areas with a population of less than 3.000 people, and with a drive time of over 60 minutes to a settlement of 10.000 or more.

Table 4: Posterior medians of the non-model-based regression coefficients for each local authority.

Local Authorities	β_0	c_0	Local Authorities	β_0	c_0
Aberdeen City	-0.03	3.59	Highland	-0.06	3.41
Aberdeenshire	-0.26	2.10	Inverclyde	0.18	3.62
Angus	-0.16	2.15	Midlothian	-0.13	2.31
Argyll & Bute	-0.07	3.05	Moray	-0.25	2.27
Clackmannanshire	-0.18	2.57	North Ayrshire	0.18	3.06
Dumfries & Galloway	-0.18	3.23	North Lanarkshire	0.18	2.93
Dundee City	0.04	3.14	Orkney Islands	-0.17	2.29
East Ayrshire	0.16	3.24	Perth & Kinross	-0.13	3.02
East Dunbartonshire	-0.15	3.14	Renfrewshire	0.06	3.59
East Lothian	-0.25	2.48	Scottish Borders	0.01	3.10
East Renfrewshire	-0.26	2.99	Shetland Islands	-0.18	3.23
Edinburgh, City of	-0.31	3.69	South Ayrshire	0.11	3.10
Eilean Siar	-0.02	2.26	South Lanarkshire	0.01	2.54
Falkirk	-0.15	2.26	Stirling	-0.21	3.43
Fife	-0.14	2.75	West Dunbartonshire	0.13	2.57
Glasgow City	0.20	3.59	West Lothian	0.09	3.08

Table 5: Posterior medians for the regression coefficients of the full model.

Covariates	β_i	c_i	Covariates	β_i	c_i
Deprivation weight	0.87	0.05	Eilean Siar	0.00	0.39
Other urban area	0.00	-0.31	Falkirk	-0.19	1.48
Accesible small town	-0.04	0.06	Fife	-0.14	1.40
Remote small town	-0.04	-0.12	Glasgow City	0.03	1.72
Very remote small town	0.15	0.91	Highland	0.00	1.87
Accesible rural area	-0.08	-0.01	Inverclyde	0.04	1.39
Remote rural area	-0.19	0.11	Midlothian	-0.11	1.39
Very remote rural area	-0.09	0.51	Moray	-0.12	1.44
Aberdeen City	0.05	1.26	North Ayrshire	0.07	1.19
Aberdeenshire	-0.05	1.49	North Lanarkshire	0.06	1.08
Angus	-0.11	1.93	Orkney Islands	-0.11	-0.33
Argyll & Bute	-0.02	1.62	Perth & Kinross	0.03	1.26
Clackmannanshire	-0.21	1.68	Renfrewshire	0.04	1.70
Dumfries & Galloway	-0.10	2.42	Scottish Borders	0.12	2.38
Dundee City	-0.07	1.58	Shetland Islands	-0.13	1.04
East Ayrshire	0.07	1.87	South Ayrshire	0.12	1.17
East Dunbartonshire	0.05	1.57	South Lanarkshire	-0.01	1.45
East Lothian	-0.16	0.69	Stirling	-0.08	1.75
East Renfrewshire	-0.03	1.29	West Dunbartonshire	-0.03	1.05
Edinburgh, City of	-0.20	1.92	West Lothian	0.05	1.68

Table 6: Posterior medians for the variance and spatial autocorrelation of the random effects.

	Median	2.5%	97.5%
τ^2	0.004	0.002	0.008
ρ	0.927	0.783	0.971

Table 7: Percentages for refinements when updating the parameter μ.

Refinements	Accepted	Rejected	Still need refinement
0	9.6%	43.2%	47.2%
1	11.5%	20.1%	15.6%
≥ 2	7.3%	8.3%	0%
Total	28.4%	71.6%	

Table 8: Percentages for refinements when updating the parameter ν.

