Hierarchical Bayesian LASSO for a negative binomial regression

Shuai Fu

Dalle Molle Institute for Artificial Intelligence, SUPSI, Switzerland

May 11, 2015

Abstract

Numerous researches have been carried out to explain the relationship between the count data $\mathbf{y}$ and numbers of covariates $\mathbf{x}$ through a generalized linear model (GLM). This paper proposes a hierarchical Bayesian LASSO solution using six different prior models to the negative binomial regression. Latent variables $\mathbf{Z}$ have been introduced to simplify the GLM to a standard linear regression model. The proposed models regard two conjugate zero-mean Normal priors for the regression parameters and three independent priors for the variance: the Exponential, Inverse-Gamma and Scaled Inverse-χ^2 distributions. Different types of priors result in different amounts of shrinkage. A Metropolis-Hastings-within-Gibbs algorithm is used to compute the posterior distribution of the parameters of interest through a data augmentation process. Based on the posterior samples, an original Double Likelihood Ratio Test (DLRT) statistic have been proposed to choose the most relevant covariates and shrink the insignificant coefficients to zero. Numerical experiments on a real-life data set prove that Bayesian LASSO methods achieved significantly better predictive accuracy and robustness than the classical maximum likelihood estimation and the standard Bayesian inference.

Keywords. Hierarchical Bayesian inference, LASSO, Latent variable, Generalized linear regression, Negative binomial, Markov chain Monte Carlo.

1 Introduction

There is a growing interest in explaining the count data $\mathbf{y} = (y_1, \ldots, y_n)^T$ from numerous covariates $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ldots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$ through a generalized linear model (GLM) (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989)

$$\mathbb{E}(\mathbf{y} \,|\, \mathbf{x}) \;=\; g^{-1}(\mu \mathbf{1}_n + \mathbf{x}\boldsymbol{\beta}), \tag{1}$$

where g denotes the canonical link function, μ is the constant and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of p regression coefficients. Due to the specific properties of the count data such as discreteness, nonnegativity and overdispersion, y_i is assumed to follow the Negative Binomial (NB) distribution instead of the Poisson distribution through the logarithmic link function

$$y_i \;\sim\; \mathrm{NB}\left(\phi_i = \exp\left(\mu + \mathbf{x}_i\boldsymbol{\beta}\right),\ \psi\right),\ i = 1, \ldots, n, \tag{2}$$

where ϕ_i is the mean and ψ is the dispersion parameter. Thus $\mathbb{E}(y_i \mid \mathbf{x}_i) = \exp(\mu + \mathbf{x}_i\boldsymbol{\beta})$. It especially allows an overdispersion property, ie. the variance exceeds the mean, which is often the case. For instance, the average daily road accidents in Canton Ticino (Switzerland) in 2007 is 16.57 while its variance is 26.91, considerably exceeding the mean value. Another example arises in stock management, where the recorded daily sale of a clothing item is 159, averaged on 2923 observations summed in 10 retail scores of Switzerland. But its variance is enormous: $3.8e05$.

The frequentist approach is considered to be the classical approach for the GLM, where the regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ are estimated by maximizing the nonlinear log-likelihood, named the MLE. However, the MLE provides only a point estimate which may not be robust, or even has convergence concerns for the quasi-Newton or conjugate-gradient algorithms when solving the embedded optimization problem, if the sample size is small or if the dispersion parameter is much larger than the mean. As an alternative, the Bayesian inference accounts for prior expert knowledge on variables of interest, which can lead to more stable estimates, and it yields a sample of posterior estimators which may be helpful for the uncertainty analysis. The R package *Bayesm* provides a Bayesian solution to the GLM. A hierarchical Bayesian analysis for the NB regression can equally be found in Fu (2014, unpublished results).

As shown in Tibshirani (1996), it would be possible to improve the predictive performance of the Bayesian inference through LASSO by shrinking the insignificant coefficients to zero. LASSO highlights the importance of favoring the sparseness in estimating $\boldsymbol{\beta}$. Mathematically, the LASSO estimates of regression parameters can be achieved by

$$\min_{\boldsymbol{\beta}} \left[(\mathbf{Z} - \mu - \mathbf{x}\boldsymbol{\beta})^T (\mathbf{Z} - \mu - \mathbf{x}\boldsymbol{\beta}) \; + \; \lambda \sum_{k=1}^{p} |\beta_k| \right], \tag{3}$$

where $\lambda \geq 0$ is the LASSO parameter controlling the degree of shrinkage and $\mathbf{Z}$ is the $n \times 1$ vector of responses explained by

$$Z_i \;\; = \;\; \mu + \sum_{k=1}^{p} x_{ik}\beta_k + \epsilon_i \; = \; \mu + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \tag{4}$$

with ϵ_i the independent white noise distributed by $\mathrm{N}(0, \sigma^2)$. As in Tibshirani (1996) and Park and Casella (2008), the LASSO estimates for the regression coefficients $\boldsymbol{\beta}$ are equivalent to Bayesian posterior mode estimates when assigning an independent Laplace prior for it. The Laplace prior, equivalently the Double-exponential prior, is known to promote sparseness (Williams, 1995). Furthermore, either the marginal-likelihood maximization (Park and Casella, 2008) or the full Bayesian analysis (Bae and Mallick, 2004; Yi and Xu, 2008) can be performed on approximating the LASSO parameters. The full Bayesian analysis has been applied in this paper.

However, up to now there are not yet Bayesian LASSO methods which focus on the current negative binomial regression: they treat either the standard linear regression (Park and Casella, 2008; Yi and Xu, 2008) or the probit regression with $\mathbf{y}$ taking values in $\{0, 1\}$ (Bae and Mallick, 2004), and none of those methods propose a formal and model-specific way to select the relevant variables. With help of latent variables, this paper provides an extension of the Bayesian LASSO from the linear regression model to the negative binomial GLM. It proposes the use of six different priors to promote different degrees of sparseness following a three-level hierarchical model (see Figure 1). The hierarchical structure helps accounting for the dependence among different groups of observed data $\mathbf{y}$ by assuming common hyperparameters in higher probabilistic levels.

The six prior distributions regard two options for β_k: the independent zero-mean Gaussian distributions conditioning or not on the residual variance σ^2; three options for the variance τ_k^2 of β_k: the Exponential distribution, the Inverse-Gamma distribution and the Scaled Inverse-χ^2 distribution. Mixing the zero-mean Gaussian prior for β_k with each of those priors for τ_k^2 is equivalent to the Laplace prior (Tibshirani, 1996), the Student's t-prior (Li *et al.*, 2002) and the Scaled Student's t-prior (Sorense and Gianola, 2002; Gelman *et al.*, 2003), respectively. The hyperparameters in the priors are treated as unknowns and estimated from a noninformative hyperprior distribution. The advantage is that the degree of shrinkage can be totally justified by the data without any subjective influence.

Based on different priors for the regression parameters and the hyperprior for hyperparameters, we developed a full Bayesian analysis using the hybrid MCMC algorithm, namely the Gibbs sampler combined with a Metropolis-Hastings algorithm (*Metropolis-Hastings-within-Gibbs algorithm*) (see Robert and Casella, 2004). The embedded Metropolis-Hastings algorithm enables the simulation of the unknown conditional posterior distributions through a Markov chain. The convergence of the Markov chain was checked with help of the Brooks-Gelman statistic (Brooks and Gelman, 1998) and the Hellinger distance statistic (Boone et *al.*, 2012). The computation of the optimal number of relevant variables and the selection of those influential ones were then carried out. We proposed the Double Likelihood Ratio Test (DLRT) statistics for variable selection. Numerical experiments on a real-life data set confirm the outstanding performance of Bayesian LASSO methods for the negative binomial regression, compared with the MLE and the standard Bayesian inference. The paper is organized as follows. In Section 2, we introduce our hierarchical model. Different prior distributions have been introduced in Section 3 and the corresponding posterior distributions have been computed in Section 4. Section 5 explains the LASSO approach to select the significantly important variables. After a typical application presented in Section 6, a conclusion section ends this paper.

2 Modeling

In the regression analysis of counts, our major objective is to explain the count data from numerous covariates, especially to identify the most tightly linked covariates and to estimate the magnitudes of those covariates. For simplicity purposes, we explain our Bayesian context on the example of regional crash counts.

In each region j, we have n independently observed crash counts $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn})^T$. By introducing n independent latent variables $\mathbf{Z}_j = (Z_{j1}, \ldots, Z_{jn})^T$, the distribution of y_{ji} can be written as $y_{ji} \,|\, Z_{ji} \sim \text{NB}\ (\exp(Z_{ji}),\ \psi)$, where ψ is the common dispersion parameter for all regions and Z_{ji} can be explained by a linear regression model

$$Z_{ji} \;=\; \mu_j + \mathbf{x}_{ji}\boldsymbol{\beta}_j + \epsilon_{ji} \;=\; \mu_j + \sum_{k=1}^{p} x_{jik}\beta_{jk} + \epsilon_{ji}, \tag{5}$$

with μ_j the overall mean for region j, p the number of covariates, x_{jik} the k-th covariate of period i for region j, β_{jk} the effect of the k-th covariate for region j and $\epsilon_{ji} \sim \text{N}(0, \sigma_j^2)$ the white noise. The probability density of y_{ji} is

$$f(y_{ji} \,|\, Z_{ji}) \;=\; \frac{\Gamma(y_{ji}+\psi)}{y_{ji}!\,\Gamma(\psi)} \left[\frac{\exp(Z_{ji})}{\exp(Z_{ji})+\psi}\right]^{y_{ji}} \left[\frac{\psi}{\exp(Z_{ji})+\psi}\right]^{\psi}.$$

We have $\mathbb{E}(y_{ji} \,|\, Z_{ji}) = \exp(Z_{ji})$ and $\text{Var}(y_{ji} \,|\, Z_{ji}) = \exp(Z_{ji})\,[1 + \exp(Z_{ji})/\psi]$, where the variance exceeds the mean since $\psi > 0$. Decreasing the dispersion parameter ψ

towards 0 corresponds to increasing the overdispersion effect; increasing ψ towards $+\infty$ leads to the convergence towards the Poisson distribution Pois $(\exp(Z_{ji}))$. Introducing latent variables $\mathbf{Z}_j$ simplifies the GLM to a standard linear regression model based on $\mathbf{Z}_j$ as observations. In fact, we have

$$\mathbb{E}(y_{ji}) = \mathbb{E}[\mathbb{E}(y_{ji} \mid Z_{ji})] = \mathbb{E}[\exp(Z_{ji})] \quad (6)$$

$$= \exp\left(\mu_j + \mathbf{x}_{ji}\boldsymbol{\beta}_j + \frac{\sigma_j^2}{2}\right) \simeq \exp(\mu_j + \mathbf{x}_{ji}\boldsymbol{\beta}_j), \quad (7)$$

which mainly insures the essential property of GLM that $\mathbb{E}(y_{ji}) = g^{-1}(\mu_j + \mathbf{x}_{ji}\boldsymbol{\beta}_j) = \exp(\mu_j + \mathbf{x}_{ji}\boldsymbol{\beta}_j)$. Note that the passage from (6) to (7) is derived from the fact that $U := \exp Z_{ji} \sim \log \mathrm{N}(\mu_j + \mathbf{x}_{ji}\boldsymbol{\beta}_j, \sigma_j^2)$ a Log-normal distribution, and the last almost equality stands because the residual variance σ_j^2 is usually extremely small. For instance, it is of order $o(10^{-8} - 10^{-5})$ from an empirical study of the dispersed count data in Fu (2014, unpublished results) and a simulation study in uncertainty analysis in Fu et *al.* (2015).

3 Prior distributions

This study is based on a three-level hierarchical Bayesian modeling: the first level involves the NB regression model with parameters ψ and Z_{ji}, the second level is developed from Z_{ji} as a linear regression model with parameters μ_j, $\boldsymbol{\beta}_j$ and σ_j^2 and the third level regards the variance parameter τ_j^2 of $\boldsymbol{\beta}_j$ and the related hyperparameters θ_j. As shown in Figure 1, we assume that different regions share an overall dispersion parameter ψ which allows an inter-dependence among them and in each region, Z_{ji} is the bridge between the NB regression and the linear regression. In the GLM, the prior distribution of Z_{ji} is naturally chosen as $\mathrm{N}(\mu_j + \mathbf{x}_{ji}\boldsymbol{\beta}_j, \sigma_j^2)$ and ψ is typically assumed to follow a Gamma distribution $\mathrm{G}(u, v)$ because of its positive nature. In the linear model, Z_{ji} is used as the observed data which is usually sufficient to estimate the overall mean μ_j and the residual variance σ_j^2. Thus we can choose noninformative priors for them: an independent flat prior for μ_j as $\pi(\mu_j) \propto 1$ and the Jeffreys noninformative prior for σ_j^2 as $\pi(\sigma_j^2) \propto 1/\sigma_j^2$.

