

Fast Bayesian Estimation of Spatial Count Data Models

15 October 2020

PRATEEK BANSAL* (corresponding author)

Transport Strategy Centre, Department of Civil and Environmental Engineering
Imperial College London, UK
prateek.bansal@imperial.ac.uk

RICO KRUEGER*

Transport and Mobility Laboratory
Ecole Polytechnique Fédérale de Lausanne, Switzerland
rico.krueger@epfl.ch

DANIEL J. GRAHAM

Transport Strategy Centre, Department of Civil and Environmental Engineering
Imperial College London, UK
d.j.graham@imperial.ac.uk

* Equal contribution.

arXiv:2007.03681v2 [stat.ME] 16 Oct 2020

Abstract

Spatial count data models are used to explain and predict the frequency of phenomena such as traffic accidents in geographically distinct entities such as census tracts or road segments. These models are typically estimated using Bayesian Markov chain Monte Carlo (MCMC) simulation methods, which, however, are computationally expensive and do not scale well to large datasets. Variational Bayes (VB), a method from machine learning, addresses the shortcomings of MCMC by casting Bayesian estimation as an optimisation problem instead of a simulation problem. Considering all these advantages of VB, a VB method is derived for posterior inference in negative binomial models with unobserved parameter heterogeneity and spatial dependence. Pólya-Gamma augmentation is used to deal with the non-conjugacy of the negative binomial likelihood and an integrated non-factorised specification of the variational distribution is adopted to capture posterior dependencies. The benefits of the proposed approach are demonstrated in a Monte Carlo study and an empirical application on estimating youth pedestrian injury counts in census tracts of New York City. The VB approach is around 45 to 50 times faster than MCMC on a regular eight-core processor in a simulation and an empirical study, while offering similar estimation and predictive accuracy. Conditional on the availability of computational resources, the embarrassingly parallel architecture of the proposed VB method can be exploited to further accelerate its estimation by up to 20 times.

Keywords: Variational Bayes; spatial count data; negative binomial regression; Pólya-Gamma data augmentation; accident analysis.

1. Introduction

Spatial count data models are widely used in disciplines such as ecology, epidemiology, geography, regional science as well as transportation planning and engineering to explain and predict non-negative integer-valued outcome variables such as species and disease counts, patenting and innovation activities as well as crime and accident rates in geographically distinct entities such as local government areas, census tracts or traffic analysis zones (e.g. [Acs et al., 2002](#); [Dormann et al., 2007](#); [Glaser, 2017](#); [Marshall, 1991](#); [Ver Hoef et al., 2018](#); [Wakefield, 2007](#)).

Models of spatial count data typically pivot on Poisson lognormal and negative binomial regressions, in which the spatial arrangement of the investigated units is explicitly specified. These models generally consider two types of spatial effects, namely *spatial heterogeneity* and *spatial dependence* ([Simões and Natário, 2016](#)). While *spatial heterogeneity* accounts for the spatially-varying effect of covariates on the dependent variable, *spatial dependence* captures the systematic correlation across neighbouring spatial units. In spatial count data models, unobserved spatial heterogeneity is operationalised through the inclusion of random link function parameters ([Mannering et al., 2016](#)); spatial dependence can be represented through different variants of autoregressive specifications including the spatial and conditional autoregressive and matrix exponential spatial specifications ([Whittle, 1954](#); [Besag, 1974](#); [LeSage and Pace, 2007](#)). Ignoring these spatial effects may result in biased parameter estimates and inaccurate inference due to higher type-I error ([Anselin, 2013](#); [Dormann, 2007](#); [Dormann et al., 2007](#)). However, accounting for spatial heterogeneity and dependence also renders the estimation of spatial count data models computationally expensive.

Spatial count data models are predominantly estimated using Markov Chain Monte Carlo (MCMC) methods ([Banerjee et al., 2014](#); [Haining and Li, 2020](#)), aside from few exceptions which rely on maximum likelihood estimation ([Castro et al., 2012](#); [Narayanamoorthy et al., 2013](#)). MCMC methods guarantee asymptotically exact inference, but succumb to three important limitations, namely computationally intensive estimation, high storage costs for the posterior draws, and difficulties in assessing convergence ([Bansal et al., 2020](#)). Furthermore, state-of-practice Gibbs samplers for spatial count data models also include Metropolis-Hastings steps to sample from high-dimensional conditional distributions, since conjugate priors for the parameters of Poisson lognormal and negative binomial regressions are not known. Sampling via the Metropolis-Hastings algorithm suffers from a variety of inefficiencies including insufficient exploration of the posterior of interest and serial correlation, if it is not tuned well ([Rossi et al., 2012](#)).

To address the bottlenecks of MCMC in the estimation of spatial econometric models, [Bivand et al. \(2014\)](#) propose the integrated nested Laplace approximation (INLA) method, under which the model parameters are first segregated into hyper-parameters and latent variables. Then, a discrete distribution is specified on the hyper-parameters using a multi-dimensional grid, and the posterior distribution of the latent variables is approximated via Laplace's method. This analytical approximation comes at the cost of the assumption that conditional on the hyper-parameters, the latent variables are normally distributed. INLA reduces the estimation times of typical spatial econometric models from hours to minutes, but the conditional normality assumption restricts the flexibility of the posterior approximation ([Han et al., 2013](#)).

In machine learning and computational statistics, variational Bayes (VB) methods have also emerged as a promising alternative to MCMC for the estimation of complex econometric models ([Bansal et al., 2020](#); [Blei et al., 2017](#); [Braun and McAuliffe, 2010](#); [Jordan et al., 1999](#); [Tan et al., 2013](#)). Whilst

MCMC treats Bayesian inference as a simulation problem, in which the posterior distribution of interest is approximated through samples from a Markov chain, VB recasts Bayesian inference into an optimisation problem, which consists of minimising the probability distance between an approximating variational distribution and the targeted posterior distribution. Translating Bayesian inference into an optimisation problem accelerates estimation, admits a straightforward assessment of convergence and alleviates storage requirements.

VB methods have been introduced for the estimation of non-spatial count data models and of linear spatial models. Yet, no VB method exists for the estimation of spatial count data models. Several studies present VB methods for variants of count data models, but none of the proposed approaches accounts for spatial dependencies between units (Klami, 2015; Luts et al., 2015; Tan et al., 2013; Zhou et al., 2012). Kabisa et al. (2016), Ren et al. (2011) and Wu (2018) devise VB methods for the estimation of models with spatial dependence; however, the proposed methods are limited to linear models with continuous outcome variables.

In this paper, we propose a VB method for the fast estimation of a spatial count data model, which accommodates both spatial heterogeneity and dependence. To be specific, we consider a negative binomial (NB) model with random link function parameters and a matrix exponential spatial specification of spatial dependence (LeSage and Pace, 2007). To address the non-conjugacy of the NB model, we also adopt the Pólya-Gamma data augmentation (PGDA) technique in the proposed inference method. PDGA introduces auxiliary latent variables into the models. Conditional on these variables, the NB likelihood of the observed counts is translated into a heteroskedastic Gaussian likelihood, which admits closed-form conjugate posterior updates for nearly all model parameters. Only a few studies employ the PGDA technique in VB estimation (Durante et al., 2019; Klami, 2015; Park et al., 2016; Wenzel et al., 2019; Zhou et al., 2012).

We first derive a mean-field variational Bayes (MFVB) method, which posits a factorised representation of the joint variational distributions, for the Pólya-Gamma-augmented spatial NB model. MFVB is the workhorse approach for the specification of the approximating variational distribution in VB inference. However, in the current application, the mean-field assumption oversimplifies posterior dependencies and leads to a high bias in the recovery of the spatial model parameters. Alternatively, the variational distribution can be specified according to the integrated non-factorised variational Bayes (INFVB; Han et al., 2013) approach, which generalises INLA by relaxing the conditional normality assumption. Motivated by the superior finite sample properties of INFVB for linear spatial models, we devise an INFVB method to allow for richer representations of relevant posterior dependencies in the considered spatial count data model. We benchmark the performance of INFVB against MCMC using simulated data and real data on youth pedestrian injury counts in New York City. The results indicate that INFVB is able to emulate the performance of MCMC in terms of posterior recovery and in-sample predictive accuracy. Furthermore, the embarrassingly parallel nature of the proposed INFVB algorithm makes INFVB substantially faster than MCMC, which, in turn, suggests that INFVB is scalable to large datasets of spatial counts.

We organise the remainder of the paper as follows. In the subsequent section, we formulate the considered spatial negative binomial model, and in Section 3, we derive MCMC and VB estimators for the model. In Section 4, we benchmark computational efficiency and finite sample properties of the proposed estimators in a Monte Carlo study. Section 5 further compares VB and MCMC in estimating youth pedestrian injury counts in the census tracts of New York City. The findings of this empirical

application corroborate the insights derived from the simulation study. Conclusions and avenues for future research are presented in Section 6.

2. Model formulation

Let y_i denote the non-negative integer-valued outcome variable observed for spatial unit $i \in \{1, \dots, N\}$. We assume that y_i is drawn from a negative binomial (NB) distribution with probability parameter p_i and shape parameter r . We model p_i , using a logit link function, which depends on predictors \mathbf{M}_i with fixed parameters $\boldsymbol{\gamma}$, predictors \mathbf{X}_i with spatially-varying parameters $\boldsymbol{\beta}_i$ and a spatial random effect ϕ_i . The resulting NB model is succinctly summarised below:

$$y_i \sim \text{NB}(r, p_i), \quad i = 1, \dots, N \quad (1)$$

$$p_i = \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}, \quad i = 1, \dots, N \quad (2)$$

$$\psi_i = \mathbf{M}_i^\top \boldsymbol{\gamma} + \mathbf{X}_i^\top \boldsymbol{\beta}_i + \phi_i. \quad i = 1, \dots, N \quad (3)$$

2.1. Spatial heterogeneity and dependence

To accommodate spatial heterogeneity in the model, i.e. to allow for spatially varying effects of \mathbf{X}_i on y_i , we place a multivariate Gaussian prior on $\boldsymbol{\beta}_i$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Furthermore, we apply the matrix exponential spatial specification (MESS; [LeSage and Pace, 2007](#)) to the random effect vector $\boldsymbol{\phi} = (\phi_1, \dots, \phi_N)^\top$ to capture spatial dependence between units. MESS is an attractive representation of spatial error dependence, as it implies a simple likelihood. Alternative specifications spatial dependence such as the spatial and conditional autoregressive ones, are similar to MESS with the key difference that MESS assumes an exponential decay instead of a geometric decay of spatial correlation (see [Strauss et al., 2017](#), for a detailed comparison). The spatial aspects of the considered model are succinctly restated below:

$$\boldsymbol{\beta}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, \dots, N \quad (4)$$

$$\mathbf{S}\boldsymbol{\phi} = \exp(\tau \mathbf{W})\boldsymbol{\phi} = \boldsymbol{\epsilon}, \quad (5)$$

$$\boldsymbol{\epsilon} \sim \text{Normal}(0, \sigma^2 \mathbf{I}_N). \quad (6)$$

Here, \mathbf{W} is a row-normalised spatial weight matrix, τ is the spatial association parameter, $\boldsymbol{\epsilon}$ is a homoskedastic Gaussian error with scale σ , and \mathbf{I}_N is an identity matrix of size $N \times N$. $\exp(\tau \mathbf{W})$ is a matrix of size $N \times N$ given by a power series: $\sum_{k=0}^{\infty} \frac{\tau^k}{k!} \mathbf{W}^k$, where \mathbf{W}^0 is an identity matrix. We compute this matrix exponential using the Pade approximation ([Al-Mohy and Higham, 2010](#)).