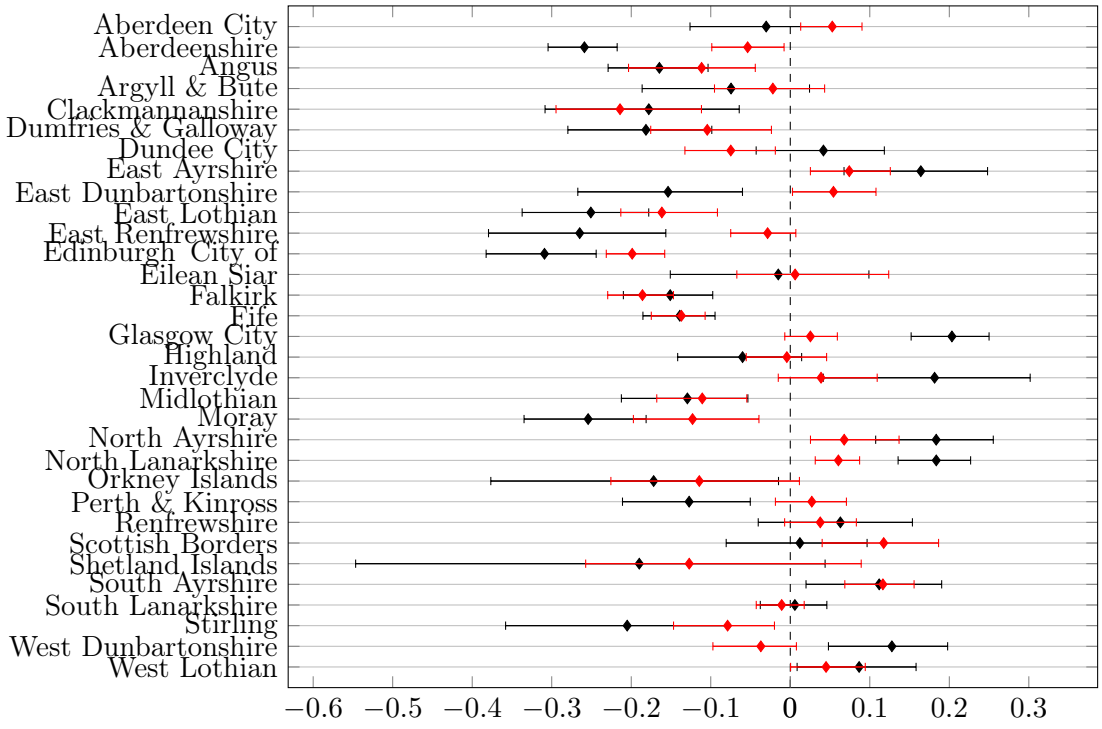
Refinements	Accepted	Rejected	Still need refinement
0	11.9%	27.2%	60.9%
1	16.3%	23.8%	20.8%
≥ 2	9.6%	11.2%	0%
Total	37.8%	62.2%	

Table 9: Mean values for the difference of the log bounds and the computed terms when updating the parameter μ.