As presented in Section 1, the LASSO method is applied to the current problem to remove the irrelevant coefficients which contribute little in predicting the count data but introduce additional modeling error. For this purpose, the prior distribution of $\boldsymbol{\beta}_j$ is chosen to promote sparseness. We consider three options.

- Park and Casella (2008) suggested a commonly used prior as the centered Laplace distribution, or equivalently the Double-exponential distribution as

$$\pi(\boldsymbol{\beta}_j) = \text{Laplace}\left(\boldsymbol{\beta}_j; 0, \frac{1}{\lambda_j}\right) = \prod_{k=1}^{p} \frac{\lambda_j}{2} \exp(-\lambda_j |\beta_{jk}|), \quad (8)$$

  where λ_j is the inverse scale parameter. By applying this Laplace prior, the posterior mode estimate of $\boldsymbol{\beta}_j$ is equivalent to the LASSO estimate (Tibshirani, 1996; Park and Casella, 2008).

- Another centered and wide-tailed distribution is the Student's t-distribution as

$$\pi(\boldsymbol{\beta}_j) = \mathrm{St}(\boldsymbol{\beta}_j; e_j) \propto \prod_{k=1}^{p} \left(1 + \frac{\beta_{jk}^2}{e_j}\right)^{-(e_j+1)/2}, \quad (9)$$

  where $e_j > 0$ is the degrees of freedom.

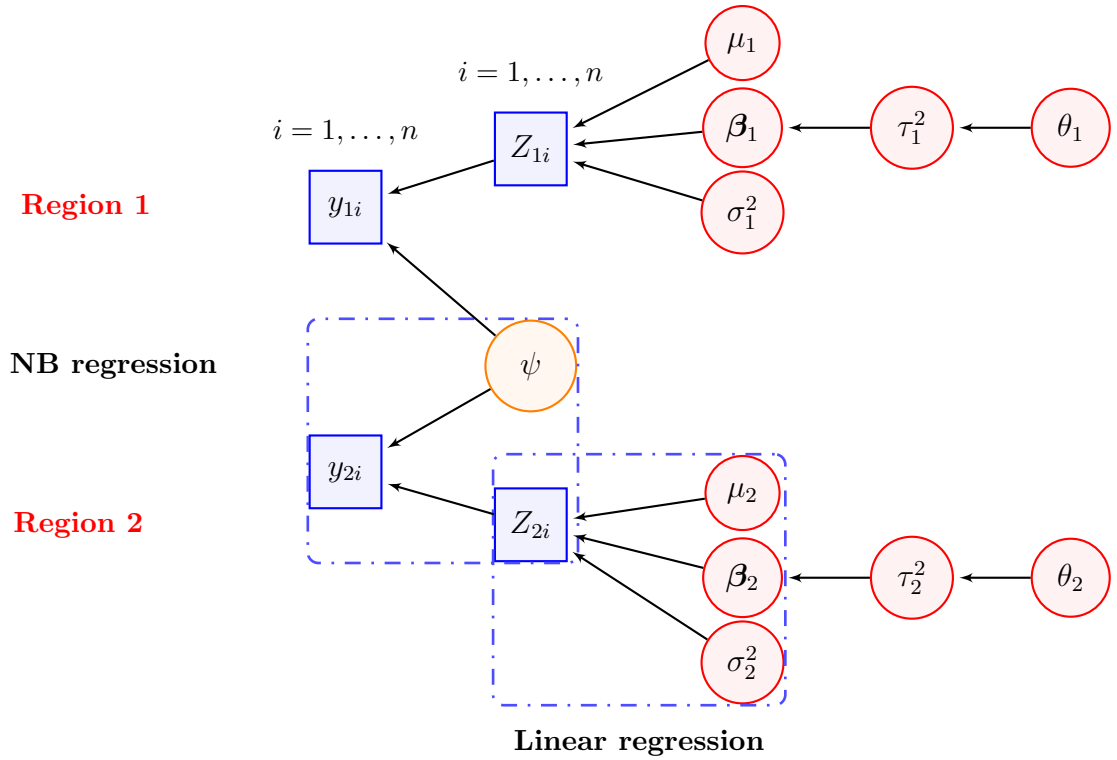

Figure 1: DAG (Directed Acyclic Graph) of three-level hierarchical Bayesian model in 2 regions.

- The third wide-tailed distribution is the Scaled Student's t-distribution (Sorensen and Gianola, 2002; Gelman *et al.*, 2003) as

$$\pi(\boldsymbol{\beta}_j) \quad = \quad \mathrm{SSt}(\boldsymbol{\beta}_j;\, \nu_j, s_j^2) \;\propto\; \prod_{k=1}^{p} \left(1 + \frac{\beta_{jk}^2}{\nu_j s_j^2}\right)^{-(\nu_j+1)/2}, \tag{10}$$

  where $\nu_j > 0$ is the degrees of freedom and s_j^2 is the scale parameter.

To ease the analysis, we present those priors as a two-level hierarchical model (see for instance Griffin and Brown, 2006) where the first level is in common, assuming an independent Normal distribution for the coefficients β_{jk} with mean zero and variance γ_{jk}^2

$$\boldsymbol{\beta}_j \,|\, \tau_j^2 \quad \sim \quad \prod_{k=1}^{p} \mathrm{N}\left(\beta_{jk};\, 0, \tau_{jk}^2\right), \tag{11}$$

and the second level ensures an independent prior for the variances $\tau_j^2 = (\tau_{j1}^2, \ldots, \tau_{jp}^2)$

$$\tau_j^2 \,|\, \theta_j \quad \sim \quad \prod_{k=1}^{p} \pi(\tau_{jk}^2 \,|\, \theta_j), \tag{12}$$

with θ_j the hyperparameters. The two-level hierarchical model multiplies the densities of those two distributions (11) and (12) according to the formula of Bayes. Different priors of τ_j^2 induce different amounts of shrinkage. Apart from the common Normal prior (11) of $\boldsymbol{\beta}_j$, the prior for the variance τ_j^2 takes three forms according to the three mixed priors (8), (9) and (10).

- The Laplace distribution (8) can be presented as a scale mixture of (11) with an Exponential mixing density for τ_{jk}^2

$$\pi(\tau_{jk}^2 \,|\, \theta_j) \;=\; \text{Expon}\left(\tau_{jk}^2;\, \frac{\lambda_j^2}{2}\right) \;=\; \frac{\lambda_j^2}{2} \exp\left(-\frac{\lambda_j^2 \tau_{jk}^2}{2}\right), \tag{13}$$

  where $\lambda_j^2/2$ is the inverse scale parameter and λ_j represents the LASSO parameter in formula (3) for region j, which is essential to control the shrinkage. Careful considerations are required for its computation. Rather than selecting through the marginal-likelihood maximization as in Park and Casella (2008), we assign a prior distribution to λ_j so that it can be simulated along with other parameters and the uncertainty with the shrinkage can be automatically accounted for. Regarding hyperpriors, a conjugate Gamma prior $\text{G}(a, b)$ has been assigned to λ_j^2 (not λ_j but λ_j^2). By setting a and b to small values ($a = b = 0.1$ for instance), the prior for λ_j^2 becomes quite weakly informative.

- The Student's t-distribution (9) can be written as a scale mixture of (11) with an Inverse-Gamma mixing distribution $\text{InvGamma}(e_j/2, e_j/2)$. Following Bae and Mallick (2004), we assign a more general prior density $\text{InvGamma}(e_j/2, d_j/2)$ as

$$\pi(\tau_{jk}^2 \,|\, \theta_j) \;=\; \text{InvGamma}\left(\tau_{jk}^2;\, \frac{e_j}{2}, \frac{d_j}{2}\right) \;=\; \left(\frac{1}{\tau_{jk}^2}\right)^{e_j/2+1} \exp\left(-\frac{d_j}{2\tau_{jk}^2}\right), \tag{14}$$

  where $e_j/2$ and $d_j/2$ are the shape and scale parameters to be adjusted. We treat them as unknowns by assigning a Uniform prior on $1/e_j$ in the range $]0, 1]$ and a Uniform prior on d_j in the range $]0, D]$, with D large enough (see Gelman *et al.*, 2003).

- The scaled Student's t-distribution (10) can be expressed as a scale mixture of (11) with the Scaled Inverse-χ^2 density $\text{SInv-}\chi^2(\nu_j, s_j^2)$, or equivalently the Inverse-Gamma density $\text{InvGamma}(\nu_j/2,\, \nu_j s_j^2/2)$

$$\pi(\tau_{jk}^2 \,|\, \theta_j) \;=\; \text{SInv-}\chi^2\left(\tau_{jk}^2;\, \nu_j, s_j^2\right) \;=\; \left(\tau_{jk}^2\right)^{-\nu_j/2-1} \exp\left(-\frac{\nu_j s_j^2}{2\tau_{jk}^2}\right), \tag{15}$$

  with ν_j the degrees of freedom and s_j^2 the scale parameter. As in the previous case, we assign a Uniform prior on $1/\nu_j$ in the range $]0, 1]$ and a Uniform prior on s^2 in the range $]0, D]$ with D a large number. We can treat those parameters as unknowns and simulate them along with others. e_j, d_j, ν_j and s_j^2 are indeed LASSO parameters as they affect directly the shrinkage for β_{jk} through the variance τ_{jk}^2.

As proposed in Park and Casella (2008), an alternative prior distribution for $\boldsymbol{\beta}_j$ which accounts for the residual variance σ_j^2 is as follows

$$\boldsymbol{\beta}_j \,|\, \tau_j^2 \;\sim\; \prod_{k=1}^{p} \text{N}\left(\beta_{jk};\, 0, \sigma_j^2 \tau_{jk}^2\right). \tag{16}$$

The advantage of the Normal prior (16) over the prior (11) is that it can ensure the unimodal posterior of $\boldsymbol{\beta}_j$ (see Park and Casella, 2008 for the demonstration). Both prior choices are numerically compared in the experiment section. Up to now, we have proposed two different priors for β_{jk} and three different priors for τ_{jk}^2 which are

$$\begin{aligned}
\pi_1(\beta_{jk} \,|\, \tau_{jk}^2) &= \mathrm{N}(0, \tau_{jk}^2), & \pi_1(\tau_{jk}^2 \,|\, \theta_j) &= \mathrm{Expon}(\lambda_j^2/2), \\
\pi_2(\beta_{jk} \,|\, \tau_{jk}^2) &= \mathrm{N}(0, \sigma_j^2\tau_{jk}^2); & \pi_2(\tau_{jk}^2 \,|\, \theta_j) &= \mathrm{InvGamma}(e_j/2, d_j/2), \\
& & \pi_3(\tau_{jk}^2 \,|\, \theta_j) &= \text{SInv-}\chi^2(\nu_j, s_j^2).
\end{aligned}$$

They yield six possible combinations and thus six different Bayesian LASSO priors as in Table 1. In the next section, the posterior distribution is computed for each of the priors.

Table 1: Six Bayesian LASSO priors and their respective compositions.