2.2. Model likelihood

Suppose that there are Q fixed parameters and K random parameters. Equation 3 can be rewritten in vector form as follows:

$$\boldsymbol{\psi} = \mathbf{M}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\phi}, \quad (7)$$

where

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_N \end{bmatrix}_{N \times 1}, \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_1^\top \\ \vdots \\ \mathbf{M}_N^\top \end{bmatrix}_{N \times Q}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{X}_N^\top \end{bmatrix}_{N \times NK}, \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix}_{NK \times 1}.$$

Furthermore, note that $\tilde{\boldsymbol{\Omega}} = \frac{\mathbf{S}^\top \mathbf{S}}{\sigma^2}$ and $\det(\mathbf{S}) = 1$, and thus, $\det(\tilde{\boldsymbol{\Omega}}) = (\sigma^2)^{-N}$ (Wu, 2018). Consequently, the likelihood of the model is:

$$\begin{aligned} P(\mathbf{y}|r, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma^2, \tau) &= P(\mathbf{y}|r, \boldsymbol{\psi})P(\boldsymbol{\psi}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\phi})P(\boldsymbol{\phi}|\sigma^2, \tau)P(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ &= P(\mathbf{y}|r, \boldsymbol{\psi})P(\boldsymbol{\psi}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2, \tau)P(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{aligned} \quad (8)$$

where

$$\begin{aligned} P(\mathbf{y}|r, \boldsymbol{\psi}) &= \prod_{i=1}^N \frac{\Gamma(y_i + r)}{\Gamma(r)y_i!} \frac{\exp(\psi_i)^{y_i}}{[1 + \exp(\psi_i)]^{r+y_i}}, \\ P(\boldsymbol{\psi}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2, \tau) &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{[\boldsymbol{\psi} - \mathbf{M}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\beta}]^\top \tilde{\boldsymbol{\Omega}}[\boldsymbol{\psi} - \mathbf{M}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\beta}]}{2}\right), \\ P(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= [2\pi\det(\boldsymbol{\Sigma})]^{-\frac{N}{2}} \prod_{i=1}^N \exp\left(-\frac{1}{2}[\boldsymbol{\beta}_i - \boldsymbol{\mu}]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\beta}_i - \boldsymbol{\mu}]\right). \end{aligned} \quad (9)$$

3. Model estimation

3.1. Pólya-Gamma data augmentation

Conjugate priors for the parameters of the NB model are generally unknown. As a consequence, the conditional distributions of the link function parameters and the shape parameter do not constitute known distributions, and no closed-form updates for the respective model parameters exist (Klami, 2015; Zhou et al., 2012). To address this issue, Polson et al. (2013) suggest to introduce Pólya-Gamma-distributed auxiliary variables $\omega_i \sim \text{PG}(y_i + r, 0)$, $i \in \{1, 2, \dots, N\}$ into the model. Using the identity derived by Polson et al. (2013), $P(\mathbf{y}|r, \boldsymbol{\psi})$ can be written as:

$$P(\mathbf{y}|r, \boldsymbol{\psi}) = \prod_{i=1}^N \frac{\Gamma(y_i + r)}{\Gamma(r)y_i!} 2^{-(r+y_i)} \exp\left(\frac{(y_i - r)\psi_i}{2}\right) \mathbb{E}_{\omega_i} \left[\exp\left(\frac{-\omega_i \psi_i^2}{2}\right) \right]. \quad (10)$$

Furthermore, conditional on the auxiliary variables $\boldsymbol{\omega}$, equation 10 can be restated as:

$$\begin{aligned} P(\mathbf{y}|\boldsymbol{\psi}, r, \boldsymbol{\omega}) &\propto \prod_{i=1}^N \exp\left(-\frac{\omega_i}{2} \left[\psi_i - \frac{y_i - r}{2\omega_i}\right]^2\right), \\ P(\mathbf{y}|\boldsymbol{\psi}, r, \boldsymbol{\omega}) &\propto \exp\left(-\frac{1}{2}[\boldsymbol{\psi} - \mathbf{Z}]^\top \boldsymbol{\Omega}[\boldsymbol{\psi} - \mathbf{Z}]\right), \end{aligned} \quad (11)$$

where

$$\mathbf{Z} = \begin{bmatrix} \frac{y_1 - r}{2\omega_1} \\ \vdots \\ \frac{y_N - r}{2\omega_N} \end{bmatrix}_{N \times 1}, \quad \boldsymbol{\Omega} = \begin{bmatrix} \omega_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \omega_N \end{bmatrix}_{N \times N},$$

$$\mathbf{Z} = \boldsymbol{\psi} + \boldsymbol{\alpha} = \mathbf{M}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\phi} + \boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Omega}^{-1}). \quad (12)$$

The main result of Pólya-Gamma data augmentation is that conditional on r and $\boldsymbol{\omega}$, the likelihood of the observed counts is converted into a heteroskedastic Gaussian likelihood, which considers \mathbf{Z} as outcome variable. As a consequence, we are able to obtain closed-form updates for the link function parameters and the shape parameter of the spatial NB model.

3.2. Prior specification and augmented likelihood

Prior distributions on latent variables are succinctly stated below:

$$\begin{aligned} \boldsymbol{\mu} &\sim \text{Normal}(\boldsymbol{\zeta}_\mu, \boldsymbol{\Delta}_\mu), & \boldsymbol{\gamma} &\sim \text{Normal}(\boldsymbol{\zeta}_\gamma, \boldsymbol{\Delta}_\gamma), & \tau &\sim \text{Normal}(\boldsymbol{\zeta}_\tau, \sigma_\tau^2), \\ \sigma^{-2} &\sim \text{Gamma}(b_{\sigma^2}, c_{\sigma^2}), & r|h &\sim \text{Gamma}(r_0, h), & h &\sim \text{Gamma}(b_0, c_0), \\ \{a_k\}_{k=1}^K &\sim \text{Gamma}(s, \eta_k), & \boldsymbol{\Sigma}|\mathbf{a} &\sim \text{IW}(\rho, \mathbf{B}), \end{aligned}$$

where $\rho = \nu + K - 1$, $\mathbf{a} = [a_1 \ \dots \ a_K]^\top$, $\mathbf{B} = 2\nu \text{diag}(\mathbf{a})$, $s = \frac{1}{2}$ and $\eta_k = A_k^{-2}$. We specify Huang's half-t prior on the covariance matrix of random parameters $\boldsymbol{\Sigma}$ by introducing \mathbf{a} (Huang et al., 2013). Here $\{\boldsymbol{\zeta}_\mu, \boldsymbol{\Delta}_\mu, \boldsymbol{\zeta}_\gamma, \boldsymbol{\Delta}_\gamma, \boldsymbol{\zeta}_\tau, \sigma_\tau^2, b_{\sigma^2}, c_{\sigma^2}, r_0, b_0, c_0, \nu, \{A_k\}_{k=1}^K\}$ is a set of hyper-parameters and $\boldsymbol{\Theta} = \{\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}, \mathbf{a}, \boldsymbol{\Sigma}, \sigma^2, \boldsymbol{\omega}, r, h, \tau\}$ is a set of latent variables. The joint distribution of latent and observed variables is:

$$\begin{aligned} P(\mathbf{y}, \boldsymbol{\Theta}) &= P(\mathbf{Z}|r, \boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\phi})P(\boldsymbol{\phi}|\sigma^2, \tau) \left(\prod_{i=1}^N P(\boldsymbol{\beta}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right) P(r|r_0, h) \dots \\ &\dots P(h|b_0, c_0) \left(\prod_{i=1}^N P(\omega_i|r) \right) P(\boldsymbol{\gamma}|\boldsymbol{\zeta}_\gamma, \boldsymbol{\Delta}_\gamma) P(\sigma^{-2}|b_{\sigma^2}, c_{\sigma^2}) \dots \\ &\dots P(\tau|\boldsymbol{\zeta}_\tau, \sigma_\tau^2) P(\boldsymbol{\mu}|\boldsymbol{\zeta}_\mu, \boldsymbol{\Delta}_\mu) \left(\prod_{k=1}^K P(a_k|s, \eta_k) \right) P(\boldsymbol{\Sigma}|\rho, \mathbf{B}). \end{aligned} \quad (13)$$

Finally, to obtain conjugate posterior updates of the dispersion parameter r , we use a compound Poisson representation of negative binomial distribution (see Appendix A).

3.3. Markov chain Monte Carlo estimation

MCMC estimation approximates a posterior distribution of interest through simulation of a Markov chain. In the present application, a Markov chain can be constructed by iteratively sampling from the conditional distributions of the parameters collected in $\boldsymbol{\Theta}$. As a results of Pólya-Gamma data augmentation, the conditional distributions of all model parameters, with the exception of the conditional distribution of the spatial association parameter τ , are conjugate to their prior and belong to known families of standard parametric distribution. Since the conditional distribution of τ does not correspond to any recognisable distribution, we adopt the random-walk Metropolis algorithm to generate samples of it. The resulting Gibbs sampler is presented in Algorithm 1. In the algorithm, ϖ_τ is the step size of the random-walk Metropolis algorithm, which needs to be tuned.

Initialization:

Set hyper-parameters: $\{\zeta_\mu, \Delta_\mu, \zeta_\gamma, \Delta_\gamma, \zeta_\tau, \sigma_\tau^2, b_{\sigma^2}, c_{\sigma^2}, r_0, b_0, c_0, \nu, \{A_k\}_{k=1}^K\}$;

Initialize latent variables: $\{\phi, \gamma, \beta, \mu, \mathbf{a}, \Sigma, \sigma^2, \omega, r, h, \tau\}$;

for 1 to max-iteration sample from

- $\phi|- \sim \text{Normal}\left((\Omega + \tilde{\Omega})^{-1}\Omega(Z - M\gamma - X\beta), (\Omega + \tilde{\Omega})^{-1}\right)$;
- $\gamma|- \sim \text{Normal}\left((\Delta_\gamma^{-1} + M^\top\Omega M)^{-1}[M^\top\Omega(Z - X\beta - \phi) + \Delta_\gamma^{-1}\zeta_\gamma], (\Delta_\gamma^{-1} + M^\top\Omega M)^{-1}\right)$;
- $\{\beta_i|\}_{i=1}^N \sim \text{Normal}\left(\left([\omega_i X_i X_i^\top]^{-1} + \Sigma\right)[\omega_i(Z_i - M_i^\top\gamma - \phi_i)X_i + \Sigma^{-1}\mu], [\omega_i X_i X_i^\top]^{-1} + \Sigma\right)$;
- $\mu|- \sim \text{Normal}\left((N\Sigma^{-1} + \Delta_\mu^{-1})^{-1}\left(\Sigma^{-1}\sum_{i=1}^N\beta_i + \Delta_\mu^{-1}\zeta_\mu\right), (N\Sigma^{-1} + \Delta_\mu^{-1})^{-1}\right)$;
- $\{a_k|\}_{k=1}^K \sim \text{Gamma}\left(\frac{\nu+K}{2}, \frac{1}{A_k^2} + \nu(\Sigma^{-1})_{kk}\right)$;
- $\Sigma|- \sim \text{IW}\left(\nu + N + K - 1, \mathbf{B} + \sum_{i=1}^N[\beta_i - \mu][\beta_i - \mu]^\top\right)$;
- $\sigma^{-2}|- \sim \text{Gamma}\left(b_{\sigma^2} + \frac{N}{2}, c_{\sigma^2} + \frac{\phi^\top S^{-1} S \phi}{2}\right)$;
- $\{\omega_i|\}_{i=1}^N \sim \text{PG}(y_i + r, \psi_i)$;
- $r|- \sim \text{Gamma}\left(r_0 + \sum_{i=1}^N L_i, h + \sum_{i=1}^N \ln(1 + \exp(\psi_i))\right)$ (see details in Appendix A) ;
- $h|- \sim \text{Gamma}(r_0 + b_0, r + c_0)$;
- $\tau|-$ (random-walk Metropolis step)
 - Propose $\tilde{\tau} = \tau + \sqrt{\omega_\tau}\sigma_\tau\zeta$, where $\zeta \sim \text{Normal}(0, 1)$;
 - Compute $\xi = \frac{P(\tilde{\tau}|\zeta_\tau, \sigma_\tau^2)P(\phi|\tilde{\tau}, \sigma^2)}{P(\tau|\zeta_\tau, \sigma_\tau^2)P(\phi|\tau, \sigma^2)}$;
 - Draw $u \sim \text{Uniform}(0, 1)$. If $\xi \leq u$, accept the proposal, else reject it.

end

Algorithm 1: Gibbs sampler for posterior inference in the spatial negative binomial model

3.4. Variational Bayes estimation

In this section, we propose a variational Bayesian (VB) method to estimate the spatial negative binomial regression model. The goal of VB is to find a variational distribution $q(\Theta)$, which approximates the posterior distribution of interest, via minimisation of the probability distance between the variational distribution and the actual posterior distribution (Jordan et al., 1999; Blei et al., 2017). The probability distance is conveniently measured by Kullback-Leibler (KL) divergence, which is defined as follows:

$$\begin{aligned} \text{KL}(q(\Theta)||P(\Theta|y)) &= \int \ln\left(\frac{q(\Theta)}{P(\Theta|y)}\right)q(\Theta)d\Theta \\ &= \mathbb{E}_q[\ln q(\Theta)] - \mathbb{E}_q[\ln P(\Theta|y)] \\ &= \mathbb{E}_q[\ln q(\Theta)] - \mathbb{E}_q[\ln P(\Theta, y)] + \ln P(y). \end{aligned} \tag{14}$$

VB aims to minimise the KL divergence, which implies that

$$q^*(\Theta) = \arg \min_q \text{KL}(q(\Theta)||P(\Theta|y)). \tag{15}$$

However, since $\ln P(y)$ has no closed form expression, the KL divergence is not analytically tractable. Recognising that $\mathbb{E}_q[\ln q(\Theta)] - \mathbb{E}_q[\ln P(\Theta, y)]$ is negative of the evidence lower bound (ELBO), we rearrange Equation 14 as follows:

$$\text{ELBO} = \ln P(y) - \text{KL}(q(\Theta)||P(\Theta|y)). \tag{16}$$

Since the KL divergence is always positive, equation 16 shows that the optimal variational distribution can be equivalently obtained by maximising the ELBO.