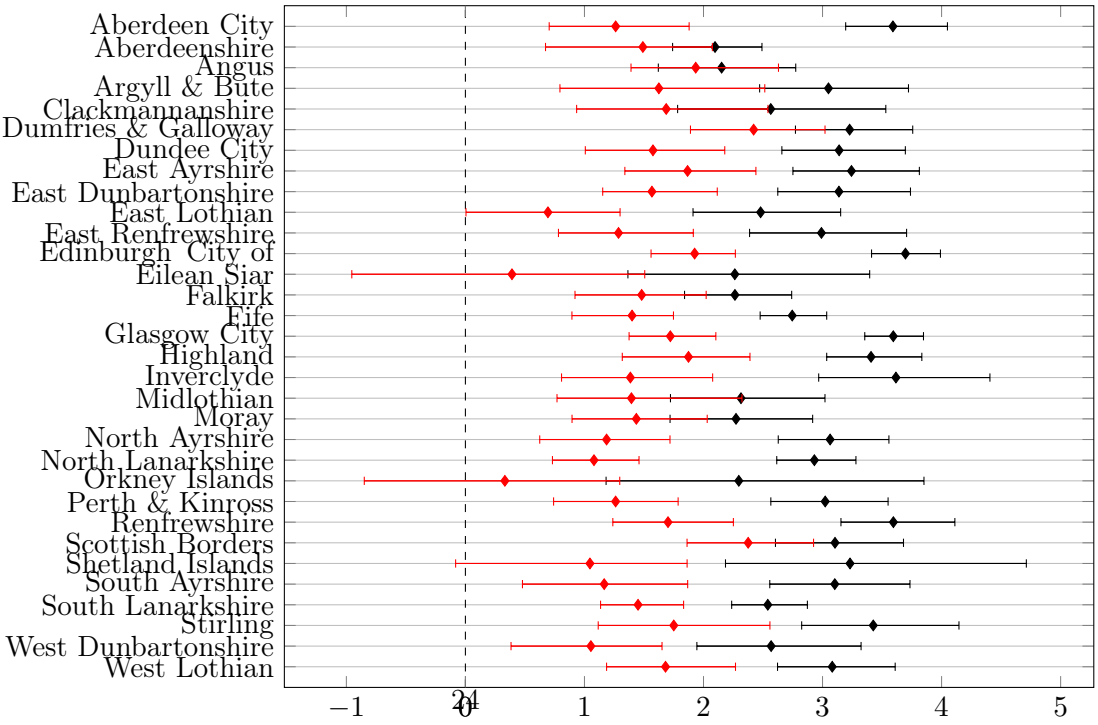
Refinements	Accepted	Acceptance $\log \hat{p}_n - \log \check{p}_n$	Computed terms	Rejected	Rejection $\log \hat{p}_n - \log \check{p}_n$	Computed terms
0	33.9%	3.51	240	60.3%	3.55	240
1	40.5%	1.11	264.71	28.1%	1.13	264.66
≥ 2	25.6%	0.40	426.15	11.6%	0.43	420.55
Weighted average		1.74	297.84		2.51	267.75

Table 10: Mean values for the difference of the log bounds and the computed terms when updating the parameter ν.

Refinements	Accepted	Acceptance $\log \hat{p}_n - \log \check{p}_n$	Computed terms	Rejected	Rejection $\log \hat{p}_n - \log \check{p}_n$	Computed terms
0	31.5%	3.52	240	43.7%	3.55	240
1	43.1%	1.11	264.72	38.2%	1.13	264.70
≥ 2	25.4%	0.41	425.16	18.1%	0.43	420.06
Weighted average		1.69	297.82		2.07	281.52