Priors	$\pi_1(\tau_{jk}^2 \,\vert\, \theta_j)$	$\pi_2(\tau_{jk}^2 \,\vert\, \theta_j)$	$\pi_3(\tau_{jk}^2 \,\vert\, \theta_j)$
$\pi_1(\beta_{jk} \,\vert\, \tau_{jk}^2)$	prior I	prior II	prior III
$\pi_2(\beta_{jk} \,\vert\, \tau_{jk}^2)$	prior IV	prior V	prior VI

4 Posterior computation

In the hierarchical Bayesian inference, the hybrid MCMC algorithm is applied on both the NB regression and linear regression models. It is also named the Metropolis-Hastings-within-Gibbs algorithm (see Robert and Casella, 2004), which is based on the joint conditional posterior distribution of all the parameters, derived from

$$\pi(\mathbf{Z}_j, \psi \,|\, \mathbf{y}_j, \mu_j, \boldsymbol{\beta}_j, \sigma_j^2, u, v) \propto \prod_{i=1}^{n} \pi(y_{ji} \,|\, Z_{ji}, \psi) \cdot \pi(Z_{ji} \,|\, \mu_j, \boldsymbol{\beta}_j, \sigma_j^2) \cdot \pi(\psi \,|\, u, v) \tag{17}$$

$$\begin{aligned}
\pi(\mu_j, \boldsymbol{\beta}_j, \sigma_j^2, \tau_j^2, \theta_j \,|\, \mathbf{Z}_j) \propto & \prod_{i=1}^{n} \pi(Z_{ji} \,|\, \mu_j, \boldsymbol{\beta}_j, \sigma_j^2) \cdot \pi(\mu_j) \cdot \pi(\sigma_j^2) \\
& \cdot \prod_{k=1}^{p} \pi(\beta_{jk} \,|\, \tau_{jk}^2) \cdot \pi(\tau_{jk}^2 \,|\, \theta_j) \cdot \pi(\theta_j),
\end{aligned} \tag{18}$$

in region j, where $\theta_j = \lambda_j$ for priors I and IV, $\theta_j = (e_j, d_j)$ for priors II and V, and $\theta_j = (\nu_j, s_j^2)$ for priors III and VI. In the NB regression model (17), the conditional posterior distributions of $\mathbf{Z}_j$ and ψ do not belong to any known family of distributions. Numerical methods, for instance the Metropolis-Hastings (MH) algorithm, are thus necessary for their simulations. In Fu (2014, unpublished results), several alternatives of instrumental distributions have been numerically compared and the Gaussian density using the previous value as mean and an originally constructed positive-definite matrix as variance is well fitting. We inherit this instrumental distribution in our MH step. In the second linear regression model (18), the conditional posterior distributions of μ_j and σ_j^2 are derived from their noninformative priors as Normal and Scaled Inverse-χ^2 distributions

$$\pi(\mu_j \,|\, \mathbf{Z}_j, \sigma_j^2, \boldsymbol{\beta}_j) \propto \mathrm{N}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_{ji} - \mathbf{x}_{ji}\boldsymbol{\beta}_j),\ \frac{1}{n}\sigma_j^2\right), \tag{19}$$

$$\pi(\sigma_j^2 \,|\, \mathbf{Z}_j, \mu_j, \boldsymbol{\beta}_j) \propto \text{SInv-}\chi^2\left(n,\ \frac{1}{n}\sum_{i=1}^{n}(Z_{ji} - \mu_j - \mathbf{x}_{ji}\boldsymbol{\beta}_j)^2\right). \tag{20}$$

With zero-mean Gaussian priors for $\boldsymbol{\beta}_j$, its conditional posterior distribution is

$$\pi(\boldsymbol{\beta}_j \,|\, \mathbf{Z}_j, \mu_j, \sigma_j^2) \quad \propto \quad \mathrm{N}\left(A_j^{-1}\mathbf{x}_j\,(\mathbf{Z}_j - \mu_j \mathbf{1}_n),\ \sigma_j^2 A_j^{-1}\right), \tag{21}$$

$$\text{with } A_j \quad = \quad \begin{cases} \mathbf{x}_j^T\mathbf{x}_j + \sigma_j^2 \mathrm{Diag}^{-1}(\tau_{j1}^2, \ldots, \tau_{jp}^2), & \text{for priors I,II,III;} \\ \mathbf{x}_j^T\mathbf{x}_j + \mathrm{Diag}^{-1}(\tau_{j1}^2, \ldots, \tau_{jp}^2), & \text{for priors IV,V,VI.} \end{cases}$$

For priors I and IV, the conditional posterior distribution of τ_{jk}^{-2} (not τ_{jk}^2 but its inverse) is an Inverse-Gaussian distribution

$$\pi(\tau_{jk}^{-2} \,|\, \mathbf{Z}_j, \beta_{jk}, \sigma_j^2, \lambda_j^2) \quad \propto \quad \mathrm{InvGauss}\left(\frac{\sqrt{\lambda_j^2}}{|\gamma_{jk}|},\ \lambda_j^2\right),$$

where $\gamma_{jk} = \beta_{jk}$ for prior I and $\gamma_{jk} = \beta_{jk}/\sigma_j$ for prior IV. For priors II and V, the conditional posterior distribution of τ_{jk}^2 is still an Inverse-Gamma distribution

$$\pi(\tau_{jk}^2 \,|\, \mathbf{Z}_j, \beta_{jk}, \sigma_j^2, e_j, d_j) \quad \propto \quad \mathrm{InvGamma}\left(\frac{e_j + 1}{2},\ \frac{\gamma_{jk}^2 + d_j}{2}\right),$$

where $\gamma_{jk}^2 = \beta_{jk}^2$ for prior II and $\gamma_{jk}^2 = \beta_{jk}^2/\sigma_j^2$ for prior V. For priors III and VI, the conditional posterior distribution of τ_{jk}^2 is still a Scaled Inverse-χ^2 distribution as follows

$$\pi(\tau_{jk}^2 \,|\, \mathbf{Z}_j, \beta_{jk}, \sigma_j^2, \nu_j, s_j^2) \quad \propto \quad \text{SInv-}\chi^2\left(\nu_j + 1,\ \frac{\gamma_{jk}^2 + \nu_j s_j^2}{\nu_j + 1}\right),$$

where $\gamma_{jk}^2 = \beta_{jk}^2$ for prior III and $\gamma_{jk}^2 = \beta_{jk}^2/\sigma_j^2$ for prior VI. Regarding the hyperparameters, we update λ_j^2 from the Gamma distribution $\mathrm{G}\left(p + a,\ \sum_{k=1}^p \tau_{jk}^2/2 + b\right)$ for priors I and IV; update d_j from the Gamma distribution $\mathrm{G}\left(e_j p/2 + 1,\ \sum_{k=1}^p 1/\,2\tau_{jk}^2\right)$ and e_j using a Metropolis-Hastings step for priors II and V; and update s_j^2 from the Gamma distribution $\mathrm{G}\left(\nu_j p/2 + 1,\ \sum_{k=1}^p \nu_j/\,2\tau_{jk}^2\right)$ and ν_j using a Metropolis-Hastings step for priors III and VI. See Appendix A for more details about the posterior computation. Applying the improper priors for μ and σ^2, their posterior distributions can be proven to be proper. See Appendix B for the detailed proof. The hybrid MCMC algorithm has then been carried out. The convergence has been checked using the Brooks-Gelman (BG) statistic (Brooks and Gelman, 1998) computed on three parallel Markov chains. A classic rule of thumb is to suppose quasi-stationarity once the statistic stably remains under 1.1 (Brooks and Gelman, 1998). It has been obtained by using the second half run of Metropolis-Hastings iterations and Gibbs iterations after the chosen burn-in periods. An alternative to the BG statistic for MCMC diagnostic is the Hellinger distance (HD) approach (see Boone et *al.*, 2014), which can determine a discrepancy between two distributions and also diagnose some "sticky" MCMC chains. It does not own an universal cut-off value for the dissimilarity, it has thus been applied as an additional diagnostic for the MCMC convergence. Once the convergence has been accepted for each parameter of interest, we can start to sample the drawn variables and further statistical tests can be carried out on those collected samples.

5 Variable selection

This particular issue of the Bayesian LASSO approach focuses on selecting the influential covariates by shrinking the insignificant coefficients to zero in order to improve

the performance of prediction. This selection is based on posterior simulated samples. Hoti and Sillanpää (2006) suggested to set up a threshold $c > 0$ such that β_{jk} is to be kept for region j if the absolute value of its estimate exceeds c. However, a formal choice of this shrinkage criterion was missing. In gene selection, Li *et al.* (2002), Bae and Mallick (2004) observed that with prior I many simulations of the LASSO parameter λ_{jk} approached zero and they proposed to remove the k-th coefficient β_{jk} if $\lambda_{jk}^2 < 10^{-12}$. However, this fixed threshold is neither model-specific nor purpose-oriented. Here we propose a model-depending alternative named the Double Likelihood Ratio Selection (DLRS), which is based on the Double Likelihood Ratio Test statistics (DLRT). It is achieved in four steps:

1. In region j, for each k-th covariate we compute the Likelihood Ratio Test (LRT_{jk}) statistic from the posterior estimates $(\hat{\mu}_j, \hat{\boldsymbol{\beta}}_j, \hat{\psi}_j)$

$$\begin{aligned}\mathrm{LRT}_{jk} &= 2\log\frac{\prod_{i=1}^{n}\mathrm{NB}\left(y_{ji};\,\exp\left(\hat{\mu}_j+\mathbf{x}_{ji}\hat{\boldsymbol{\beta}}_j\right),\,\hat{\psi}_j\right)}{\prod_{i=1}^{n}\mathrm{NB}\left(y_{ji};\,\exp\left(\hat{\mu}_j+\mathbf{x}_{ji}\hat{\boldsymbol{\beta}}_j-x_{jik}\hat{\beta}_{jk}\right),\,\hat{\psi}_j\right)} \\ &= 2\log\mathscr{L}_{\mathrm{NB}}(\hat{\mu}_j,\hat{\boldsymbol{\beta}}_j,\hat{\psi}_j) - 2\log\mathscr{L}_{\mathrm{NB}}(\hat{\mu}_j,\hat{\boldsymbol{\beta}}_{j,-k},\hat{\beta}_{jk}=0,\hat{\psi}_j),\end{aligned}$$

   with $\hat{\boldsymbol{\beta}}_{j,-k} = (\hat{\beta}_{j1},\ldots,\hat{\beta}_{jk-1},\hat{\beta}_{jk+1},\ldots,\hat{\beta}_{jp})$ and $\mathscr{L}_{\mathrm{NB}}$ the likelihood of the NB distribution. LRT_{jk} means the difference between the log-likelihood of the NB regression containing all the covariates and the log-likelihood removing the k-th covariate out. It is positive-defined and can be used as a good measurement of the significance of covariates: the larger LRT_{jk}, the more relevant the k-th covariate.

2. We sort the vector $(\mathrm{LRT}_{j1},\ldots,\mathrm{LRT}_{jp})$ in decreasing order as $(\mathrm{LRT}_{j(1)} > \cdots > \mathrm{LRT}_{j(p)})$ and construct p embedding groups of estimated coefficients starting from the most influential one, ie. $\hat{\boldsymbol{\beta}}'_{j1} = \{\hat{\beta}_{j(1)}\} \subseteq \hat{\boldsymbol{\beta}}'_{j2} = \{\hat{\beta}_{j(1)},\hat{\beta}_{j(2)}\} \subseteq \cdots \subseteq \hat{\boldsymbol{\beta}}'_{jp} = \{\hat{\beta}_{j(1)},\ldots,\hat{\beta}_{j(p)}\}$. Therefore, $\hat{\boldsymbol{\beta}}'_{jk}$ groups the estimates of k most significant coefficients.

3. For each significant group of size k, we recompute the Likelihood Ratio Test statistics as

$$\mathrm{DLRT}_{jk} = 2\log\mathscr{L}_{\mathrm{NB}}(\hat{\mu}_j,\hat{\boldsymbol{\beta}}_j,\hat{\psi}_j) - 2\log\mathscr{L}_{\mathrm{NB}}(\hat{\mu}_j,\hat{\boldsymbol{\beta}}'_{jk},\hat{\boldsymbol{\beta}}'_{j,-k}=0,\hat{\psi}_j), \tag{22}$$

   where $\hat{\boldsymbol{\beta}}'_{j,-k} = (\hat{\beta}_{j(k+1)},\ldots,\hat{\beta}_{j(p)})$. DLRT_{jk} is in fact a positive-define decreasing-to-zero function, which measures the fitting quality using each influential group $\hat{\boldsymbol{\beta}}'_{jk}$ compared with the goodness using the full sample $\hat{\boldsymbol{\beta}}_j$. The positivity and homogeneity come from the fact that a model with more parameters will always fit better or at least as well as the model with less parameters, and $\mathrm{DLRT}_{jp} = 0$ since $\hat{\boldsymbol{\beta}}'_{jp} = \hat{\boldsymbol{\beta}}_j$. Especially, DLRT_{jk} is asymptotically distributed by the χ^2 distribution with $(p-k)$ degrees of freedom (Huelsenbeck and Crandall, 1997). The full sample $\hat{\boldsymbol{\beta}}_j$ fits *significantly* better than $\hat{\boldsymbol{\beta}}'_{jk}$ only if DLRT_{jk} exceeds the related percentile at a certain level of significance. The next step thus follows.