The variational distribution must be selected by the analyst. Its specification determines both the quality of the posterior approximation as well as the complexity of the optimisation problem (Blei et al., 2017). In the following subsections, we describe two approaches for the specification of the variational distribution and suitable methods for ELBO maximisation.

3.4.1. Mean field variational Bayes (MFVB)

MFVB specifies the density of the variational distribution as a product of the component-specific variational densities:

$$q(\Theta) = \prod_{j=1}^J q(\Theta_j), \quad (17)$$

where $j \in \{1, \dots, J\}$ are indexes of model parameter blocks. This specification imposes posterior independence between blocks of model parameters. The optimal variational density of a latent factor can be obtained using the following expression (Ormerod and Wand, 2010):

$$q^*(\Theta_j) \propto \exp\left(\mathbb{E}_{-\Theta_j} [\ln P(\mathbf{y}, \Theta)]\right). \quad (18)$$

If the conditional conjugacy holds for a model parameter, its variational distribution belongs to a recognisable family and can be easily obtained using the above equation. In case of non-conjugacy, the optimal variational density $q^*(\Theta_j)$ of a model parameters can be obtained using quasi-Newton methods, non-conjugate variational message passing (Knowles and Minka, 2011), stochastic linear regression (Salimans et al., 2013), or Laplace approximation (see Wang and Blei, 2013, for a comprehensive review).

In the Pólya-Gamma-augmented spatial NB model, the conditional conjugacy holds for all model parameters, except for τ . We thus obtain the optimal variational density of τ using non-conjugate variational message passing, while the optimal variational density of the remaining model parameters are obtained using equation 18. The results of MFVB indicate that the variational distributions of all variables, except τ and σ^2 , closely resemble the posterior estimates of MCMC. This observation is well aligned with the findings of Wu (2018) in linear spatial models. However, in accordance with Wu (2018), we also find that τ and σ^2 are poorly recovered by MFVB because of the untenable assumption of posterior independence.

3.4.2. Integrated non-factorised variational Bayes (INFVB)

To address the bottlenecks of MFVB in the estimation of the considered spatial NB model, we propose INFVB method (Han et al., 2013; Wu, 2018). INFVB decomposes latent variables Θ into two disjoint subsets $\{\Theta_c, \Theta_d\}$ to specify a flexible variational distribution:

$$q_{\text{INFVB}}(\Theta) = q(\Theta_c | \Theta_d) q(\Theta_d). \quad (19)$$

Since direct maximization of ELBO to find optimal variational density $q_{\text{INFVB}}^*(\Theta)$ is computationally challenging, a discrete distribution is specified on Θ_d by discretising its domain using a multi-dimensional grid. We adopt a two-step procedure to obtain the optimal variational density $q_{\text{INFVB}}^*(\Theta)$:

1. For each grid point $\boldsymbol{\Theta}_d^{(g)} \in \{\boldsymbol{\Theta}_d^{(1)}, \dots, \boldsymbol{\Theta}_d^{(G)}\}$, we obtain $q^*(\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)})$ and $q^*(\boldsymbol{\Theta}_d^{(g)})$ (up to a multiplicative constant) using equations 20 and 21, respectively (Han et al., 2013):

$$q^*(\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)}) = \operatorname{arg\,min}_{q(\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)})} \mathbb{E}_q \left[\ln q(\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)}) \right] - \mathbb{E}_q \left[\ln P(\mathbf{y}, \boldsymbol{\Theta}_c^{(g)}, \boldsymbol{\Theta}_d^{(g)}) \right], \quad (20)$$

$$q^*(\boldsymbol{\Theta}_d^{(g)}) \propto \exp \left(\mathbb{E} \left[\ln P(\mathbf{y}, \boldsymbol{\Theta}_c^{(g)}, \boldsymbol{\Theta}_d^{(g)}) \right] - \mathbb{E} \left[\ln q^*(\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)}) \right] \right). \quad (21)$$

2. We then compute optimal variational densities of $\boldsymbol{\Theta}_d$ and $\boldsymbol{\Theta}_c$ using equation 22:

$$\begin{aligned} q^*(\boldsymbol{\Theta}_d) &= \sum_{g=1}^G q^*(\boldsymbol{\Theta}_d^{(g)}) \mathbb{1}(\boldsymbol{\Theta}_d = \boldsymbol{\Theta}_d^{(g)}), \\ q^*(\boldsymbol{\Theta}_c) &= \sum_{g=1}^G q^*(\boldsymbol{\Theta}_d^{(g)}) q^*(\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)}), \end{aligned} \quad (22)$$

$$\text{where } q^*(\boldsymbol{\Theta}_d^{(g)}) = \frac{\exp \left(\mathbb{E} \left[\ln P(\mathbf{y}, \boldsymbol{\Theta}_c^{(g)}, \boldsymbol{\Theta}_d^{(g)}) \right] - \mathbb{E} \left[\ln q^*(\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)}) \right] \right)}{\sum_{e=1}^G \exp \left(\mathbb{E} \left[\ln P(\mathbf{y}, \boldsymbol{\Theta}_c^{(e)}, \boldsymbol{\Theta}_d^{(e)}) \right] - \mathbb{E} \left[\ln q^*(\boldsymbol{\Theta}_c^{(e)}|\boldsymbol{\Theta}_d^{(e)}) \right] \right)}.$$

We highlight three important features of INFVB. First, the optimal density update of $\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)}$ using equation 20 results into similar updates as obtained in MFVB (see equation 18). As a consequence, computation of $q^*(\boldsymbol{\Theta}_c^{(g)}|\boldsymbol{\Theta}_d^{(g)})$ is straightforward if conditional conjugacy holds for $\boldsymbol{\Theta}_c$. Second, the first step of INFVB includes embarrassingly parallel tasks. The communications overhead of these tasks is negligible, because the results of each task are only combined once during estimation. These characteristics make INFVB computationally efficient and scalable for large datasets. Third, if we consider $\boldsymbol{\Theta}_d$ as a vector of hyper-parameters, INFVB can be viewed as a generalised version of INLA. Specifically, INFVB relaxes the INLA's strict assumption on the normality of the conditional distribution $q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d)$ (see section 2.3 of Han et al., 2013, for a detailed discussion on the superiority of INFVB over INLA).

3.4.3. INFVB for the spatial negative binomial model

On the basis of the findings of MFVB, we consider $\boldsymbol{\Theta}_d = \{\tau, \sigma^2\}$ and $\boldsymbol{\Theta}_c = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_d$. We specify a nonparametric distribution on $\boldsymbol{\Theta}_d$ by discretising its domain using a two-dimensional grid and consider the following product form representation of $q(\boldsymbol{\Theta}_c)$:

$$\begin{aligned} q(\boldsymbol{\Theta}_c) &= q(\phi|\lambda_\phi, \Lambda_\phi)q(\gamma|\lambda_\gamma, \Lambda_\gamma)q(\beta|\lambda_\beta, \Lambda_\beta)q(\mu|\lambda_\mu, \Lambda_\mu) \prod_{k=1}^K q(a_k|\tilde{b}_{a_k}, \tilde{c}_{a_k}) \dots \\ &\dots q(\Sigma|\tilde{\rho}, \tilde{B}) \prod_{i=1}^N q(\omega_i|\tilde{b}_{\omega_i}, \tilde{c}_{\omega_i})q(h|\tilde{b}_h, \tilde{c}_h)q(r) \prod_{i=1}^N q(L_i). \end{aligned} \quad (23)$$

We find that variational distributions of model parameters blocks in Θ_c belong to known families of distributions due to conjugacy:

$$\begin{aligned}
q(\phi) &\sim \text{Normal}(\lambda_\phi, \Lambda_\phi), & q(\gamma) &\sim \text{Normal}(\lambda_\gamma, \Lambda_\gamma), & \{q(\beta_i)\}_{i=1}^N &\sim \text{Normal}(\lambda_{\beta_i}, \Lambda_{\beta_i}), \\
q(\mu) &\sim \text{Normal}(\lambda_\mu, \Lambda_\mu), & \{q(a_k)\}_{k=1}^K &\sim \text{Gamma}(\tilde{b}_{a_k}, \tilde{c}_{a_k}), & q(\Sigma) &\sim \text{IW}(\tilde{\rho}, \tilde{\mathbf{B}}), \\
\{q(\omega_i)\}_{i=1}^N &\sim \text{PG}(\tilde{b}_{\omega_i}, \tilde{c}_{\omega_i}), & q(h) &\sim \text{Gamma}(\tilde{b}_h, \tilde{c}_h), & q(r) &\sim \text{Gamma}(\tilde{b}_r, \tilde{c}_r), \\
\{q(L_i)\}_{i=1}^N &= \sum_{j=0}^{y_i} R_{\tilde{r}}(y_i, j) \delta_j, & q(\psi) &\sim \text{Normal}(\lambda_\psi, \Lambda_\psi).
\end{aligned}$$

Set hyper-parameters: $\{\zeta_\mu, \Delta_\mu, \zeta_\gamma, \Delta_\gamma, \zeta_\tau, \sigma_\tau^2, b_{\sigma^2}, c_{\sigma^2}, r_0, b_0, c_0, \nu, \{A_k\}_{k=1}^K\}$;
Compute fixed variational parameters: $\tilde{b}_{a_k} = \frac{\nu+K}{2}$; $\tilde{\rho} = \nu + N + K - 1$; $\tilde{b}_h = r_0 + b_0$;
Specify a two-dimensional grid $\Theta_d^{(g)} \in \{\Theta_d^{(1)}, \dots, \Theta_d^{(G)}\}$ on the domain of $\Theta_d = \{\tau, \sigma^2\}$;

Step: 1

for g in 1 to G obtain $q^*(\Theta_c^{(g)}|\Theta_d^{(g)})$ and $q^*(\Theta_d^{(g)})$ in parallel