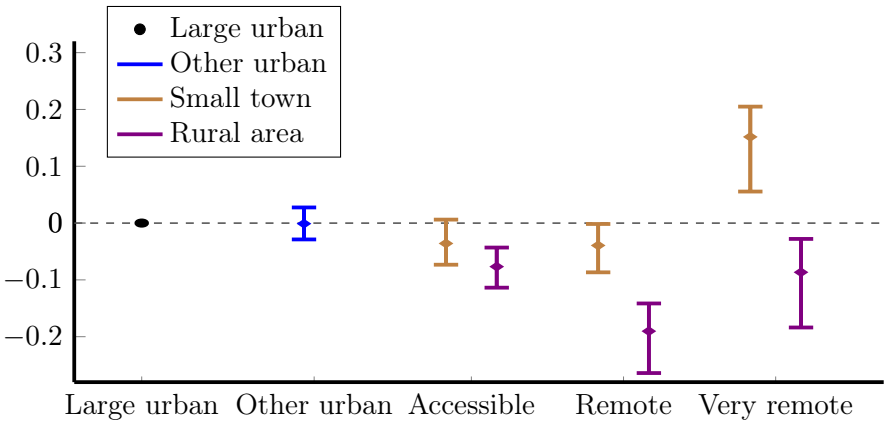


(a) Regression coefficients for μ

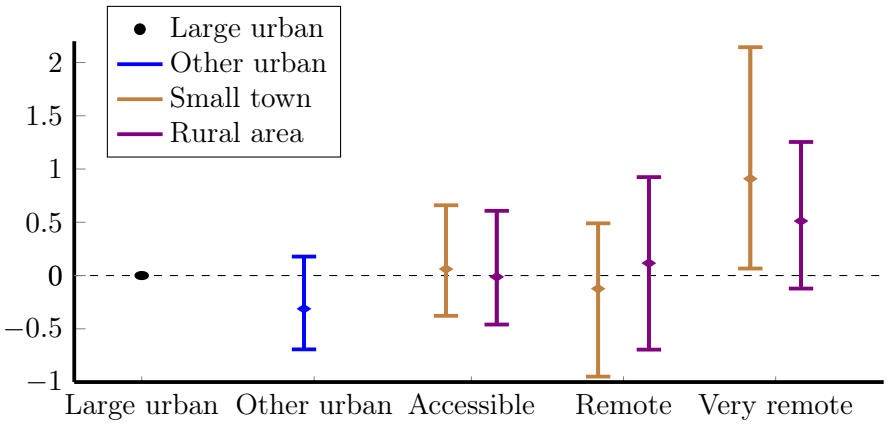


(b) Regression coefficients for ν

Figure 3: Credible intervals for the regression coefficients of the local authorities. The non-model-based regression coefficients of table 4 are shown in black and the full model of table 5 in red.



(a) Regression coefficients for μ



(b) Regression coefficients for ν

Figure 4: Credible intervals for the regression coefficients for each class in table 3.

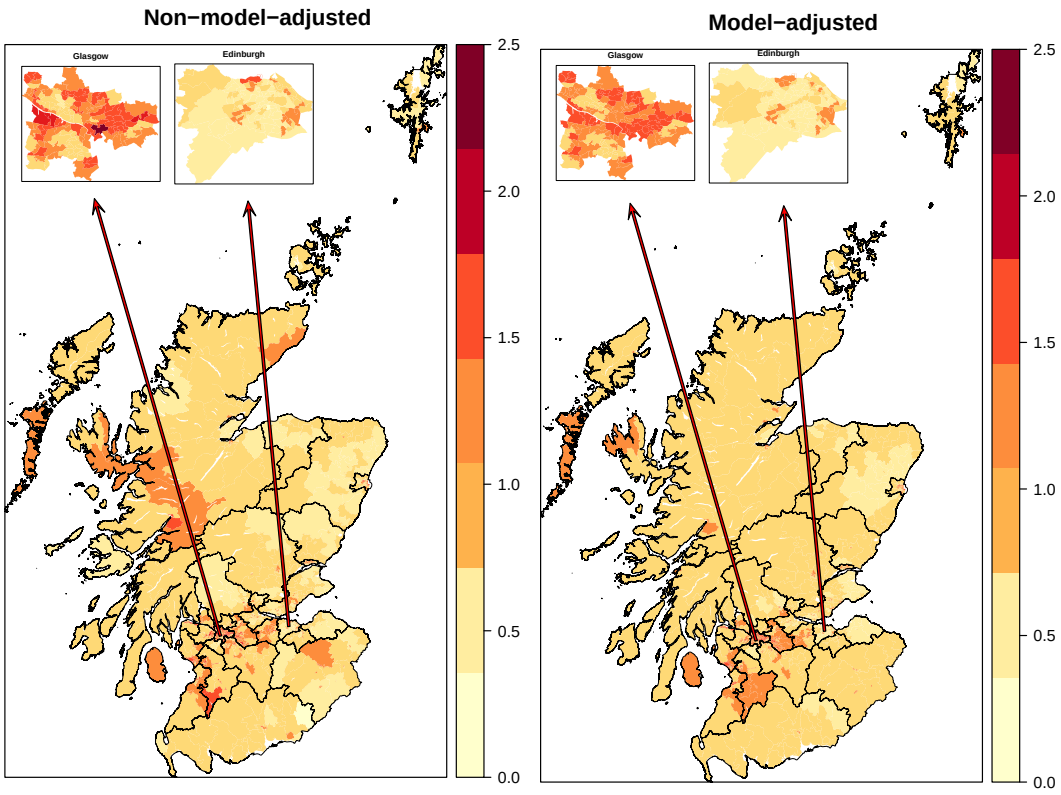


Figure 5: SIR for emergency hospital admissions.