4. We search the optimal significant group $\hat{\boldsymbol{\beta}}'_{jk^*}$ such that it involves as few coefficients as possible and there will be little fitting improvement by enlarging the group size from k^*. Thus,

$$k^* = \arg\max_k\left[\mathrm{DLRT}_{jk} > \mathrm{F}^{-1}_{\chi^2_{p-k}}(0.95)\right] \tag{23}$$

   and the k^* selected coefficients would be $\boldsymbol{\beta}'_{jk^*} = \{\beta_{j(1)},\ldots,\beta_{j(k^*)}\}$. Consequently, the model becomes statistically equivalent with any additional coefficient other than $\boldsymbol{\beta}'_{jk^*}$. The covariates additionally considered will not improve the model but introduce unnecessary modeling error and simulation uncertainty. It is worth noting that $\mathrm{DLRT}_{jp} = \mathrm{F}^{-1}_{\chi^2_0}(0.95) = 0$ and if we have

$\mathrm{DLRT}_{j1} < \mathrm{F}^{-1}_{\chi^2_{p-1}}(0.95)$ for $k = 1$, it means that without considering any coefficient the model is already well fitting. Thus, $\hat{y}_{ji} = \exp(\mu_j)$ and $k^* = 0$ in this extreme case.

6 Application: regional road crash counts

This experiment regards $1,024$ daily road crash counts registered in 3 years (2007, 2008 and 2010) and $R = 4$ regions of Switzerland. The first 344 counts were used as the training set and the last 680 counts constituted the test set. Thus, we predicted the number of accidents in 2008 and 2010 by using the registered accident counts in 2007. The 8 considerable covariates included the rainfall, the temperature, their interaction, the daily traffic on the main highway, the crash counts on the previous day, the total crash counts on the previous two days, those on the previous three days and previous four days. The summary statistics about the crash counts and the main covariates on the training set and test set are provided in Table 2. As the major daily traffic owned extremely high average and high standard deviation, we chose thus its logarithmic form to make it comparable with other covariates. Moreover, we replace all the temperatures higher than 10 by 10, since the effect is not linear and there will be little additional effect on the crash counts if the temperature is elevate enough (higher than 10). It is worth noting that the interaction $\mathbf{x}_3$ between the temperature and rainfall often varies highly from the training set to the test set, as marked in **black** in Table 2. For instance, in region 3, the minimum is -17.2 in the training set and -239.1 in the test set, the maximum varies from 73.4 to 215.9 and the standard deviation varies from 5.51 to 16.38. This nonhomogeneity raises the difficulty of getting a stable solution to the NB regression, as the prediction error can be exponentially amplified due to the exponential link function of the GLM. Besides, the last column the coefficient of dispersion does not vary much from region to region, which confirms our assumption that all the four regions share an overall dispersion parameter.

Table 2: Summary statistics of regional crash counts and four main covariates in the training set and test set.

Regions	Data & Covariates	Min.		Max.		Average		St. dev.		Dispersion (σ^2/μ)	
		Train.	Test	Train.	Test	Train.	Test	Train.	Test	Train.	Test
1	Counts $\mathbf{y}_1$	0	1	22	22	8.74	8.61	3.64	3.38	1.52	1.33
	Rainfall	0	0	90.9	102.0	3.32	5.37	10.16	12.16	/	/
	Temperature	-2.6	-5.0	10.0	10.0	7.54	6.65	3.37	3.97	/	/
	Interaction	-0.4	-4.4	**15.9**	**193.2**	**0.08**	**1.08**	**1.05**	**10.56**	/	/
	Log. Traffic	8.64	8.63	9.66	9.74	9.40	9.43	0.25	0.26	/	/
2	Counts $\mathbf{y}_2$	0	0	4	4	0.36	0.41	0.63	0.67	1.10	1.10
	Rainfall	0	0	98.5	104.5	3.42	7.12	9.42	19.53	/	/
	Temperature	-6.0	-7.7	10.0	10.0	6.18	5.08	4.33	5.00	/	/
	Interaction	**-17.2**	**-239.1**	**73.4**	**215.9**	0.64	0.91	**5.51**	**16.48**	/	/
	Log. Traffic	8.33	7.89	9.50	9.52	9.00	8.95	0.21	0.24	/	/
3	Counts $\mathbf{y}_3$	0	0	17	15	5.72	5.36	2.70	2.57	1.28	1.24
	Rainfall	0	0	82.5	123.0	3.74	5.35	10.75	13.26	/	/
	Temperature	-7.8	-11.4	10.0	10.0	5.47	4.78	5.18	5.48	/	/
	Interaction	**-2.56**	**-92.8**	159.8	121.2	1.34	0.79	11.86	10.77	/	/
	Log. Traffic	9.68	9.63	10.71	10.72	10.42	10.40	0.18	0.20	/	/
4	Counts $\mathbf{y}_4$	0	0	6	8	1.68	1.69	1.31	1.39	1.02	1.15
	Rainfall	0	0	56.6	83.2	2.80	4.25	7.03	10.33	/	/
	Temperature	-6.5	-9.2	10.0	10.0	5.89	4.85	4.34	5.07	/	/
	Interaction	**-5.2**	**-52.1**	**28.4**	**169.1**	0.46	0.61	**2.62**	**9.66**	/	/
	Log. Traffic	9.55	9.11	10.95	11.02	10.22	10.21	0.26	0.27	/	/

The regression coefficients were estimated through the MLE, the Bayesian regression (by the Bayesm package) with the six Bayesian LASSO priors. In the Bayesian context, we assumed that no expert knowledge was available by taking noninformative or weakly informative priors. The posterior estimates were thus comparable with the MLEs. Moreover, three parallel Markov chains were simulated within 60,000 Gibbs sampling iterations, with a discarded burn-in period of 10,000 iterations. The convergence was verified by the BG statistic computed on the parallel Markov chains and the Hellinger distance computed on one randomly chose Markov chain. Once the convergence had been reached, every 50-th sample was collected to generate 1,000 quasi-i.i.d. samples. The lag 50 was chosen from an ACF statistical test. In the following, we first present the posterior simulations of the main variables and the shrinkage procedure with help of the DLRT statistic. Then, we propose two criteria which help evaluating the quality of the prediction.

Posterior estimates Figure 2 shows how the logarithm of the data $\mathbf{y}_1$ in the training set was duplicated by the latent variables $\mathbf{Z}_1$ (we take region 1 for an example). Their moving averages with past windows of 5 and 10 were respectively displayed. We can see that $\mathbf{Z}_1$ duplicated quite well $\mathbf{y}_1$, by keeping almost each instant trend and the mean value as well. As illustrated in Figure 1 that $\mathbf{Z}_1$ link the GLM to the classical linear model, where the latter model is completely based on $\mathbf{Z}_1$ instead of $\mathbf{y}_1$ as observation. Thus, the quality of the duplication of $\mathbf{y}_1$ is of extreme importance.

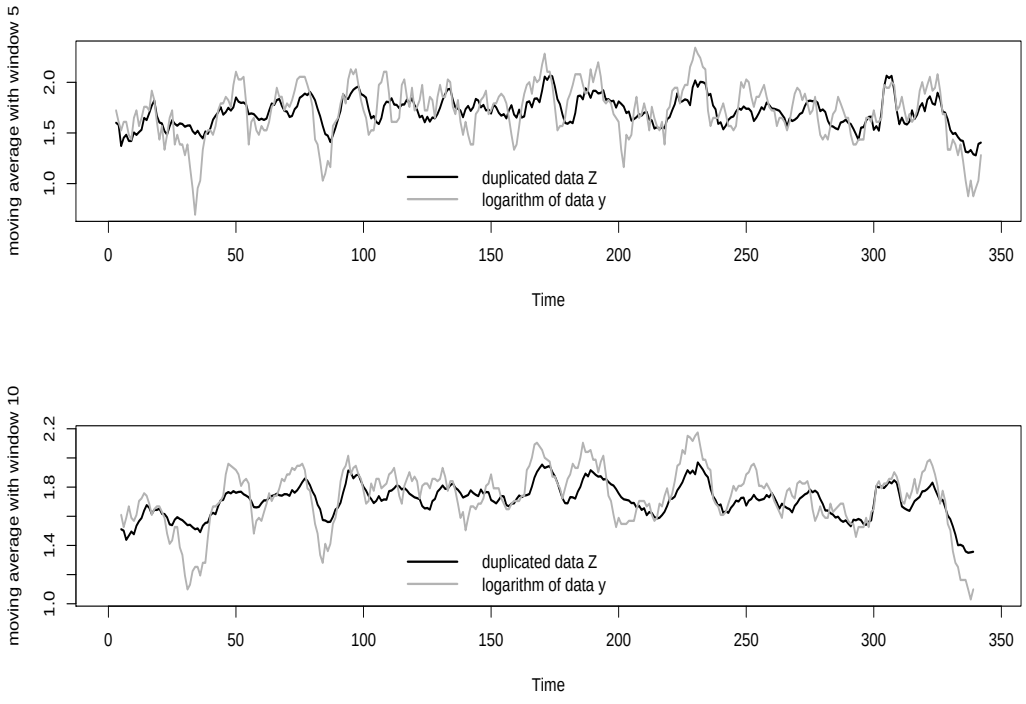

Figure 2: Moving average with past windows of 5 and 10 of latent variables $\mathbf{Z}_1$ which duplicate the logarithm of the true observed data $\mathbf{y}_1$ (for region 1).

Figure 3 shows posterior medians and 95% credible intervals for the NB regression coefficients using the Bayesian LASSO prior I without any shrinkage. For comparison,

the MLE and the Bayesian estimates are equally marked. We can see that the three methods returned close-to-zero estimates for most of the regression coefficients, but they gave quite different estimates for β_4, ie. the major daily traffic, which corresponded to the most influential covariate, . Note that we don't show more LASSO estimates because of their similarity before shrinkage.

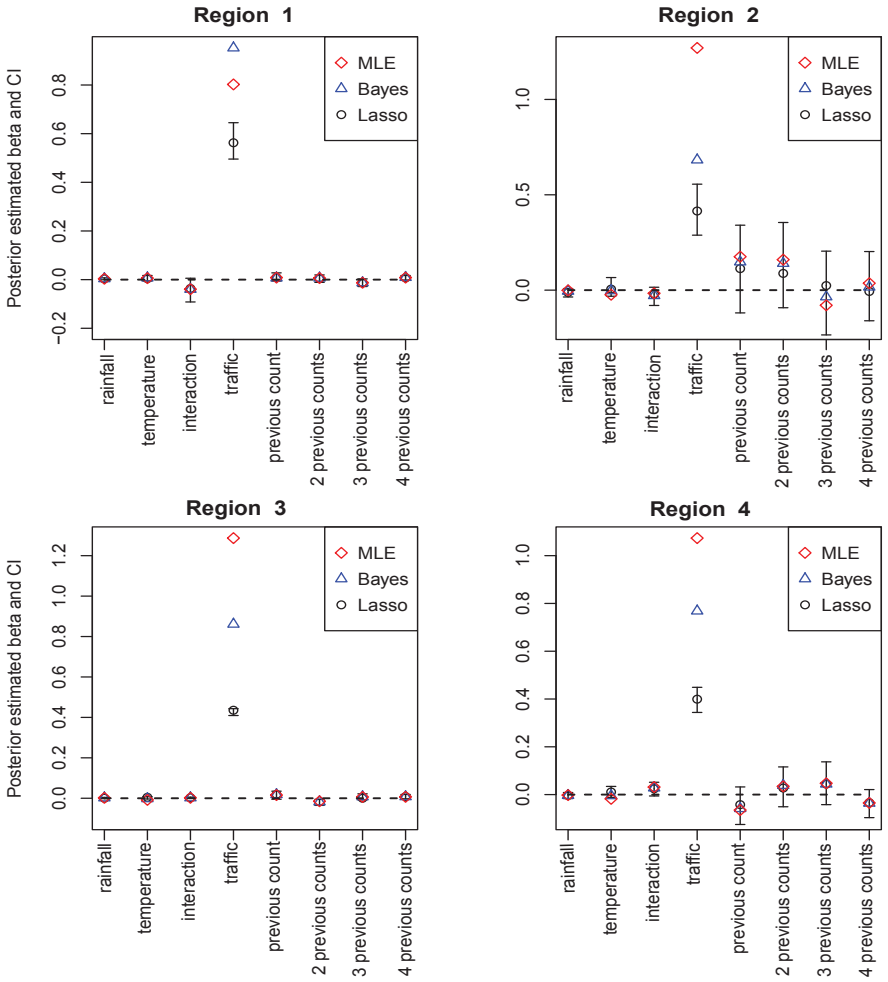

Figure 3: Posterior median of Bayesian LASSO estimates based on prior I and corresponding 95% credible interval. The least squares estimates and Bayesian estimates are equally marked.