Initialize $\{\lambda_\phi^{(g)}, \Lambda_\phi^{(g)}, \lambda_\gamma^{(g)}, \Lambda_\gamma^{(g)}, \{\lambda_{\beta_i}^{(g)}, \Lambda_{\beta_i}^{(g)}\}_{i=1}^N, \lambda_\mu^{(g)}, \Lambda_\mu^{(g)}, \{\tilde{c}_{a_k}^{(g)}\}_{k=1}^K, \tilde{\mathbf{B}}^{(g)}, \tilde{c}_h^{(g)}, \tilde{b}_r^{(g)}, \tilde{c}_r^{(g)}\}$;

while not converged do

$$\begin{aligned}
\Lambda_\phi^{(g)} &= (\mathbb{E}[\Omega]^{(g)} + \tilde{\Omega}^{(g)})^{-1}; \\
\lambda_\phi^{(g)} &= \Lambda_\phi^{(g)} (\mathbb{E}[\{Z^*\}^{(g)}] - \mathbb{E}[\Omega]^{(g)} M \lambda_\gamma^{(g)} - \mathbb{E}[\Omega]^{(g)} X \lambda_\beta^{(g)}); \\
\Lambda_\gamma^{(g)} &= (\Delta_\gamma^{-1} + M^T \mathbb{E}[\Omega]^{(g)} M)^{-1}; \\
\lambda_\gamma^{(g)} &= \Lambda_\gamma^{(g)} (M^T (\mathbb{E}[\{Z^*\}^{(g)}] - \mathbb{E}[\Omega]^{(g)} X \lambda_\beta^{(g)} - \mathbb{E}[\Omega]^{(g)} \lambda_\phi^{(g)}) + \Delta_\gamma^{-1} \zeta_\gamma); \\
\{\Lambda_{\beta_i}^{(g)}\}_{i=1}^N &= (\mathbb{E}[\omega_i^{(g)}] X_i X_i^T + \tilde{\rho} \{\tilde{\mathbf{B}}^{(g)}\}^{-1})^{-1}; \\
\{\lambda_{\beta_i}^{(g)}\}_{i=1}^N &= \Lambda_{\beta_i}^{(g)} [(\mathbb{E}[\{Z_i^*\}^{(g)}] - \mathbb{E}[\omega_i^{(g)}] M_i^T \lambda_\gamma^{(g)} - \mathbb{E}[\omega_i^{(g)}] \lambda_{\phi_i}^{(g)}) X_i + \tilde{\rho} \{\tilde{\mathbf{B}}^{-1} \lambda_\mu\}^{(g)}]; \\
\Lambda_\mu^{(g)} &= [N \tilde{\rho} \{\tilde{\mathbf{B}}^{(g)}\}^{-1} + \Delta_\mu^{-1}]^{-1}; \\
\lambda_\mu^{(g)} &= \Lambda_\mu^{(g)} [(\tilde{\rho} \{\tilde{\mathbf{B}}^{(g)}\}^{-1}) \sum_{i=1}^N \lambda_{\beta_i}^{(g)} + \Delta_\mu^{-1} \zeta_\mu]; \\
\{\tilde{c}_{a_k}^{(g)}\}_{k=1}^K &= \left[\frac{1}{A_k^2} + \nu \tilde{\rho} (\{\tilde{\mathbf{B}}^{(g)}\}^{-1})_{kk} \right]; \\
\tilde{\mathbf{B}}^{(g)} &= 2\nu \text{diag}\left(\frac{\tilde{b}_a}{\tilde{c}_a^{(g)}}\right) + N \Lambda_\mu^{(g)} + \sum_{i=1}^N (\Lambda_{\beta_i} + [\lambda_{\beta_i} - \lambda_\mu][\lambda_{\beta_i} - \lambda_\mu]^T)^{(g)}; \\
\tilde{c}_h^{(g)} &= \left(\frac{\tilde{b}_r}{\tilde{c}_r}\right)^{(g)} + c_0; \\
\tilde{b}_r^{(g)} &= r_0 + \sum_{i=1}^N \mathbb{E}(L_i^{(g)}); \\
\tilde{c}_r^{(g)} &= \frac{\tilde{b}_h}{\tilde{c}_h^{(g)}} + \sum_{i=1}^N \mathbb{E}[\log(1 + \exp(\psi_i^{(g)}))]; \\
\lambda_\psi^{(g)} &= M \lambda_\gamma^{(g)} + X \lambda_\beta^{(g)} + \lambda_\phi^{(g)}; \\
\Lambda_\psi^{(g)} &= M \Lambda_\gamma^{(g)} M^T + X \Lambda_\beta^{(g)} X^T + \Lambda_\phi^{(g)};
\end{aligned}$$

end

Compute $q^*(\Theta_d^{(g)})$ up to a multiplicative constant by inserting expectations computed using equation 31 (see appendix B.2) into equation 21;

end

Step: 2

Obtain optimal variational densities of Θ_d and Θ_c using equation 22;

Algorithm 2: Integrated non-factorized variational Bayes (INFVB) method for the spatial NB model

We reiterate that a compound Poisson representation of negative binomial distribution is used to ensure conjugate posterior updates for the dispersion parameter r (see Appendix A for details).

Accordingly, we adopt the variational distribution used by Zhou et al. (2012) on L_i , where δ_j is an indicator. The INFVB method to estimate the spatial count model is summarised in Algorithm 2; supplementary identities and expressions are presented in Appendix B.1. The expression for the conditional ELBO, i.e. the negative of the function minimised in equation 20 is presented in Appendix B.2.

4. Simulation study

To evaluate computational efficiency and finite sample properties of INFVB and MCMC estimators, we conduct a Monte Carlo study. In this section, we present details of the data generating process (DGP), followed by performance measures, implementation details and results of the simulation study.

4.1. Data and experimental setup

We generate data according to the following DGP:

$$\begin{aligned}
\beta_i &\sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), & i = 1, \dots, N \\
\epsilon &\sim \text{Normal}(0, \sigma^2 \mathbf{I}_N), \\
\mathbf{S}\boldsymbol{\phi} &= \exp(\boldsymbol{\tau}\mathbf{W})\boldsymbol{\phi} = \epsilon, \\
\psi_i &= \mathbf{M}_i^\top \boldsymbol{\gamma} + \mathbf{X}_i^\top \boldsymbol{\beta}_i + \phi_i, & i = 1, \dots, N \\
p_i &= \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}, & i = 1, \dots, N \\
y_i &\sim \text{NB}(r, p_i). & i = 1, \dots, N
\end{aligned}$$

We consider eight simulation scenarios defined through combinations of $N = \{1000, 1500\}$, $\tau = \{-0.7, 0.7\}$, and $\sigma = \{0.2, 0.4\}$. Ten resamples of each simulation scenario are generated, i.e. we estimate the spatial NB model using MCMC and INFVB on a total of 80 simulated datasets. For all simulation scenarios, we set $\boldsymbol{\mu} = [0.2 \quad -0.2 \quad 0.2]^\top$, $\boldsymbol{\Sigma} = \text{diag}(\tilde{\boldsymbol{\sigma}})\tilde{\boldsymbol{\Omega}}\text{diag}(\tilde{\boldsymbol{\sigma}})$ with $\tilde{\boldsymbol{\sigma}} = [0.141 \quad 0.141 \quad 0.141]^\top$ and $\tilde{\boldsymbol{\Omega}} = \begin{bmatrix} 1 & 0.2 & 0 \\ 0.2 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{bmatrix}^\top$ as well as $\boldsymbol{\gamma} = [1.0 \quad 0.3 \quad -0.3 \quad 0.3]^\top$, and $r = 1.5$. Furthermore, we let $M_{i,1} = 1$ and $M_{i,q} \sim \text{Normal}(0, 1)$ for $q = 2, 3, 4$ as well as $X_{i,k} \sim \text{Normal}(0, 1)$ for $k = 1, 2, 3$. To construct the row-normalised spatial weights matrix \mathbf{W} , we calculate an 8-nearest neighbour matrix for N points, which are randomly located in a unit square.

4.2. Performance metrics

We evaluate the estimation accuracy of the INFVB and MCMC methods by calculating the mean of the absolute percent bias (APB) of model parameters across resamples. APB is a normalised measure of the finite sample bias and is given by $\text{APB} = \left| \frac{\text{MPM} - \text{True value}}{\text{True value}} \right| \times 100$, where the mean posterior mean (MPM) is the average of the posterior mean across resamples. In addition, we also report the standard deviation of the posterior mean (SDPM) and the mean of posterior standard deviation (MPSD) across resamples.

4.3. Implementation and estimation practicalities

We implement the MCMC and INFVB methods for the spatial NB model by writing our own Python code. To draw from the Pólya-Gamma distribution, we use an existing implementation (Linderman et al., 2015, 2016a,b) of the sampling techniques proposed by Polson et al. (2013) and Windle et al. (2014).¹

The MCMC sampler is executed with two parallel Markov chains and 40,000 iterations for each chain, whereby the initial 20,000 iterations are discarded for burn-in. After burn-in, every fifth draw is retained. The random-walk Metropolis step to generate samples from the conditional distribution of the spatial association parameter τ is adaptively scaled such that the average acceptance rate is approximately 44%, which is the recommended acceptance ratio for a uni-dimensional target density (see Roberts et al., 1997). Convergence of the MCMC simulation is assessed with the help of the potential scale reduction factor (Gelman et al., 1992).

For INFVB, a two-dimensional search space over $\{\tau, \sigma\}$ is defined via the Cartesian product of two uni-dimensional grids. The grid over τ consists of 15 equidistant points in the interval $[0, 1.4]$ or $[-1.4, 0]$ (depending on the true value of τ), while the grid over σ consists of 10 equidistant points in the interval $[0.05, 0.8]$. We exploit the embarrassingly parallel computations of the INFVB method by distributing step 1 of Algorithm 2 over an eight-core processor.

4.4. Results

Before comparing INFVB with MCMC, we demonstrate the accuracy of our analytical derivation and implementation of the INFVB method. In one resample of one specific simulation scenario, we plot the evolution of the conditional ELBO (presented in Appendix B.2) over the number of iterations for ten randomly selected grid points in Figure 1. It can be seen that the conditional ELBOs of the ten randomly grid points are monotonically increasing over iterations, which illustrates the correctness of the proposed INFVB estimator.

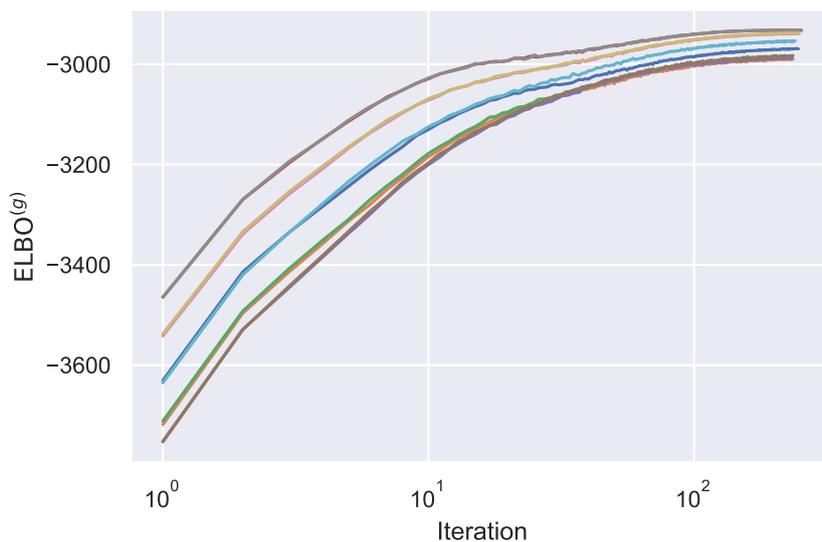


Figure 1.: **Sequence of conditional ELBOs of ten randomly selected grid points for simulation scenario $\tau = -0.7$, $\sigma = 0.2$, $N = 1500$**

¹The estimation code is publicly available at https://github.com/RicoKrueger/infvb_spatial_count.

Table 1 enumerates the computation times of the MCMC and INFVB estimators for all DGP instances. INFVB is approximately 50 times faster than MCMC for all instances of the DGP. Considerably low standard deviations of the estimation time across resamples underscore the robustness of this result. Further reductions in the estimation time of INFVB could be realised by distributing step 1 of Algorithm 2 over more than eight compute cores.

Next, we present the results of the other performance measures for four out of the eight simulation scenarios in Tables 2 to 5.² Similar and considerably low APB values (below 10% for most of the parameters), and small SDPM values indicate that INFVB and MCMC not only recover the true parameters quite well but also with an identical precision across all the considered simulation scenarios. As an exception, the recovery of σ is poor in INFVB and a similar bias is observed for τ in MCMC. However, both τ and σ are recovered equally well by MCMC and INFVB in the empirical study (see Figure 4 in the next section). Furthermore, for most model parameters, MPSD is substantially lower for INFVB than for MCMC. This result corroborates the findings of earlier studies, which suggest that VB underestimates the posterior uncertainty (Blei et al., 2017; Giordano et al., 2018).