Figures 4 and 5 explain how the variable selection was achieved with help of DLRT, for priors I,II,III and priors IV,V,VI respectively. With each prior, DLRT_{jk} was drawn with respect to the significant group size k changing from 1 to 8 for each region j. The 95% percentile curve of the corresponding χ^2_{p-k} distribution was added. We know that once the DLRT curve decreases under the percentile curve, adding more covariates brings no longer significant improvement but more modeling uncertainty. The optimal selected number of coefficients is thus the last point above the percentile curve. Comparing Figures 4 and 5, priors I,II,III forced stronger shrinkage as they got less points above the red curve and priors IV,V,VI admitted weaker shrinkage.

Figure 6 displays the histograms of the posterior samples of the LASSO parameters with the six combined priors for region 1 (we take region 1 as an example).

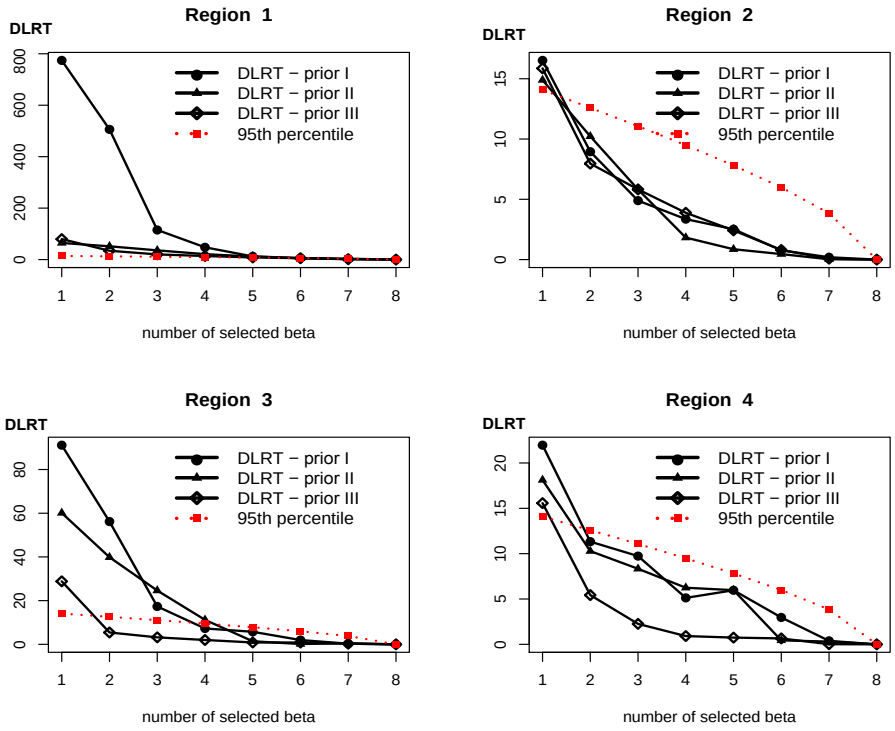

Figure 4: DLRT and the 95% percentile of χ^2_{p-k} distribution with respect to the number of selected coefficients $\boldsymbol{\beta}_j$, in Bayesian LASSO based on priors I, II and III.

More precisely, they regard the inverse scale parameter $\lambda^2/2$ for priors I and VI, the shape parameter $e/2$ and the scale parameter $d/2$ for priors II and V, the degrees of freedom ν and the scale parameter s^2 for priors III and VI. The top and bottom plots show inference for the prior distributions $\mathrm{N}(\beta; 0, \tau^2)$ and $\mathrm{N}(\beta; 0, \sigma^2\tau^2)$, which correspond respectively to priors I.II,III and priors IV,V,VI. We see that $\mathrm{N}(\beta; 0, \tau^2)$ returned a larger estimate of the inverse scale, a smaller estimate of the shape, scale and degrees of freedom, which induced stronger shrinkage (refer to Figure 4). Vice versa for $\mathrm{N}(\beta; 0, \sigma^2\tau^2)$ (refer to Figure 5).

Prediction evaluation Two criteria have been applied in our experiments to evaluate the performance of prediction for each method.

- Mean squared error (MSE). The MSE criterion measures the accuracy of prediction, defined as the expected value of the squared difference between the fitted data $\widehat{\mathbf{y}}_j = (\hat{y}_{j1}, \ldots, \hat{y}_{jm})$ of the test set and the true future observations $\mathbf{y}_j^{\mathrm{F}} = (y_{j1}^{\mathrm{F}}, \ldots, y_{jm}^{\mathrm{F}})$. For each region j, it can be easily estimated as

$$\widehat{\mathrm{MSE}}_j \quad = \quad \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_{ji} - y_{ji}^{\mathrm{F}})^2, \quad \text{with } \hat{y}_{ji} \; = \; \exp\left(\mathbf{x}_{ji}^{\mathrm{F}}\hat{\boldsymbol{\beta}}_j\right).$$

  In this expression, m is the test sample size, $\mathbf{x}_{ji}^{\mathrm{F}} = (x_{ji1}^{\mathrm{F}}, \ldots, x_{jip}^{\mathrm{F}})$ is the i-th future covariate vector for region j and $\hat{\boldsymbol{\beta}}_j$ denotes the regression coefficients estimated from the previous data $\mathbf{y}_j$. In the Bayesian LASSO approach, $\hat{\boldsymbol{\beta}}_j$ is either calculated by $\frac{1}{G}\sum_{g=1}^{G}\boldsymbol{\beta}_j^{(g)}$, with $\boldsymbol{\beta}_j^{(g)} \sim \pi(\cdot|\mathbf{y}_j)$ the posterior sample

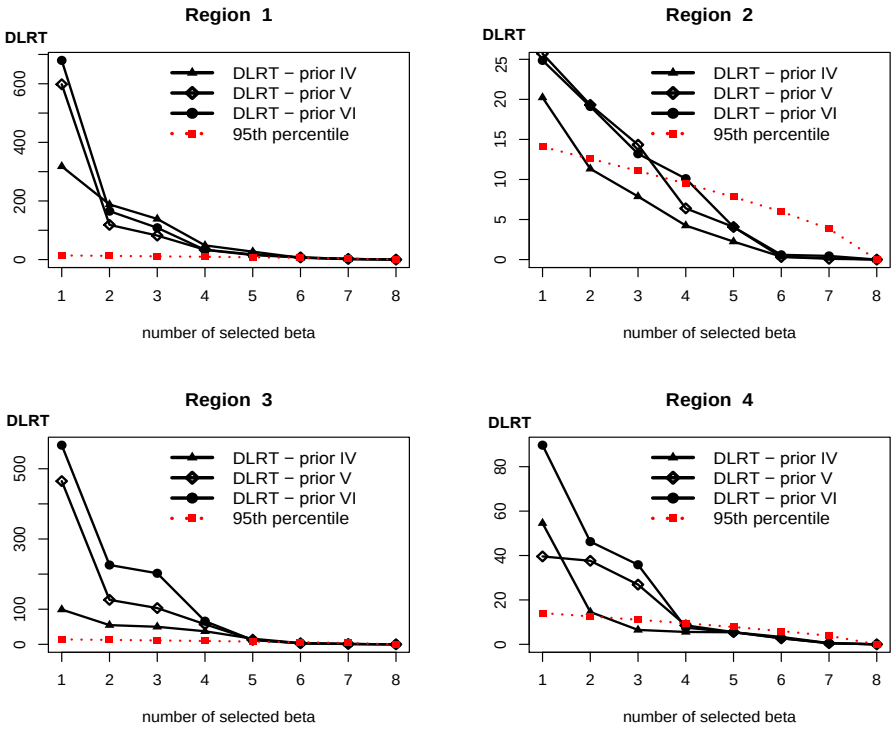

Figure 5: DLRT and the 95% percentile of χ^2_{p-k} distribution with respect to the number of selected coefficients $\boldsymbol{\beta}_j$, in Bayesian LASSO based on priors IV, V and VI.

resulting from the hybrid MCMC algorithm or shrinked to 0 if it is related to the irrelevant covariates.

- Logarithmic score (LS). LS is a strictly proper scoring rule, which measures the predictive accuracy at a probabilistic level. It can be expressed as the logarithm of the forecast probability that the event is realized. In Bayesian LASSO, let $\{\theta_j^{(1)}, \ldots, \theta_j^{(G)}\}$ be G quasi-i.i.d. posterior samples for region j with each $\theta_j^{(g)} = (\boldsymbol{\beta}_j^{(g)}, \psi_j^{(g)})$ and $\mathbf{y}_j^{\mathrm{F}} = (y_{j1}^{\mathrm{F}}, \ldots, y_{jm}^{\mathrm{F}})$ be the m future observations. Thus, for each region j,

$$\mathrm{LS}_j \;=\; \int \log \mathbb{P}_{\mathrm{NB}}(\mathbf{y}_j^{\mathrm{F}}|\theta_j)\, \pi(\theta_j|\mathbf{y}_j)\, d\theta \;\simeq\; \frac{1}{mG} \sum_{g=1}^{G} \sum_{i=1}^{m} \log \mathbb{P}_{\mathrm{NB}}(y_{ji}^{\mathrm{F}} \,|\, \theta_j^{(g)}),$$

where $\mathbb{P}_{\mathrm{NB}}(y_{ji}^{\mathrm{F}} \,|\, \theta_j^{(g)})$ is the probability of $\mathbf{y}_j^{\mathrm{F}}$ derived from $\mathrm{NB}\left(\exp(\mathbf{x}_j^{\mathrm{F}} \boldsymbol{\beta}_j^{(g)}),\ \psi^{(g)}\right)$ and $\mathbf{x}_j^{\mathrm{F}} = \begin{pmatrix} \mathbf{x}_{j1}^{\mathrm{F}} \\ \vdots \\ \mathbf{x}_{jm}^{\mathrm{F}} \end{pmatrix} = \begin{pmatrix} x_{j11}^{\mathrm{F}} & \cdots & x_{j1p}^{\mathrm{F}} \\ \vdots & \ldots & \vdots \\ x_{jm1}^{\mathrm{F}} & \cdots & x_{jmp}^{\mathrm{F}} \end{pmatrix}$ denotes m future covariate vectors.

10-fold cross validation To check if our model is over-fit, a 10-fold cross validation adapted to the time series structure was applied. Instead of randomly breaking the dataset into 10 partitions, we chose 10%, 20% until 90% of the dataset to fit the Bayesian model, then the prediction was made on the next 10% data. In other words, we increased the training sample size and kept the same test sample size. Figure 7 showed the predictive mean squared error (MSE) averaged on the 4 regions

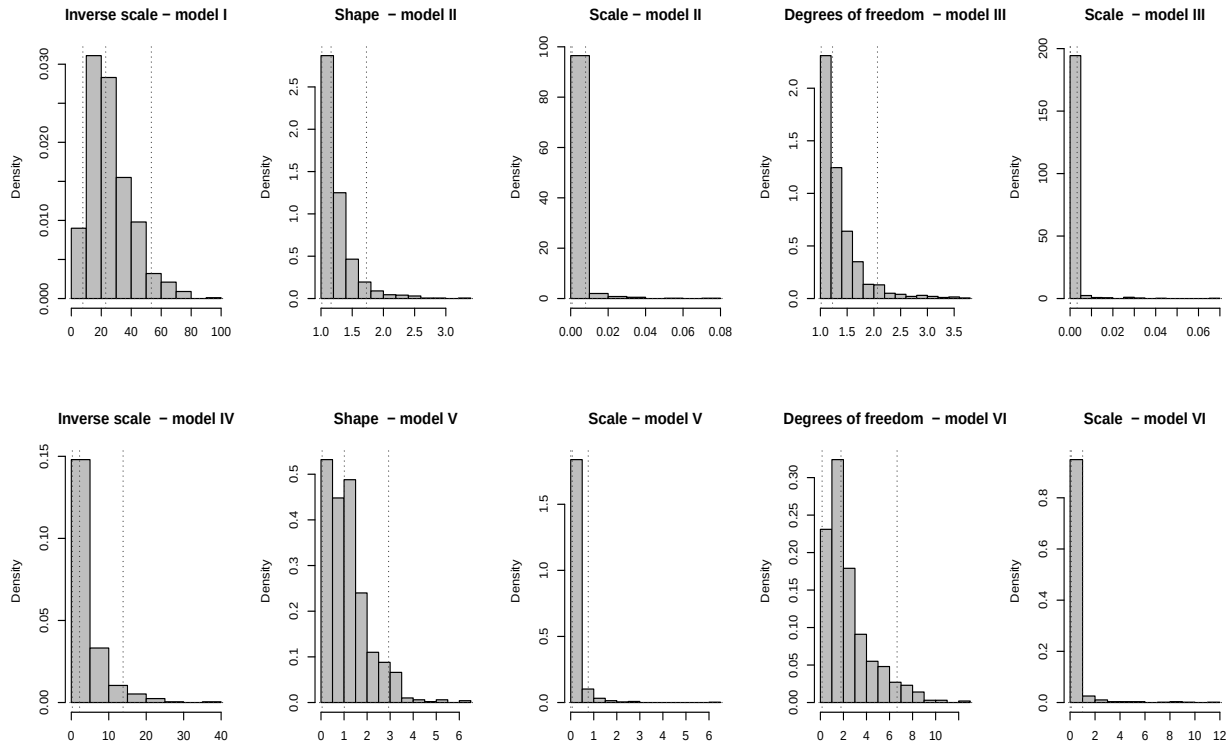

Figure 6: Histograms of the posterior samples for the LASSO parameters with different priors for region 1. The dotted lines denote the posterior 5%, 50% and 95% percentiles. The upper and lower plots correspond to the prior distributions (11) and (16) on $\boldsymbol{\beta}_1$, respectively.

applying the six different priors, with or without the variable selection procedure, respectively noted by Bayesian LASSO and the classical Bayesian approach. We found that Bayesian LASSO was much more stable than the classical Bayesian approach through removing the potentially dangerous covariates, which confirmed one conclusion of the paper. Moreover, there was a slight decreasing trend of the averaged MSE when the training sample size increased. This also confirmed the good property of the Bayesian approach, which works already well in a small sample size setting.

Accounting for prior information Assuming that we have additional prior information about the average crash counts as well as the relationship between the count data and covariates, for each region j we replace the flat prior for μ_j by a normal prior centered on the prior mean μ_{j0} with variance proportional to that of Z_{ji}, as $\mu_j \sim \mathrm{N}(\mu_{j0}, \sigma_j^2/a_0)$. The Jeffreys noninformative prior for σ_j^2 has been kept, which is $\sigma_j^2 \propto 1/\sigma_j^2$. Derived from this *partially* informative prior, the conditional posterior distributions of μ_j and σ_j^2 can be proven as follows

$$\begin{aligned}
\pi(\mu_j \,|\, \mathbf{Z}_j, \sigma_j^2, \boldsymbol{\beta}_j) &\propto \mathrm{N}\left(\frac{n}{n+a_0}(\bar{\mathbf{Z}}_{jn} - \bar{\mathbf{x}}_{jn}\boldsymbol{\beta}_j) + \frac{a_0}{n+a_0}\mu_{j0},\ \frac{\sigma_j^2}{n+a_0}\right), \\
\pi(\sigma_j^2 \,|\, \mathbf{Z}_j, \mu_j, \boldsymbol{\beta}_j) &\propto \text{SInv-}\chi^2\left(n+1,\ \frac{\sum_{i=1}^n (Z_{ji} - \mu_j - \mathbf{x}_{ji}\boldsymbol{\beta}_j)^2 + a_0(\mu_j - \mu_{j0})^2}{n+1}\right),
\end{aligned}$$

with $\bar{\mathbf{Z}}_{jn} = \sum_{i=1}^n Z_{ji}/n$ and $\bar{\mathbf{x}}_{jn} = \sum_{i=1}^n x_{ji}/n$. The conditional posterior mean of μ_j mixes the empirical value $\bar{\mathbf{Z}}_{jn} - \bar{\mathbf{x}}_{jn}\boldsymbol{\beta}_j$ and the prior value μ_{j0} with respective

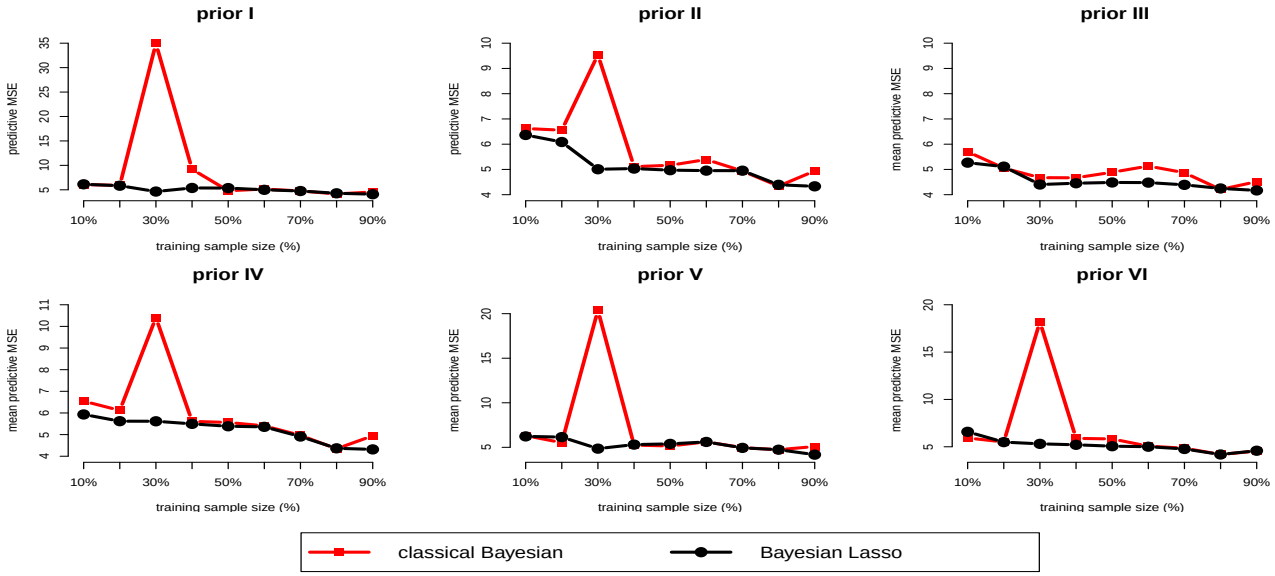

Figure 7: The predictive MSE averaged on 4 regions, based on 10-fold cross validation, applying the classical Bayesian approach and Bayesian LASSO approach with 6 different priors.

importance weights n and a_0. a_0 can thus be regarded as the size of a virtual sample to be adjusted with respect to our belief on the prior information. When a_0 is close to 0, the impact of the prior distribution disappears; when a_0 is large, there is no more impact of the data. Now we present how to choose μ_{j0} from the expert choice of the mean crash count $\bar{\mathbf{y}}_{j\text{Exp}}$ and the regression coefficients $\boldsymbol{\beta}_{j\text{Exp}}$ for each region j. As proven in Eq.(7) for $i = 1, \ldots, n$,

$$\log \mathbb{E}(y_{ji}) \quad \simeq \quad \mu_j + \mathbf{x}_{ji}\boldsymbol{\beta}_j \;=\; \mu_{j0} + \mathbf{x}_{ji}\boldsymbol{\beta}_{j\text{Exp}} + (\mu_j - \mu_{j0}) + \mathbf{x}_{ji}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j\text{Exp}}),$$

by taking expectation and averaging on both sides, we have

$$\log\left(\bar{\mathbf{y}}_{j\text{Exp}}\right) \quad \simeq \quad \mu_{j0} + \bar{\mathbf{x}}_{jn}\boldsymbol{\beta}_{j\text{Exp}}, \text{ with } \bar{\mathbf{x}}_{jn} = \frac{1}{n}\sum_{i=1}^{n}(x_{j1} + \cdots + x_{jp}).$$

We choose thus $\mu_{j0} \simeq \log\left(\bar{\mathbf{y}}_{j\text{Exp}}\right) - \bar{\mathbf{x}}_{jn}\boldsymbol{\beta}_{j\text{Exp}}$, which is completely determined by the expert knowledge and available data. In our experiments, we varies the virtual size a_0 to adjust the importance of the prior information. Three different values have been chosen: $a_0 = 0$ for a noninformative prior, $a_0 = 5$ for a highly informative prior and $a_0 = 0.1$ for a weakly informative prior.

Table 3 summarizes the predictive performances of the six proposed Bayesian LASSO methods by applying noninformative and informative priors, as well as the MLE and the standard Bayesian inference. In terms of both the MSE and LS, the Bayesian LASSO methods significantly improved the MLE and the Bayesian regression. The number of selected coefficients was precised for each LASSO model. Among the six LASSO priors, we notice that the first three priors allowed stronger shrinkage and with fewer selected coefficients, they achieved better accuracy. The other three priors made weaker shrinkage and the predictive accuracy was also reduced. Compared with the noninformative prior, informative priors permit to consider more covariates, ie. weaker shrinkage, and slightly improve the prediction accuracy. The weakly informative prior works better that the highly informative prior. The reason

Table 3: Comparison of the predictive performances of the six Bayesian LASSO methods, with noninformative prior(-N), highly informative prior (-H) and weakly informative prior (-W), the standard Bayesian inference and the maximum likelihood estimation, in terms of the number of selected $\boldsymbol{\beta}$, MSE and LS.

	Region 1			Region 2			Region 3			Region 4		
Methods	$\#\boldsymbol{\beta}$	MSE	LS	$\#\boldsymbol{\beta}$	MSE	LS	$\#\boldsymbol{\beta}$	MSE	LS	$\#\boldsymbol{\beta}$	MSE	LS
LASSO-N I	5	9.925	-2.544	1	0.463	-0.863	3	5.444	-2.317	1	1.974	-1.664
LASSO-H I	6	9.873	-2.542	1	0.466	-0.869	7	5.449	-2.308	1	1.973	-1.663
LASSO-W I	6	9.860	-2.540	1	0.462	-0.862	4	5.436	-2.301	1	1.973	-1.663
LASSO-N II	5	9.854	-2.540	1	0.465	-0.865	4	4.909	-2.272	1	1.964	-1.661
LASSO-H II	5	9.859	-2.550	1	0.469	-0.868	4	4.998	-2.276	1	1.962	-1.579
LASSO-W II	5	9.832	-2.533	2	0.463	-0.862	**4**	**4.385**	**-2.263**	**1**	**1.962**	**-1.576**
LASSO-N III	5	9.635	-2.532	1	0.462	-0.861	1	4.998	-2.281	1	1.972	-1.665
LASSO-H III	6	9.569	-2.497	1	0.459	-0.857	2	4.961	-2.274	1	1.972	-1.664
LASSO-W III	**5**	**9.542**	**-2.449**	**1**	**0.452**	**-0.852**	2	4.976	-2.276	1	1.970	-1.660
LASSO-N IV	6	11.034	-2.599	1	0.472	-0.878	5	5.827	-2.353	2	2.079	-1.701
LASSO-H IV	6	10.371	-2.585	4	0.467	-0.872	4	5.583	-2.351	5	2.084	-1.711
LASSO-W IV	6	10.132	-2.566	3	0.471	-0.874	4	5.482	-2.348	5	1.985	-1.682
LASSO-N V	6	10.719	-2.582	3	0.471	-0.876	5	5.558	-2.331	3	2.192	-1.750
LASSO-H V	6	10.320	-2.575	4	0.471	-0.876	5	5.169	-2.318	3	2.108	-1.733
LASSO-W V	6	10.356	-2.579	4	0.468	-0.873	5	5.260	-2.320	3	2.060	-1.724
LASSO-N VI	6	10.779	-2.585	4	0.468	-0.872	5	5.539	-2.329	3	2.168	-1.739
LASSO-H VI	6	10.280	-2.582	4	0.468	-0.872	5	5.275	-2.317	3	2.076	-1.728
LASSO-W VI	6	10.254	-2.573	4	0.465	-0.871	5	5.200	-2.312	3	2.049	-1.722
Bayesian	8	11.385	-2.844	8	20.484	-0.939	8	5.733	-2.361	8	29.348	-1.817
MLE	8	11.046	-2.824	8	0.502	-0.867	8	5.644	-2.350	8	100.838	-2.149

is twofold: the sample size is large enough in our experiments; the prior information from the expert knowledge is not perfect. Regarding the MLE and Bayesian inference, the improvement achieved by LASSO especially resided in regions 2 and 4, where the mean crash counts were close to 0 with many 0's as observations and the covariates were not homogeneous between the training and test sets (see Table 2). The MLE and the Bayesian inference made a large prediction error, while the Bayesian LASSO model wih all the six priors worked with robustness by removing some noninfluential but dangerous covariates. As marked in black, priors II and III were the preferred LASSO priors. More precisely, the priors III with weakly prior information was preferred for regions 1 and 2, the priors II with weakly prior information was preferred for regions 3 and 4, and both priors II and III permit strong shrinkage.