	INFVB		MCMC	
	Mean	Std. dev.	Mean	Std. dev.
<i>N</i> = 1000				
$\tau = -0.7; \sigma = 0.2$	9.1	0.2	494.0	17.6
$\tau = 0.7; \sigma = 0.2$	9.2	0.2	512.8	1.2
$\tau = -0.7; \sigma = 0.4$	9.4	0.1	525.3	4.1
$\tau = 0.7; \sigma = 0.4$	9.3	0.1	506.7	1.6
<i>N</i> = 1500				
$\tau = -0.7; \sigma = 0.2$	28.2	0.3	1397.1	14.9
$\tau = 0.7; \sigma = 0.2$	28.2	0.4	1423.2	6.4
$\tau = -0.7; \sigma = 0.4$	29.0	0.4	1491.8	10.4
$\tau = 0.7; \sigma = 0.4$	21.5	1.1	1343.3	8.5

Table 1.: **Estimation time in minutes across ten resamples by estimation method and simulation scenario**

²The results for the remaining for simulation scenarios with $N = 1000$ offer similar insights and are thus included as supplementary material.

	True	INFVB				MCMC			
		MPM	SDPM	APB	MPSD	MPM	SDPM	APB	MPSD
γ_1	1.000	1.005	0.038	0.5	0.030	1.021	0.037	2.1	0.043
γ_2	0.300	0.285	0.040	4.9	0.030	0.291	0.037	2.9	0.043
γ_3	-0.300	-0.294	0.028	1.9	0.031	-0.298	0.031	0.8	0.043
γ_4	0.300	0.301	0.038	0.4	0.030	0.308	0.043	2.7	0.043
μ_1	0.200	0.197	0.021	1.3	0.003	0.202	0.022	1.2	0.026
μ_2	-0.200	-0.205	0.037	2.3	0.003	-0.208	0.036	3.9	0.026
μ_3	0.200	0.199	0.034	0.4	0.003	0.205	0.037	2.6	0.026
$\tilde{\sigma}_1$	0.141	0.123	0.017	13.0	0.004	0.146	0.064	3.2	0.057
$\tilde{\sigma}_2$	0.141	0.120	0.013	15.4	0.004	0.135	0.053	4.3	0.065
$\tilde{\sigma}_3$	0.141	0.116	0.011	18.1	0.004	0.111	0.044	21.7	0.061
τ	-0.700	-0.604	0.110	13.7	0.390	-0.159	0.145	77.3	0.435
σ	0.200	0.119	0.020	40.3	0.046	0.152	0.069	23.8	0.071
r	1.500	1.514	0.053	0.9	0.040	1.477	0.057	1.5	0.083

Note: MPM = mean of posterior mean; SDPM = standard deviation of posterior mean; APB = absolute percent bias; MPSD = mean of posterior standard deviation. All statistics are calculated across ten resamples.

Table 2.: **Simulation results for $\tau = -0.7$, $\sigma = 0.2$, $N = 1500$**

	True	INFVB				MCMC			
		MPM	SDPM	APB	MPSD	MPM	SDPM	APB	MPSD
γ_1	1.000	0.986	0.026	1.4	0.030	1.003	0.029	0.3	0.044
γ_2	0.300	0.297	0.046	0.9	0.030	0.304	0.048	1.3	0.043
γ_3	-0.300	-0.282	0.031	6.0	0.030	-0.287	0.028	4.4	0.043
γ_4	0.300	0.279	0.033	7.1	0.030	0.283	0.037	5.7	0.043
μ_1	0.200	0.184	0.023	7.8	0.003	0.192	0.023	3.8	0.027
μ_2	-0.200	-0.198	0.030	0.8	0.003	-0.202	0.030	1.2	0.027
μ_3	0.200	0.202	0.027	0.8	0.003	0.204	0.030	2.2	0.027
$\tilde{\sigma}_1$	0.141	0.122	0.013	13.8	0.004	0.134	0.056	5.6	0.065
$\tilde{\sigma}_2$	0.141	0.129	0.016	8.7	0.004	0.150	0.057	6.2	0.070
$\tilde{\sigma}_3$	0.141	0.118	0.014	16.5	0.004	0.132	0.042	6.9	0.058
τ	0.700	0.633	0.041	9.6	0.421	-0.045	0.150	106.5	0.435
σ	0.200	0.116	0.017	41.8	0.046	0.153	0.052	23.4	0.084
r	1.500	1.531	0.061	2.0	0.039	1.497	0.061	0.2	0.089

Note: For an explanation of the table headers see Table 2.

Table 3.: **Simulation results for $\tau = 0.7$, $\sigma = 0.2$, $N = 1500$**

	True	INFVB				MCMC			
		MPM	SDPM	APB	MPSD	MPM	SDPM	APB	MPSD
γ_1	1.000	0.979	0.032	2.1	0.031	0.981	0.032	1.9	0.048
γ_2	0.300	0.317	0.049	5.6	0.031	0.304	0.050	1.3	0.047
γ_3	-0.300	-0.275	0.033	8.3	0.032	-0.299	0.036	0.4	0.047
γ_4	0.300	0.299	0.041	0.2	0.031	0.290	0.043	3.5	0.047
μ_1	0.200	0.199	0.026	0.7	0.004	0.206	0.028	2.9	0.029
μ_2	-0.200	-0.190	0.036	5.0	0.004	-0.196	0.036	1.9	0.028
μ_3	0.200	0.203	0.030	1.5	0.004	0.208	0.032	4.0	0.029
$\tilde{\sigma}_1$	0.141	0.127	0.017	10.5	0.008	0.152	0.072	7.4	0.066
$\tilde{\sigma}_2$	0.141	0.126	0.016	10.9	0.008	0.135	0.044	4.4	0.071
$\tilde{\sigma}_3$	0.141	0.126	0.018	10.9	0.007	0.152	0.054	7.5	0.070
τ	-0.700	-1.025	0.203	46.4	0.293	-0.635	0.194	9.2	0.250
σ	0.400	0.184	0.033	54.0	0.048	0.359	0.075	10.3	0.073
r	1.500	1.480	0.114	1.3	0.056	1.519	0.118	1.3	0.101

Note: For an explanation of the table headers see Table 2.

Table 4.: **Simulation results for $\tau = -0.7$, $\sigma = 0.4$, $N = 1500$**

	True	INFVB				MCMC			
		MPM	SDPM	APB	MPSD	MPM	SDPM	APB	MPSD
γ_1	1.000	1.024	0.061	2.4	0.032	1.015	0.060	1.5	0.048
γ_2	0.300	0.302	0.058	0.8	0.032	0.280	0.053	6.6	0.048
γ_3	-0.300	-0.254	0.045	15.4	0.031	-0.283	0.053	5.7	0.048
γ_4	0.300	0.313	0.026	4.3	0.032	0.292	0.031	2.7	0.048
μ_1	0.200	0.193	0.028	3.3	0.004	0.203	0.037	1.7	0.031
μ_2	-0.200	-0.189	0.026	5.6	0.003	-0.193	0.028	3.5	0.029
μ_3	0.200	0.205	0.027	2.4	0.003	0.211	0.030	5.6	0.028
$\tilde{\sigma}_1$	0.141	0.133	0.020	5.8	0.008	0.160	0.066	13.3	0.073
$\tilde{\sigma}_2$	0.141	0.128	0.019	9.3	0.007	0.134	0.052	5.0	0.065
$\tilde{\sigma}_3$	0.141	0.123	0.016	13.1	0.007	0.128	0.051	9.5	0.068
τ	0.700	0.717	0.079	2.4	0.419	0.295	0.166	57.8	0.325
σ	0.400	0.163	0.021	59.3	0.056	0.366	0.086	8.5	0.087
r	1.500	1.404	0.089	6.4	0.052	1.482	0.101	1.2	0.102

Note: For an explanation of the table headers see Table 2.

Table 5.: **Simulation results for $\tau = 0.7$, $\sigma = 0.4$, $N = 1500$**

5. Case study

In this section, we compare the performance of INFVB and MCMC in terms of computational efficiency, goodness-of-fit, and marginal posterior distributions of model parameters in an empirical application.

5.1. Data

The data consist of youth pedestrian injury counts in 603 census tracts of the New York City boroughs Bronx and Manhattan in the period from 2005 to 2014. The considered injury data were originally compiled by [Morris et al. \(2019\)](#) and contain census tract level information about reported youth pedestrian injury counts (aggregated across different levels of injury severity), social fragmentation, traffic volume and private vehicle commute mode shares. The youth pedestrian injury counts are informed by the number of 5- to 18-year-old pedestrian injured in traffic crashes. Social fragmentation is measured by a composite index which takes into account the number of vacant housing units, single-person households, non-owner occupied housing units, and the population having relocated within the past year. Traffic volume is measured in terms of the maximum annual average daily traffic in the census tract. For more information about the data compilation and the data sources, the reader is directed to [Morris et al. \(2019\)](#). We supplement the data collected by [Morris et al. \(2019\)](#) with information about the employment density (number of workers per km²), the proportion of households with poverty status and the proportion of the population that identifies as Black or African-American. The supplementary data were sourced from the 2012–2016 American Community Survey ([US Census Bureau, nd](#)). Summary statistics for the considered data are reported in Table 6. Figures 2 and 3 visualise the distribution of observed youth pedestrian injury counts across census tracts. A 5-nearest neighbour matrix for the study area is constructed using the PySAL library ([Rey and Anselin, 2010](#)) for Python.

Variable	Mean	Std.	Min.	Max.
Youth pedestrian injury count, 2005-14	9.69	8.35	0.00	44.00
Prop. of households with poverty status, 2012-16	0.24	0.15	0.00	0.57
Prop. of black or African-American alone population, 2012-16	0.24	0.22	0.00	0.91
No. of workers per km ² in 1000, 2012-16	17.96	37.34	0.02	260.40
Social fragmentation index	2.02	2.73	-4.50	18.67
Avg. annual daily traffic (AADT) in 10k, 2015	4.45	4.68	0.21	27.65
Private vehicle commute mode share, 2010-14	0.19	0.15	0.00	0.76

Table 6.: Description of youth pedestrian injury counts and explanatory variables by census tract (N = 603)

Youth pedestrian injury count, 2005-14

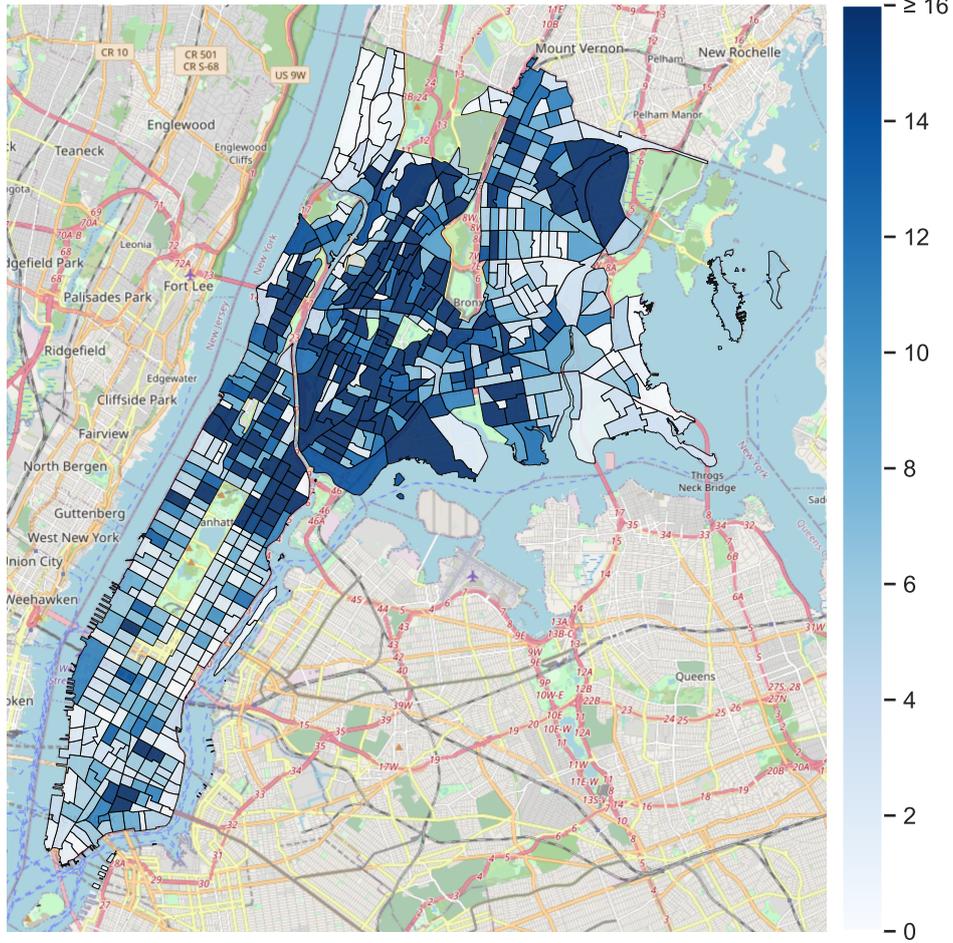


Figure 2.: Observed youth pedestrian injury counts in the Bronx and Manhattan in 2005-14 by census tract

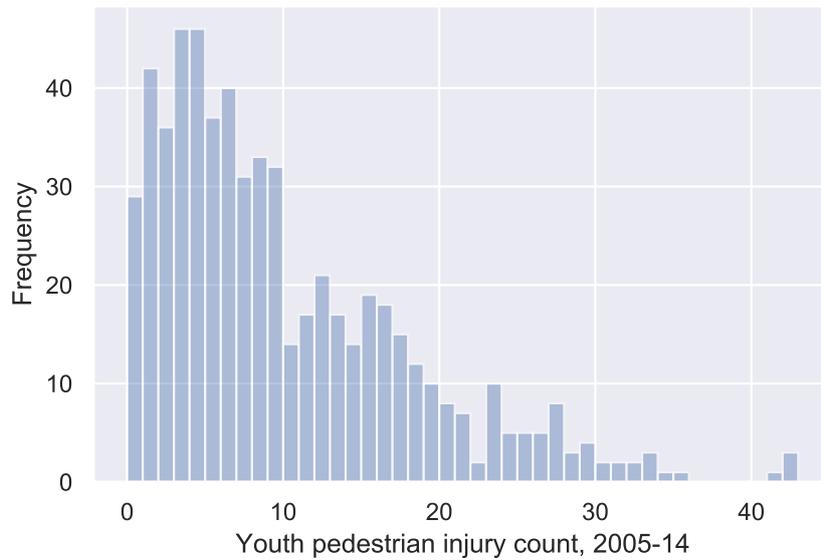


Figure 3.: Histogram of observed youth pedestrian injury counts in the Bronx and Manhattan in 2005-14 by census tract

5.2. Goodness of fit

We evaluate the estimation accuracy of the MCMC and INFVB estimators in terms of goodness of fit to the training data. To this end, we compute three proper scoring rules, namely the log-score, the Dawid-Sebastiani score and the ranked probability score. In principle, a scoring rule provides a measurement of the discrepancy between the observed outcome and the estimated predictive distribution. A scoring rule is said to be proper if the expected score is minimised by the true predictive distribution (Gneiting and Raftery, 2007; Wei and Held, 2014). The three considered scoring rules are defined and calculated as follows:

- The log-score (LS; Gneiting and Raftery, 2007; Wei and Held, 2014) corresponds to the negative pointwise log-likelihood:

$$\text{LS}(y_{\text{obs}}, \theta) = -\log f(y_{\text{obs}}|\theta). \quad (24)$$

For the NB model, the log-score is given by

$$\text{LS}(y_i, \psi_i, r) = -\ln \Gamma(y_i + r) + \ln \Gamma(r) + \ln \Gamma(y_i + 1) - y_i \psi_i + (y_i + r) \ln(1 + \exp(\psi_i)). \quad (25)$$

- The Dawid-Sebastiani score (DSS; Dawid and Sebastiani, 1999) is informed by the mean μ and the variance σ^2 of the predictive distribution:

$$\text{DSS}(y_{\text{obs}}, \mu, \sigma^2) = \frac{(y_{\text{obs}} - \mu)^2}{\sigma^2} + \log \sigma^2. \quad (26)$$

For the NB model, we have $\mu_i = \exp(\psi_i)r$ and $\sigma_i^2 = (\exp(\psi_i) + \exp(2\psi_i))r$.

- The ranked probability score (RPS; Matheson and Winkler, 1976) depends on the whole predictive distribution:

$$\text{RPS}(F, y_{\text{obs}}) = \sum_{t=0}^{\infty} (F(t) - \mathbb{1}\{y_{\text{obs}} \leq t\})^2, \quad (27)$$

where F denotes the predictive cumulative distribution function (CDF). $\mathbb{1}\{y_{\text{obs}} \leq t\}$ is an indicator which is one if the observed outcome y_{obs} is less than the threshold t and zero otherwise. Jordan et al. (2019) and Wei and Held (2014) provide expressions for the ranked probability score of the NB model:

$$\begin{aligned} \text{RPS}(F_{r,p_i}, y_i) = & y_i (2F_{r,p_i}(y_i) - 1) - \frac{rp_i}{(1-p_i)^2} \\ & \left((1-p_i)(2F_{r+1,p_i}(y_i-1) - 1) + {}_2\mathcal{F}_1\left(r+1, \frac{1}{2}; 2; -\frac{4p_i}{(1-p_i)^2}\right) \right). \end{aligned} \quad (28)$$

Here, $F_{r,p}(y) = \begin{cases} 1 - I_p(y+1, r), & y \geq 0 \\ 0 & y < 0 \end{cases}$ is the CDF of the NB distribution; $I_x(a, b)$ represents the regularised incomplete beta function; ${}_2\mathcal{F}_1(a, b; c; z)$ denotes the hypergeometric function.

For simplicity, the definitions presented above pertain to a single observation. In practice, aggregate scores are computed by summing over all observations in the data. In a Bayesian context, the posterior

distributions of the scores can be obtained by evaluating the scores at the posterior samples of the model parameters.

5.3. Results

For the case study, the same estimation practicalities as for the simulation study (see Section 4.3) apply with the only a minor difference that for INFVB, the grid over τ consists of 16 equidistant points in the interval $[-1.5, 0]$.

Our first finding is that INFVB is substantially faster than MCMC. While the estimation time of MCMC is 135.9 minutes, the estimation of INFVB is only 2.9 minutes. The computation time of INFVB can be further decreased by distributing step 1 of Algorithm 2 over more than eight computer cores. In theory, as many compute cores as there are grid points can be used and the estimation time of INFVB can be further decreased by a factor of 20. However, it is important to note that the MCMC simulation cannot be sped further due to the sequential and conditional nature of Gibbs sampling.

The goodness of fit results of the MCMC and INFVB estimators are compared in Table 7. For all scores, the posterior mean of INFVB is marginally smaller than the respective posterior mean of MCMC. For example, the posterior mean of the Dawid-Sebastiani score for MCMC is 2762.3, while it is 2720.2 for INFVB. For all scores, the credible intervals of MCMC are wider than those of INFVB. In fact, the credible intervals of the INFVB scores are fully contained within the MCMC credible intervals. In a nutshell, the posterior distributions of the scores indicate that MCMC and INFVB provide the same level of goodness of fit to the training data, while MCMC estimation carries greater uncertainty than INFVB estimation. Lower uncertainty in INFVB estimates is as expected and is consistent with the literature (Blei et al., 2017; Giordano et al., 2018).

Score	MCMC			INFVB		
	Mean	[2.5%;	97.5%]	Mean	[2.5%;	97.5%]
LS	1846.3	[1785.2;	1878.1]	1832.5	[1770.8;	1855.7]
DSS	2762.3	[2588.0;	2864.7]	2720.2	[2552.0;	2796.3]
RPS	2159.6	[1953.9;	2275.4]	2102.5	[1858.2;	2192.0]

Table 7.: Goodness of fit to youth pedestrian injury count data by estimation method

Figure 4 shows the marginal posterior approximations inferred by MCMC and INFVB of selected model parameters. By and large, the posterior approximations produced by the two methods exhibit a close correspondence. In particular, the posterior approximations of the fixed link function parameters, the mean and variance terms of the random link function parameters, the spatial error scale σ and the spatial association parameter τ coincide closely. For the the negative binomial shape parameter r , the posterior approximations of MCMC and INFVB overlap, but their modes differ.

Furthermore, we contrast the in-sample predictive accuracy of the MCMC and INFVB estimators by comparing the predicted injury counts for each census tract. Figure 5 shows histograms of the predicted injury counts for both MCMC and INFVB. It can be seen that the two distributions overlap closely with each other. In addition, Figure 6 visualises the difference between the youth pedestrian injury counts predicted by INFVB (\hat{y}^{INFVB}) and the corresponding MCMC prediction (\hat{y}^{MCMC}) for all census tracts. The differences in predicted youth pedestrian injury counts are generally small relative to the observed injury counts (see Figure 2).

Finally, Figure 7 shows histograms of the posterior means of the spatial errors $\{\phi_1, \dots, \phi_N\}$ for MCMC and INFVB. The figure suggests that MCMC and INFVB perform equally well at recovering the unobserved spatial dependence.

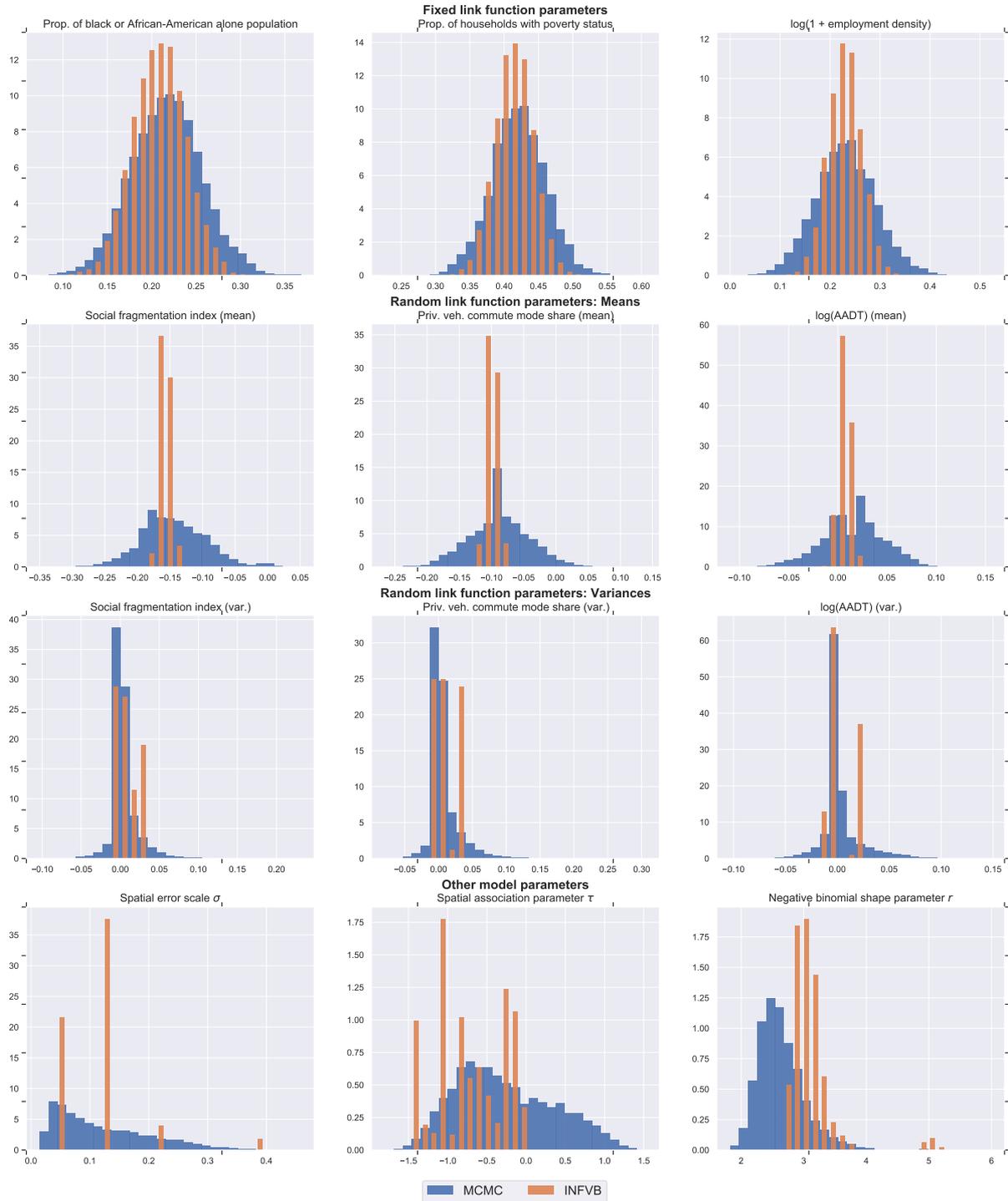


Figure 4.: Marginal posterior approximations of MCMC and INFVB for the youth pedestrian injury count data