7 Discussion

In this paper we proposed one hierarchical Bayesian LASSO model with six different priors to address the NB regression problem. Latent variables $\mathbf{Z}$ have been introduced to transform the GLM to a standard linear regression model. The hierarchical structure, longitudinally allowed modeling the uncertainty of different probabilistic levels which derived the amount of shrinkage (with help of the DLRT statistic); laterally helped considering the dependence among different regions by assuming a common dispersion parameter and common hyperpriors. A full Bayesian analysis was provided by treating all the parameters and hyperparameters as unknowns and a hybrid MCMC algorithm has been constructed to simulate their posterior distributions, by taking into account the prior knowledge. Although computationally intense compared with the MLE, it provided not only point estimates but also posterior samples of all parameters. Statistical analysis has been carried on those samples and a comparison

of predictive performances was made with help of the MSE and LS.

Our experiments showed that for the NB regression problem, all the six Bayesian LASSO methods worked better than the MLE and the Bayesian inference in both accuracy and robustness. In terms of robustness, the MLE procedure sometimes had convergence concerns when treating with noisy and varied data. We got warning messages returned by the mathematical software. On the other hand, the Gibbs sampler always reached its convergence to the stationary posterior distribution, even though it could take a long time. On a Intel® Core^{TM} i7 processor 2.60GHz computer, it took about 3h to finish 60,000 iterations of the hybrid MCMC for all the four regions. Allowing an appropriate amount of shrinkage was especially helpful in the quite tricky situation with few observations and a large number of covariates, some of which are highly nonhomogeneous. In fact, we prefer removing such "dangerous" covariates than using them. It is also suggested in the intermittent series (there are many 0's as observations) or missing-data cases.

The variable selection procedure achieved very satisfying results with help of the DLRT statistic. Between the two priors $\mathrm{N}(0, \tau^2)$ and $\mathrm{N}(0, \sigma^2\tau^2)$ for $\boldsymbol{\beta}$, we notice that the first one worked better in our real-life experiment by allowing stronger shrinkage. The second one, although recommended by Park and Casella (2008) for its unimodal posterior property, did not work as well as the first unconditional one. Moreover, for a fixed prior for $\boldsymbol{\beta}$, the three prior choices for the variance τ^2 allowed different amounts of shrinkage. Their performances were similar and the Inverse-gamma and the Scaled Inverse-χ^2 distributions worked slightly better than the Exponential distribution. Besides, accounting for prior information is valuable in a small sample size setting. Instead of applying a flat prior for the constant parameter μ, the prior knowledge can be incorporated by setting up a normal distribution with variance proportional to σ^2, which assigns a high probability on the supposed-to-be value. The virtual size a_0 helps to adjust the important of the prior information. In our experiments, introducing prior information can slightly improve the predictive accuracy. A larger improvement is envisaged if the number of available data is reduced.

The Bayesian LASSO approach can be applied to various research themes. For instance, the quantitative trait loci mapping (Yi and Xu, 2008), the gene selection (Bae and Mallick, 2004), the infectious or viral diseases affection (Gentleman et *al.*, 1994), the pollutants propagations (Ang and Tang, 1984; Zhang and Dai, 2007), etc. If there are missing values among the data $\mathbf{y}$, two alternative solutions can be considered. The first method replaces the missing data by their expected values conditioning on the observed data (Haley and Knott, 1992). The uncertainty is thus ignored. The second method is through a MCMC data augmentation process modeling the whole uncertainty. Since it is computationally slower, one can consider accelerating the MCMC through an adaptive augmentation procedure (see Pasanisi et *al.*, 2012). For future research, the multiple perspectives are as follows.

- The introduction of latent variables may be extended to other kinds of regression problems, as demonstrated in Bae and Mallick (2004) for the probit regression problem.
- It may be helpful to replace the classical Gibbs sampler or at least a part of the Gibbs sampler by a posterior mode-searching algorithm, for example the conditional maximization (Zhang and Xu, 2005), the EM algorithm (Dempster et al., 1977) and its extension. In fact, with independent Laplace priors the LASSO estimates for linear regression coefficients are equivalent to Bayesian posterior mode estimates (Tibshirani, 1996). We only need its posterior mode rather than the whole posterior sample. An acceleration of the computation is

envisaged.

- If the multivariate variable $\mathbf{y}$ is highly correlated among different regions, we could consider applying the Negative Multinomial (NM) distribution instead of the NB distribution. In fact, the NM distribution holds the same mean and variance values as the NB distribution (thus the dispersion property) and it permits an explicit correlation among the sub-components of $\mathbf{y}$. More details about the NM distribution can be found in Appendix C.

References

[1] Ang, A.H.S. and Tang, W.H. (1984). *Probability Concepts in Engineering Planning and Design*, **2**, Wiley, NY.

[2] Bae, K. and Mallick, B.K. (2004). Gene selection using a two-level hierarchical Bayesian model, *Bioinformatics*, **20**, 3423-3430.

[3] Bickel, E.J. (2007). Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules, *Decision Analysis*, **4**(2), 49-65.

[4] Boone, E.L. and Merrick, J.R.W and Krachey, M.J. (2014). A Hellinger distance approach to MCMC diagnostics, *Journal of Statistical Computation and Simulation*, **84** (4), 833-849.

[5] Bousquet, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis, *Journal of Applied Statistics*, **35**, 1011-1029.

[6] Brooks, S.P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations, *Journal of Computational and Graphical Statistics*, **7**, 434-455.

[7] Carlin, B.P. and Louis, T.A. (2008). *Bayesian Methods for Data Analysis*, Chapman & Hall/CRC, 3rd edition.

[8] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Ser. B. **39**, 1-38.

[9] Fink, D. (1995). A Compendium of Conjugate Priors, In progress report: Extension and enhancement of methods for setting data quality objectives.

[10] Fu, S. (2014). A hierarchical Bayesian approach to negative binomial regression, *unpublished paper*.

[11] Fu, S., Celeux, G., Bousquet, N. and Couplet, M. (2015). Bayesian inference for inverse problems occurring in uncertainty analysis, *International Journal for Uncertainty Quantification*, **5** (1), 73-98.

[12] Garnero, M.A. and Montgomery, N. (2006). Pronostic de la profondeur de fissuration d'un rotor de turbine (in French). Proceedings of the lambda-mu 15th congress.

[13] Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis*, Chapman & Hall, London.

[14] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, Chapman & Hall/CRC, 2nd edition.

[15] Gelman, A. and Rubin, D. (1992). Inference from Iterative Simulation using Multiple Sequences, *Statistical Science*, **7**, 457-511.

[16] Gentleman, R.C., Lawless, J.F., Lindsey, J.C. and Yan, P. (1994). Multi-state Markov models for analyzing incomplete disease history data with illustration for HIV disease, *Statistics in Medicine* **13**, 805-821.

[17] Griffin, J.E. and Brown, P,J. (2006). Alternative prior distributions for variable selection with very many more variables than observations, Technical report, University of Warwick, Coventry, UK.

[18] Hoti, F. and Sillanpää, M.J. (2006). Bayesian mapping of genotype × expression interactions in quantitative and qualitative traits, *Heredity*, **97**, 4-18.

[19] Huelsenbeck, J. P. and Crandall, K. A. (1997). Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood, *Annual Review of Ecology and Systematics*, **28**, 437-466.

[20] Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models, *Journal of the Royal Statistical Society B*, **63**, 425-464.

[21] Lawless, J. F. (1987). Negative binomial and mixed Poisson regression, *The Canadian Journal of Statistics*, **15**(3), 209-225.

[22] Li, Y., Campbell, C. and Tipping, M. (2002). Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, **18**, 1332-1339.

[23] Long, J.S. (1997). Regression Models for Categorical and Limited Dependent Variables, *Advanced Quantitative Techniques in the Social Sciences*, **7**.

[24] McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models*, Chapman & Hall/CRC, 2nd edition.

[25] Neal, R.M. (2003). Slice Sampling, *Annals of Statistics* **31**(3), 705-767.

[26] Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models, *Journal of the Royal Statistical Society*, Series A (General) (Blackwell Publishing) **135**(3), 370-384.

[27] Park, T. and Casella, G, (2008). The Bayesian Lasso, *Journal of the American Statistical Association*, **103**(482).

[28] Pasanisi, A., Fu, S. and Bousquet, N. (2012). Estimating discrete Markov models from various incomplete data schemes, *Computational Statistics & Data Analysis*, **56**(9), 2609-2625.

[29] Pillow, J.W. and Scott, J.G. (2012). Fully Bayesian inference for neural models with negative-binomial spiking, *Neural information processing systems*, **25**, 1907-1915.

[30] Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., and Lunn, D. (1994, 2003). BUGS: Bayesian inference using Gibbs sampling, *MRC Biostatistics Unit*, Cambridge, England.

[31] Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, Springer, 2nd edition.

[32] Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*, Springer, NY.

[33] Tibshirani, R. (1996). Regression shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, Ser. B, **58**, 267-288.

[34] Tierney, L. (1994). Markov chains for exploring posterior distributions, *Ann. Statist.*, **22**(4), 1701-1762.

[35] Tierney, L. (1995), Introduction to general state-space Markov chain theory, *Markov Chain Monte Carlo in Practice*, Chapman & Hall.

[36] Williams, P. (1995), Bayesian regularization and pruning using a Laplace prior, *Neural Comput.*, **7**, 117-143.

[37] Yi, N. and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping, *Genetics*, **179**, 1045-1055.

[38] Zhang, L. and Dai, S. (2007). Application of Markov model to environmental fate of phenanthrene in Lanzhou reach of Yellow river, *Chemosphere* 67, 1296-1299.

[39] Zhou, M., Li, L., Dunson, D. and Carin, L. (2012). Lognormal and gamma mixed negative binomial regression, Proceedings of the 29th International conference on machine learning, Scotland, UK.

[40] Zhang, Y.M. and Xu, S. (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL, *Heredity*, **95**, 96-104.

A Computation of conditional posterior distributions

The conditional posterior distributions for the regression parameters and hyperparameters in the standard linear regression are computed as follows.

$$\begin{aligned}
\pi(\mu \,|\, \mathbf{Z}, \sigma^2, \boldsymbol{\beta}) &\propto \pi(\mathbf{Z} \,|\, \mu, \sigma^2, \boldsymbol{\beta}) \cdot \pi(\mu) \;=\; \prod_{i=1}^{n} \mathrm{N}(\mu + \mathbf{x}_i\boldsymbol{\beta},\, \sigma^2) \cdot 1 \\
&\propto \exp\left[-\sum_{i=1}^{n} \frac{(Z_i - \mu - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2} \right] \\
&\sim \mathrm{N}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbf{x}_i\boldsymbol{\beta}),\;\; \frac{1}{n}\sigma^2 \right) \\
\pi(\sigma^2 \,|\, \mathbf{Z}, \mu, \boldsymbol{\beta}) &\propto \pi(\mathbf{Z} \,|\, \mu, \sigma^2, \boldsymbol{\beta}) \cdot \pi(\sigma^2) \;=\; \prod_{i=1}^{n} \mathrm{N}(\mu + \mathbf{x}_i\boldsymbol{\beta},\, \sigma^2) \cdot \frac{1}{\sigma^2} \\
&\propto \frac{1}{\sigma^n} \exp\left[-\sum_{i=1}^{n} \frac{(Z_i - \mu - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2} \right] \cdot \frac{1}{\sigma^2} \\
&\sim \text{SInv-}\chi^2\left(n,\; \frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu - \mathbf{x}_i\boldsymbol{\beta})^2 \right) \\
\pi(\boldsymbol{\beta} \,|\, \mathbf{Z}, \mu, \sigma^2) &\propto \pi(\mathbf{Z} \,|\, \mu, \sigma^2, \boldsymbol{\beta}) \cdot \pi(\boldsymbol{\beta}) \;=\; \prod_{i=1}^{n} \exp\left[-\frac{(Z_i - \mu - x_{ij}\beta_j)^2}{2\sigma^2} \right] \cdot \exp\left[-\frac{\beta_j^2}{2V_j^2} \right] \\
&\sim \prod_{i=1}^{n} \mathrm{N}\left(\frac{\sum_{i=1}^{n} x_{ij}(Z_i - \mu)}{\sum_{i=1}^{n} x_{ij}^2 + \sigma^2/V_j^2},\;\; \frac{\sigma^2}{\sum_{i=1}^{n} x_{ij}^2 + \sigma^2/V_j^2} \right)
\end{aligned}$$

with $V_j^2 = \tau_j^2$ for priors I,II,III and $V_j^2 = \sigma^2\tau_j^2$ for priors IV,V,VI. The conditional posterior distribution of $\boldsymbol{\beta}$ can also be written under the multivariate form

$$\begin{aligned}
\pi(\boldsymbol{\beta} \,|\, \mathbf{Z}, \mu, \sigma^2) &\sim \mathrm{N}\left(A^{-1}\mathbf{x}\,(\mathbf{Z} - \mu\mathbf{1}_n),\;\; \sigma^2 A^{-1} \right) \\
\text{with } A &= \begin{cases} \mathbf{x}^T\mathbf{x} + \sigma^2 \mathrm{Diag}^{-1}(\tau_1^2, \ldots, \tau_p^2), & \text{for priors I,II,III;} \\ \mathbf{x}^T\mathbf{x} + \mathrm{Diag}^{-1}(\tau_1^2, \ldots, \tau_p^2), & \text{for priors IV,V,VI.} \end{cases}
\end{aligned}$$