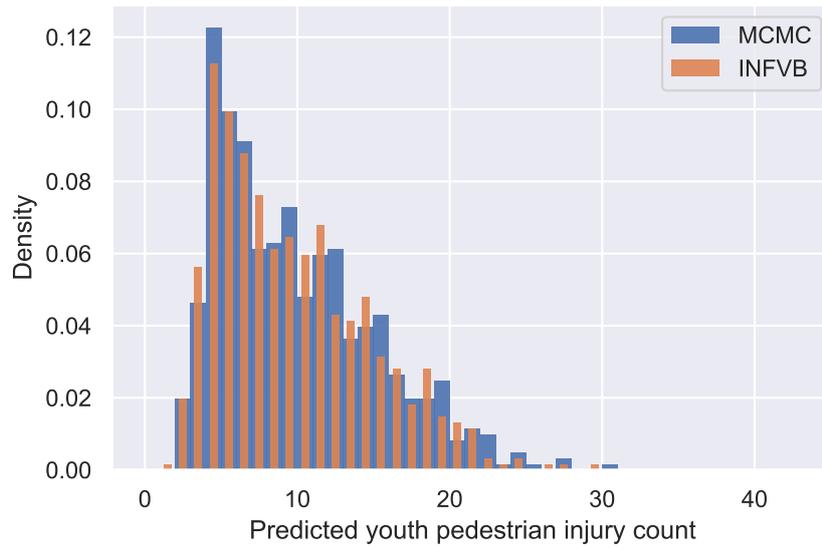


Figure 5.: Histogram of predicted youth pedestrian injury counts in the Bronx and Manhattan by census tract and estimation method

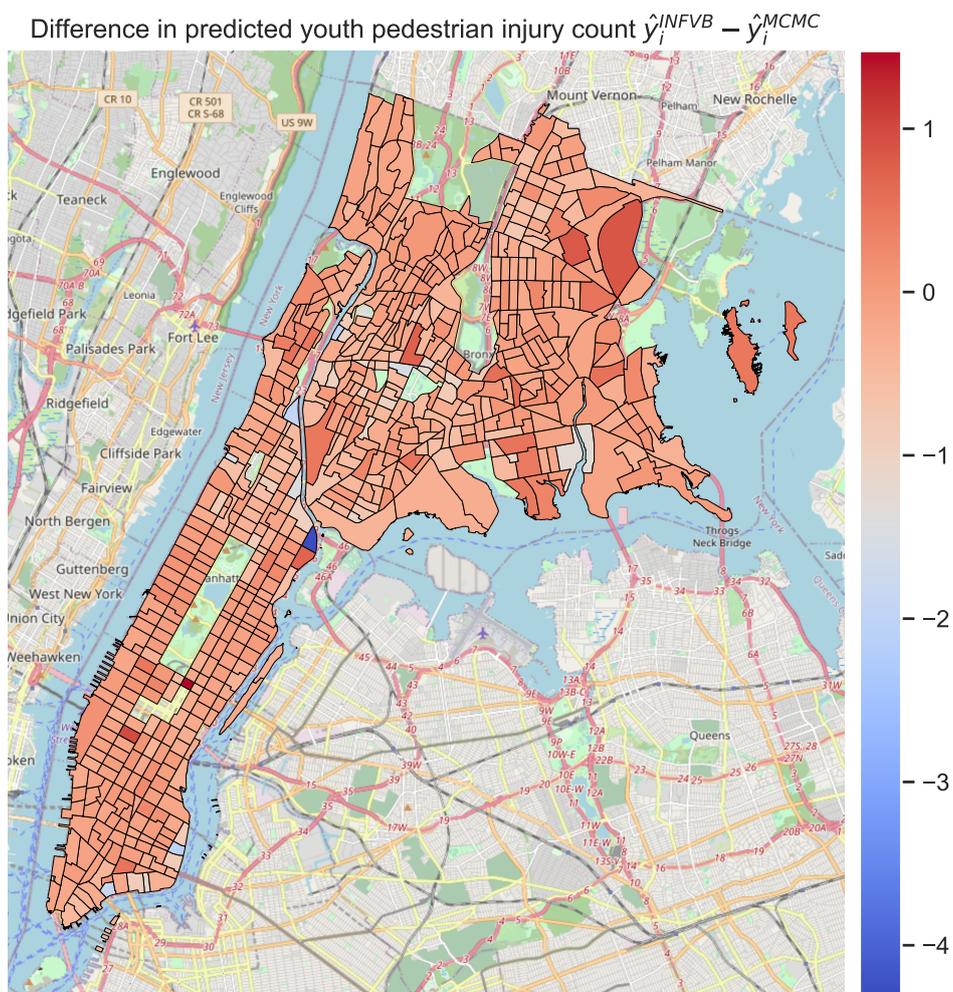


Figure 6.: Differences in youth pedestrian injury counts predicted by INFVB and MCMC in the Bronx and Manhattan by census tract

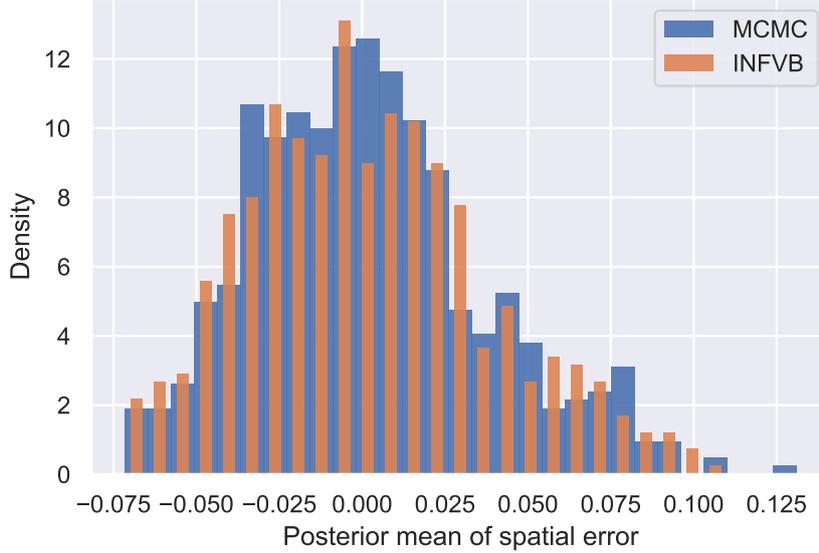


Figure 7.: **Histogram of posterior means of spatial errors $\{\phi_1, \dots, \phi_N\}$ by census tract and estimation method**

6. Conclusion

In this paper, we propose and empirically validate a variational Bayes (VB) method for posterior inference in a negative binomial model with unobserved spatial heterogeneity and dependence. The proposed VB method relies on Pólya-Gamma data augmentation to deal with the non-conjugacy of the negative binomial likelihood and an integrated non-factorised specification of the variational distribution to capture posterior dependencies. We benchmark the proposed VB method against MCMC using simulated data as well as real data on youth pedestrian injury counts in the census tracts of the New York City boroughs Bronx and Manhattan. In both applications, the VB approach is around 45 to 50 times faster than MCMC on a regular eight-core processor and emulates the estimation and predictive accuracy of MCMC. The marginal posterior approximations inferred by the VB approach and MCMC also resemble each other closely. The sequential and conditional nature of Gibbs sampling precludes improvement in computational efficiency through parallelisation. By contrast, INFVB can be further accelerated by a factor of up to 20 by taking full advantage of its embarrassingly parallel nature. Thus, INFVB is a scalable alternative to MCMC for the estimation of spatial count data models.

There are several ways in which future work can extend the research presented in the current paper. First, MCMC and VB should be compared on other data sets from other disciplines to collect additional evidence about the relative advantages of the two methods. A second directions for future work is to adapt the proposed VB approach to models with spatio-temporal dependencies. Finally, recent advances in stochastic optimisation could be leveraged to enable the application of the proposed VB method to online inference problems (Hoffman et al., 2013). Online estimation updates parameters continually, as new data points arrive, and thus facilitates the processing of very large data sets and data streams.

Acknowledgements

We would like to thank the associate editor and two anonymous reviewers for their critical assessment of our work. Furthermore, we are grateful to Michel Bierlaire for his helpful comments and suggestions.

Author contribution statement

PB: conception and design, method derivation, manuscript writing and editing. RK: conception and design, method implementation, data preparation and analysis, manuscript writing and editing. DJG: resources, manuscript editing.

References

- Abramowitz, M. and Stegun, I. A. (1948). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office.
- Acs, Z. J., Anselin, L., and Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge. *Research policy*, 31(7):1069–1085.
- Al-Mohy, A. H. and Higham, N. J. (2010). A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, 31(3):970–989.
- Anselin, L. (2013). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.
- Bansal, P., Krueger, R., Bierlaire, M., Daziano, R. A., and Rashidi, T. H. (2020). Bayesian estimation of mixed multinomial logit models: Advances and simulation-based evaluations. *Transportation Research Part B: Methodological*, 131:124–142.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.
- Bivand, R. S., Gómez-Rubio, V., and Rue, H. (2014). Approximate bayesian inference for spatial econometrics models. *Spatial Statistics*, 9:146–165.
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.
- Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335.
- Castro, M., Paleti, R., and Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation research part B: methodological*, 46(1):253–272.
- Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pages 65–81.
- Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global ecology and biogeography*, 16(2):129–138.
- Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, D. W., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.
- Durante, D., Rigon, T., et al. (2019). Conditionally conjugate mean-field variational bayes for logistic models. *Statistical Science*, 34(3):472–485.

- Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.
- Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness and variational bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.
- Glaser, S. (2017). A review of spatial econometric models for count data. Technical report, Hohenheim Discussion Papers in Business, Economics and Social Sciences.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.
- Haining, R. P. and Li, G. (2020). *Regression Modelling With Spatial and Spatial-Temporal Data: A Bayesian Approach*. CRC Press.
- Han, S., Liao, X., and Carin, L. (2013). Integrated non-factorized variational inference. In *Advances in Neural Information Processing Systems*, pages 2481–2489.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.
- Huang, A., Wand, M. P., et al. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452.
- Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoring rules. *Journal of Statistical Software*, 90(1):1–37.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.
- Kabisa, S., Dunson, D. B., and Morris, J. S. (2016). Online variational bayes inference for high-dimensional correlated data. *Journal of Computational and Graphical Statistics*, 25(2):426–444.
- Klami, A. (2015). Poly-gamma augmentations for factor models. In *Asian Conference on Machine Learning*, pages 112–128.
- Knowles, D. A. and Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*, pages 1701–1709.
- LeSage, J. P. and Pace, R. K. (2007). A matrix exponential spatial specification. *Journal of Econometrics*, 140(1):190–214.
- Linderman, S., Adams, R. P., and Pillow, J. W. (2016a). Bayesian latent structure discovery from multi-neuron recordings. In *Advances in neural information processing systems*, pages 2002–2010.
- Linderman, S., Johnson, M. J., and Adams, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the pólya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pages 3456–3464.
- Linderman, S. W., Miller, A. C., Adams, R. P., Blei, D. M., Paninski, L., and Johnson, M. J. (2016b). Recurrent switching linear dynamical systems. *arXiv preprint arXiv:1610.08466*.