With the Exponential prior for τ_j^2, the conditional posterior distribution can be calculated as

$$\begin{aligned}
\pi(\tau_j^{-2} \,|\, \mathbf{Z}, \beta_j, \sigma^2, \lambda^2) &\propto \pi(\beta_j \,|\, \tau_j^2) \cdot \pi(\tau_j^{-2} \,|\, \lambda^2) \;=\; \frac{1}{\tau_j} \exp\left[-\frac{\gamma_j^2}{2\tau_j^2} \right] \cdot \tau_j^4 \exp\left[-\frac{\lambda^2}{2}\tau_j^2 \right] \\
&\sim \text{InvGauss}\left(\frac{\sqrt{\lambda^2}}{|\gamma_j|},\;\; \lambda^2 \right),
\end{aligned}$$

with $\gamma_j = \beta_j$ for prior I and $\gamma_j = \beta_j/\sigma$ for prior IV. With the Inverse-Gamma prior for τ_j^2, its conditional posterior distribution is

$$\begin{aligned}
\pi(\tau_j^2 \,|\, \mathbf{Z}, \beta_j, \sigma^2, e, d) &\propto \pi(\beta_j \,|\, \tau_j^2) \cdot \pi(\tau_j^2 \,|\, e, d) \;=\; \frac{1}{\tau_j} \exp\left[-\frac{\gamma_j^2}{2\tau_j^2} \right] \exp\left[-\frac{d}{2\tau_j^2} \right] \left(\frac{1}{\tau_j^2} \right)^{e/2+1} \\
&\sim \text{InvGamma}\left(\frac{e+1}{2},\;\; \frac{\gamma_j^2 + d}{2} \right),
\end{aligned}$$

with $\gamma_j^2 = \beta_j^2$ for prior II and $\gamma_j^2 = \beta_j^2/\sigma^2$ for prior V. With the Scaled Inverse-χ^2 prior for τ_j^2, its conditional posterior distribution is

$$\begin{aligned}
\pi(\tau_j^2 \,|\, \mathbf{Z}, \beta_j, \sigma^2, \nu, s^2) &\propto \pi(\beta_j \,|\, \tau_j^2) \cdot \pi(\tau_j^2 \,|\, \nu, s^2) = \frac{1}{\tau_j} \exp\left[-\frac{\gamma_j^2}{2\tau_j^2}\right] \exp\left[-\frac{\nu s^2}{2\tau_j^2}\right] \frac{1}{\tau_j^{2+\nu}} \\
&\sim \text{SInv-}\chi^2\left(\nu+1,\ \frac{\gamma_j^2+\nu s^2}{\nu+1}\right),
\end{aligned}$$

with $\gamma_j^2 = \beta_j^2$ for prior III and $\gamma_j^2 = \beta_j^2/\sigma^2$ for prior VI. Regarding the hyperparameters,

$$\begin{aligned}
\pi(\lambda^2 \,|\, \mathbf{Z}, \mu, \boldsymbol{\beta}, \sigma^2, \tau^2, a, b) &\propto \prod_{j=1}^{p} \pi(\tau_j^2 \,|\, \lambda^2) \cdot \pi(\lambda^2 \,|\, a, b) \\
&\propto \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left[-\frac{\lambda^2}{2}\tau_j^2\right] \cdot \lambda^{2(a-1)} \exp\left[-b\lambda^2\right] \\
&\sim \text{G}\left(p+a,\ \sum_{j=1}^{p} \frac{\tau_j^2}{2} + b\right), \\
\pi(d \,|\, \mathbf{Z}, \mu, \boldsymbol{\beta}, \sigma^2, \tau^2, e) &\propto \prod_{j=1}^{p} \pi(\tau_j^2 \,|\, e, d) \cdot \pi(d) \propto \prod_{j=1}^{p} \left(\frac{d}{2}\right)^{e/2} \exp\left[-\frac{d}{2\tau_j^2}\right] \\
&\sim \text{G}\left(\frac{ep}{2}+1,\ \frac{1}{2}\sum_{j=1}^{p} \frac{1}{\tau_j^2}\right), \\
\pi(s^2 \,|\, \mathbf{Z}, \mu, \boldsymbol{\beta}, \sigma^2, \tau^2, \nu) &\propto \prod_{j=1}^{p} \pi(\tau_j^2 \,|\, \nu, s^2) \cdot \pi(\nu) \propto \prod_{j=1}^{p} \left(\frac{s^2\nu}{2}\right)^{\nu/2} \exp\left[-\frac{\nu s^2}{2\tau_j^2}\right] \\
&\sim \text{G}\left(\frac{\nu p}{2}+1,\ \frac{\nu}{2}\sum_{j=1}^{p} \frac{1}{\tau_j^2}\right).
\end{aligned}$$

The conditional posterior distributions of the degrees of freedom ν and the shape parameter e have no standard form. Thus, a Metropolis-Hastings step is required for their simulations.

B Are the posterior distributions proper ?

The posterior distributions of μ and σ^2 can be proven to be proper. The proof is as follows.

$$\begin{aligned}
\int \pi(\mu|\mathbf{y})d\mu &\propto \int_{\mathbb{R}}\int_{\mathbb{R}_+^n} \pi(\mathbf{y}|\mathbf{Z})\pi(\mathbf{Z}|\mu)\pi(\mu)d\mathbf{Z}\ d\mu \\
&\propto \int_{\mathbb{R}}\int_{\mathbb{R}_+^n} \prod_{i=1}^n \frac{\exp(Z_i)^{y_i}}{(\exp(Z_i)+\psi)^{y_i+\psi}} \exp\left[-\frac{(Z_i-\mu-\mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}\right] d\mathbf{Z}\ d\mu \\
&\propto \int_{\mathbb{R}_+^n} \prod_{i=1}^n \frac{\exp(Z_i)^{y_i}}{(\exp(Z_i)+\psi)^{y_i+\psi}} d\mathbf{Z} \int_{\mathbb{R}} \exp\left[-\sum_{i=1}^n \frac{(\mu-Z_i+\mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}\right] d\mu \\
&\propto \int_{\mathbb{R}_+^n} \prod_{i=1}^n \left(\frac{\exp(Z_i)}{(\exp(Z_i)+\psi)}\right)^{y_i} \left(\frac{1}{\exp(Z_i)+\psi}\right)^{\psi} d\mathbf{Z} \\
&\leq \prod_{i=1}^n \int_{\mathbb{R}_+} \left(\frac{1}{\exp(Z_i)+\psi}\right)^{\psi} dZ_i \\
&< \infty.
\end{aligned}$$

$$\begin{aligned}
\int \pi(\sigma^2|\mathbf{y})d\sigma^2 &\propto \int_{\mathbb{R}_+}\int_{\mathbb{R}_+^n} \pi(\mathbf{y}|\mathbf{Z})\pi(\mathbf{Z}|\sigma^2)\pi(\sigma^2)d\mathbf{Z}\ d\sigma^2 \\
&\propto \int_{\mathbb{R}_+}\int_{\mathbb{R}_+^n} \prod_{i=1}^n \left[\frac{\exp(Z_i)^{y_i}}{(\exp(Z_i)+\psi)^{y_i+\psi}} \frac{1}{\sigma} \exp\left[-\frac{(Z_i-\mu-\mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}\right]\right] \frac{1}{\sigma^2} d\mathbf{Z}\ d\sigma^2 \\
&\propto \int_{\mathbb{R}_+^n} \prod_{i=1}^n \frac{\exp(Z_i)^{y_i}}{(\exp(Z_i)+\psi)^{y_i+\psi}} d\mathbf{Z} \int_{\mathbb{R}_+} \frac{1}{\sigma^n} \exp\left[-\sum_{i=1}^n \frac{(Z_i-\mu-\mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}\right] \cdot \frac{1}{\sigma^2} d\sigma^2 \\
&\propto \int_{\mathbb{R}_+^n} \prod_{i=1}^n \left(\frac{\exp(Z_i)}{(\exp(Z_i)+\psi)}\right)^{y_i} \left(\frac{1}{\exp(Z_i)+\psi}\right)^{\psi} d\mathbf{Z} \\
&\leq \prod_{i=1}^n \int_{\mathbb{R}_+} \left(\frac{1}{\exp(Z_i)+\psi}\right)^{\psi} dZ_i \\
&< \infty.
\end{aligned}$$

This proof applied the fact that the conditional posterior distribution of μ is a Normal distribution and the conditional posterior distribution of σ^2 is a Scaled Inverse-χ^2 distribution.

C Negative Multinomial distribution

The i-th observation $\mathbf{y}_i = (y_{i1}, \ldots, y_{iR})^T$ consisting of R subsets follows the Negative Multinomial (NM) distribution

$$\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}, \psi \sim \text{NM}\left((\lambda_i^1, \ldots, \lambda_i^R)^T, \psi\right), \text{ with } \lambda_i^r = \exp(\mathbf{x}_i^r\beta_r)$$

where $\mathbf{x}_i = (\mathbf{x}_i^1, \ldots, \mathbf{x}_i^R)$ with $\mathbf{x}_i^r = (x_{i1}^r, \ldots, x_{ip}^r)$ denotes p covariates in the R subsets, $\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_R\end{pmatrix} = \begin{pmatrix}\beta_{11} & \cdots & \beta_{R1} \\ \vdots & \ldots & \vdots \\ \beta_{1p} & \cdots & \beta_{Rp}\end{pmatrix}$, and ψ denotes the common dispersion parameter. The probability density is

$$\pi\left(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}, \psi\right) = \frac{\Gamma(\sum_{r=1}^R y_{ir} + \psi)}{\Gamma(\psi)} \left[\frac{\psi}{\sum_{r=1}^R \lambda_i^r + \psi}\right]^{\psi} \prod_{r=1}^R \frac{1}{y_{ir}!} \left[\frac{\lambda_i^r}{\sum_{r=1}^R \lambda_i^r + \psi}\right]^{y_{ir}},$$

which indicates

$$\begin{aligned}
\mathbb{E}(y_{ir}|\mathbf{x}_i, \boldsymbol{\beta}, \psi) &= \lambda_i^r, \\
\mathrm{Var}(y_{ir}|\mathbf{x}_i, \boldsymbol{\beta}, \psi) &= \lambda_i^r(1 + \frac{\lambda_i^r}{\psi}), \\
\mathrm{Cov}(y_{ir}, y_{is}|\mathbf{x}_i, \boldsymbol{\beta}, \psi) &= \frac{\lambda_i^r \lambda_i^s}{\psi} \ (r \neq s).
\end{aligned}$$

The NB distribution is a generalization of the NB distribution. It keeps the dispersion property, ie. $\mathrm{Var}(y_{ir}|\mathbf{x}_i, \boldsymbol{\beta}, \psi) > \mathbb{E}(y_{ir}|\mathbf{x}_i, \boldsymbol{\beta}, \psi)$, and the correlation among the components of the count data $\mathbf{y}_i$ can be explicitly taken into account.