- Luts, J., Wand, M. P., et al. (2015). Variational inference for count response semiparametric regression. *Bayesian Analysis*, 10(4):991–1023.
- Mannering, F. L., Shankar, V., and Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic methods in accident research*, 11:1–16.
- Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154(3):421–441.
- Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096.
- Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., and DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the besag york mollié model in stan. *Spatial and spatio-temporal epidemiology*, 31:100301.
- Narayanamoorthy, S., Paleti, R., and Bhat, C. R. (2013). On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. *Transportation research part B: methodological*, 55:245–264.
- Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.
- Park, M., Foulds, J., Chaudhuri, K., and Welling, M. (2016). Variational bayes in private settings (vips). *arXiv preprint arXiv:1611.00340*.
- Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.
- Ren, Q., Banerjee, S., Finley, A. O., and Hodges, J. S. (2011). Variational bayesian methods for spatial data analysis. *Computational statistics & data analysis*, 55(12):3197–3217.
- Rey, S. J. and Anselin, L. (2010). Pysal: A python library of spatial analytical methods. In *Handbook of applied spatial analysis*, pages 175–193. Springer.
- Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120.
- Rossi, P. E., Allenby, G. M., and McCulloch, R. (2012). *Bayesian statistics and marketing*. John Wiley & Sons.
- Salimans, T., Knowles, D. A., et al. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882.
- Simões, P. and Natário, I. (2016). Spatial econometric approaches for count data: An overview and new directions. *International Journal of Economics and Management Engineering*, 10(1):348–357.
- Strauss, M. E., Mezzetti, M., and Leorato, S. (2017). Is a matrix exponential specification suitable for the modeling of spatial correlation structures? *Spatial statistics*, 20:221–243.

- Tan, L. S., Nott, D. J., et al. (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28(2):168–188.
- US Census Bureau (n.d.). 2012–2016 American Community Survey 5-year estimates.
- Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., and Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, 88(1):36–59.
- Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.
- Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031.
- Wei, W. and Held, L. (2014). Calibration tests for count data. *Test*, 23(4):787–805.
- Wenzel, F., Galy-Fajou, T., Donner, C., Kloft, M., and Opper, M. (2019). Efficient gaussian process classification using pòlya-gamma data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5417–5424.
- Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, pages 434–449.
- Windle, J., Polson, N. G., and Scott, J. G. (2014). Sampling pólya-gamma random variates: alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*.
- Wu, G. (2018). Fast and scalable variational bayes estimation of spatial econometric models for gaussian data. *Spatial statistics*, 24:32–53.
- Zhou, M., Li, L., Dunson, D., and Carin, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Proceedings of the International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access.

Appendix A Conditional posterior update of r in MCMC

To obtain the conditional posterior distribution of the dispersion parameter r in MCMC, we follow the strategy adopted by [Zhou et al. \(2012\)](#). We represent the negative-binomial-distributed count variable as follows:

$$y_i = \sum_{l=1}^{L_i} \chi_{li}, \quad L_i \sim \text{Poisson}(-r \ln(1 - p_i)), \quad \chi_{il} \stackrel{iid}{\sim} \text{Logarithmic}(p_i).$$

Thus, the conditional posterior update of r is:

$$P(r|-) \propto \prod_{i=1}^N P(L_i|r, p_i) P(r|r_0, h),$$

$$r|-\sim \text{Gamma}\left(r_0 + \sum_{i=1}^N L_i, h + \sum_{i=1}^N \ln(1 + \exp(\psi_i))\right). \quad (29)$$

Since the posterior update of r is conditional on L , we also update the conditional posterior of L_i using the following equation:

$$P(L_i = j|-) = R(y_i, j) \quad j = \{0, 1, \dots, y_i\}, \quad (30)$$

$$R(l, m) = \begin{cases} 1 & l = 0; m = 0 \\ \frac{F(l, m)r^m}{\sum_{j=1}^l F(l, j)r^j} & l \neq 0; m \neq 0, \end{cases}$$

$$F(m, j) = \begin{cases} 1 & m = 1 \ \& \ j = 1 \\ 0 & m < j \\ \frac{m-1}{m}F(m-1, j) + \frac{1}{m}F(m-1, j-1) & 1 \leq j \leq m. \end{cases}$$

Appendix B Supplementary material for INFVB

B.1 Important expressions and identities

$$\mathbb{E}[\mathbf{\Omega}] = \begin{bmatrix} \mathbb{E}[\omega_1] & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbb{E}[\omega_N] \end{bmatrix}_{N \times N}, \quad \mathbb{E}[\omega_i] = \left(y_i + \frac{\tilde{b}_r}{\tilde{c}_r}\right) \mathbb{E}\left[\frac{\tanh\left(\frac{\psi_i}{2}\right)}{2\psi_i}\right],$$

$$\mathbb{E}(L_i) = \sum_{j=1}^{y_i} R_{\tilde{r}}(y_i, j)j, \quad \tilde{r} = \exp(\Psi(\tilde{b}_r) - \log(\tilde{c}_r)),$$

$$\mathbb{E}[\mathbf{Z}^*] = \begin{bmatrix} \mathbb{E}[Z_1^*] \\ \vdots \\ \mathbb{E}[Z_N^*] \end{bmatrix}_{N \times 1} = \begin{bmatrix} \frac{y_1 - \tilde{b}_r}{2} \\ \vdots \\ \frac{y_N - \tilde{b}_r}{2} \end{bmatrix}_{N \times 1}, \quad \mathbf{\Lambda}_{\beta} = \begin{bmatrix} \mathbf{\Lambda}_{\beta_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{\Lambda}_{\beta_N} \end{bmatrix}_{NK \times NK},$$

where $\Psi(\cdot)$ is a digamma function. $\mathbb{E}[\log(1 + \exp(\psi_i))]$ and $\mathbb{E}\left[\frac{\tanh\left(\frac{\psi_i}{2}\right)}{2\psi_i}\right]$ are obtained using Gauss-Hermite quadrature ([Abramowitz and Stegun, 1948](#)).

B.2 Important expressions to update $q^*(\Theta_d^{(g)})$

$$\begin{aligned}
\mathbb{E} \left[\ln q(\Theta_c^{(g)} | \Theta_d^{(g)}) \right] &= -\frac{1}{2} \ln |\Lambda_\phi^{(g)}| - \frac{1}{2} \ln |\Lambda_r^{(g)}| - \sum_{i=1}^N \frac{1}{2} \ln |\Lambda_{\beta_i}^{(g)}| - \frac{1}{2} \ln |\Lambda_\mu^{(g)}| + \sum_{k=1}^K \ln \tilde{c}_{a_k}^{(g)} \\
&\quad - \frac{K+1}{2} \ln |\tilde{\mathbf{B}}^{(g)}| + \ln \tilde{c}_h^{(g)} - \tilde{b}_r^{(g)} + \ln \tilde{c}_r^{(g)} - \ln \Gamma(\tilde{b}_r^{(g)}) - (1 - \tilde{b}_r^{(g)}) \Psi(\tilde{b}_r^{(g)}). \\
\mathbb{E} \left[\ln P(\mathbf{y}, \Theta_c^{(g)}, \Theta_d^{(g)}) \right] &= \sum_{i=1}^N \left[\mathbb{E} \left[\ln \Gamma(y_i + r^{(g)}) \right] - \mathbb{E} \left[\ln \Gamma(r^{(g)}) \right] + y_i \lambda_{\psi_i}^{(g)} \right] \\
&\quad - \sum_{i=1}^N \left[\left(y_i + \left[\frac{\tilde{b}_r}{\tilde{c}_r} \right]^{(g)} \right) \mathbb{E} \left[\ln \left(1 + \exp(\psi_i^{(g)}) \right) \right] \right] \\
&\quad + \frac{1}{2} \ln |(\tilde{\Omega})^{(g)}| - \frac{1}{2} \left(\left[\boldsymbol{\lambda}_\phi^T \tilde{\Omega} \boldsymbol{\lambda}_\phi \right]^{(g)} + \text{tr}(\Lambda_\phi(\tilde{\Omega})^{(g)}) \right) - \frac{N}{2} \ln |\tilde{\mathbf{B}}^{(g)}| \\
&\quad - \frac{\tilde{\rho}}{2} \sum_{i=1}^N \left[(\boldsymbol{\lambda}_{\beta_i} - \boldsymbol{\lambda}_\mu)^T \tilde{\mathbf{B}}^{-1} (\boldsymbol{\lambda}_{\beta_i} - \boldsymbol{\lambda}_\mu) + \text{tr}(\tilde{\mathbf{B}}^{-1} \Lambda_{\beta_i}) + \text{tr}(\tilde{\mathbf{B}}^{-1} \Lambda_\mu) \right]^{(g)} \\
&\quad + r_0 \left(-\ln \tilde{c}_h^{(g)} \right) + (r_0 - 1) \left(\Psi(\tilde{b}_r^{(g)}) - \ln \tilde{c}_r^{(g)} \right) - \frac{\tilde{b}_h \tilde{b}_r^{(g)}}{\tilde{c}_h^{(g)} \tilde{c}_r^{(g)}} \\
&\quad + (1 - b_0) \ln \tilde{c}_h^{(g)} - c_0 \frac{\tilde{b}_h}{\tilde{c}_h^{(g)}} - \frac{1}{2} (\boldsymbol{\lambda}_r^{(g)} - \boldsymbol{\zeta}_r)^T \boldsymbol{\Delta}_r^{-1} (\boldsymbol{\lambda}_r^{(g)} - \boldsymbol{\zeta}_r) \\
&\quad - \frac{1}{2} \text{tr}(\boldsymbol{\Delta}_r^{-1} \Lambda_r^{(g)}) + (b_{\sigma^2} - 1) \ln \sigma_{(g)}^{-2} - c_{\sigma^2} \sigma_{(g)}^{-2} \\
&\quad - \frac{(\tau^{(g)} - \zeta_\tau)^2}{2\sigma_\tau^2} - \frac{1}{2} (\boldsymbol{\lambda}_\mu^{(g)} - \boldsymbol{\zeta}_\mu)^T \boldsymbol{\Delta}_\mu^{-1} (\boldsymbol{\lambda}_\mu^{(g)} - \boldsymbol{\zeta}_\mu) - \frac{1}{2} \text{tr}(\boldsymbol{\Delta}_\mu^{-1} \Lambda_\mu^{(g)}) \\
&\quad + \sum_{k=1}^K \left((1 - s) \ln \tilde{c}_{a_k}^{(g)} - \eta_k \frac{\tilde{b}_{a_k}}{\tilde{c}_{a_k}^{(g)}} \right) \\
&\quad - \frac{\rho}{2} \sum_{k=1}^K \ln \tilde{c}_{a_k}^{(g)} - \frac{\rho + K + 1}{2} \ln |\tilde{\mathbf{B}}^{(g)}| - \nu \tilde{\rho} \sum_{k=1}^K \frac{\tilde{b}_{a_k}}{\tilde{c}_{a_k}^{(g)}} (\tilde{\mathbf{B}}^{(g)})_{kk}^{-1}.
\end{aligned} \tag{31}$$

Thus, the conditional ELBO of INFVB for the spatial negative binomial model is obtained by inserting expressions presented in equation 31 in the following equation:

$$\text{Conditional ELBO} = -\mathbb{E}_q \left[\ln q(\Theta_c^{(g)} | \Theta_d^{(g)}) \right] + \mathbb{E}_q \left[\ln P(\mathbf{y}, \Theta_c^{(g)}, \Theta_d^{(g)}) \right]. \tag{32}$$

The optimal conditional distribution of $\Theta_c^{(g)}$ is obtained by maximising the conditional ELBO or equivalently minimising its negative at each grid point (as detailed in equation 20):

$$q^*(\Theta_c^{(g)} | \Theta_d^{(g)}) = \arg \min_{q(\Theta_c^{(g)} | \Theta_d^{(g)})} \mathbb{E}_q \left[\ln q(\Theta_c^{(g)} | \Theta_d^{(g)}) \right] - \mathbb{E}_q \left[\ln P(\mathbf{y}, \Theta_c^{(g)}, \Theta_d^{(g)}) \right]. \tag{33}$$