# A hierarchical Bayesian approach to negative binomial regression

Shuai Fu

*Dalle Molle Institute for Artificial Intelligence, SUPSI, Switzerland*

January 7, 2016

**Abstract**

There is a growing interest in establishing the relationship between the count data $\mathbf{y}$ and numerous covariates $\mathbf{x}$ through a generalized linear model (GLM), such as explaining the road crash counts from the geometry and environmental factors. This paper proposes a hierarchical Bayesian method to deal with the negative binomial GLM. The Negative Binomial distribution is preferred for modeling nonnegative overdispersed data. The Bayesian inference is chosen to account for prior expert knowledge on regression coefficients in a small sample size setting and the hierarchical structure allows to consider the dependence among the subsets. A Metropolis-Hastings-within-Gibbs algorithm is used to compute the posterior distribution of the parameters of interest through a data augmentation process. The Bayesian approach highly over-performs the classical maximum likelihood estimation in terms of goodness of fit, especially when the sample size decreases and the model complexity increases. Their respective performances have been examined in both the simulated and real-life case studies.

**Keywords.** Hierarchical Bayesian inference, Prior elicitation, Generalized linear regression, Negative binomial, Markov chain Monte Carlo.

## 1 Introduction

Considerable research has been carried out to explain the relationship between the count data $\mathbf{y}$ and different covariates $\mathbf{x} = [1, x^1, \ldots, x^P]^T$ through a generalized linear model (GLM) (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989)

$$\mathbb{E}(\mathbf{y}|\mathbf{x}) \quad = \quad g^{-1}(\mathbf{x}^T\beta), \tag{1}$$

where $g$ denotes the canonical link function and $\beta = [\beta^0, \beta^1, \ldots, \beta^P]^T$ denotes the vector of regression coefficients. The GLM generalizes the linear regression since the link function $g$ may allow an elaborate nonlinear relationship between $\mathbf{y}$ and $\mathbf{x}$. Especially, some constraints such as the nonnegativity and discreteness can be set on $\mathbf{y}$ with help of $g$. This regression model has wide applications, among which our current model is a typical one. It regards the analysis of regional crash counts with respect to covariates such as the traffic flow, the geometric design variables and environmental conditions. Due to the specific properties of counts, including discreteness, nonnegativity and overdispersion, $\mathbf{y}$ is assumed to follow the Negative Binomial (NB) distribution through the logarithmic link function. It can be specified as $\mathbf{y} \sim \mathcal{NB}\left(\lambda = \exp(\mathbf{x}^T\beta),\ \psi\right)$. Let note that a NB distribution with mean parameter

$\lambda$ and dispersion parameter $\psi$ is equivalent to a Poisson distribution with mean parameter $\gamma$, where $\gamma$ follows a Gamma distribution with shape parameter $\psi$ and rate parameter $\psi/\lambda$. Therefore, the NB distribution is also known as the Gamma-Poisson distribution. It allows an overdispersion property, i.e. the variance exceeds the mean, which favors the NB distribution over the Poisson distribution for modeling overdispersed counts. This often happens in practice. For instance, the average daily road accidents in Canton Ticino (Switzerland) in 2007 is 16.57, while its variance is 26.91 and so considerably exceeding the mean value. Another example comes from the stock management, where the recorded daily sale of a clothing item is 159, averaged on 2923 observations summed in 10 retail scores of Switzerland. Its variance is huge: $3.8e05$. In such overdispersed situations, the NB distribution is preferred whereas the Poisson distribution could introduce a large modeling error.

The frequentist approach is often considered the classical solution to the GLM, where the regression coefficients $\beta$ are usually estimated by maximizing the nonlinear log likelihood. The Newton-Raphson method can be applied to iteratively find the maximum likelihood estimator (MLE) of $\beta$ (Long, 1997). Many statistical computing packages are available for computing the MLE, such as the MASS package in R and the STATISTICS toolbox in Matlab. However, the MLE provides only a point estimate which may not be robust, or even fails to converge when the sample size is small or when the dispersion parameter is much larger than the mean. Moreover, it does not allow the consideration of prior information which may be helpful in case of lacking observations.

As an alternative, the Bayesian inference can account for prior expert knowledge on variables of interest, especially in a small sample size setting and it provides a sample of estimators which may be helpful for the uncertainty analysis. However, in the GLM, due to the complexity of the posterior distribution computation and thus lack of an efficient algorithm, the Bayesian approach has been much less developed than the MLE method. In this paper, a hierarchical Bayesian regression model has been constructed where three levels of variables have been considered: the data model, the process variable model and the parameter model (see Figure 1 to get a general idea). This hierarchical framework incorporates the whole modeling and estimation uncertainty and helps adjusting the dependence among the variables of interest in different subsets.

To compute the posterior distribution of regression coefficients, a hybrid MCMC algorithm has been proposed, namely the Gibbs sampler combined with a Metropolis-Hastings (MH) algorithm (*Metropolis-Hastings-within-Gibbs algorithm*) (see Robert and Casella, 2004). The embedded MH algorithm permits us to simulate the unknown full conditional posterior distribution through a Markov chain. The choice of the instrumental distribution is essential. Several alternatives have been numerically compared and the well fitting one in terms of accuracy and efficiency has been chosen. An original construction of the instrumental variance has been set up. The convergence of the Markov chain is accelerated, which can be checked with help of the Brooks-Gelman statistic (Brooks and Gelman, 1998).

It is worth noting that in some recent papers, a "closed-form" Bayesian inference for the regression coefficients $\beta$ has been proposed either by introducing an additional latent variable $\omega$ which follows a Polya-Gamma distribution (see for instance Pillow and Scott, 2012), or by mixing the NB model with the Lognormal-Gamma distribution (see for instance Zhou et *al.*, 2012). However, in the first case, high uncertainty has also be introduced as the prior distribution of $\omega$ is truly difficult to calibrate due to the complexity of the Polya-Gamma distribution; in the second case, the original

GLM degenerates. For instance, $\exp(\mathbf{x}^T\beta)$ no longer corresponds to the mean of $\mathbf{y}$ but involves the dispersion parameter $\psi$ and the variance parameter of the Lognormal distribution. This would mix several uncertainties to the prediction of $\mathbf{y}$. More importantly, in both cases, the calculation of some parameters in this closed-form posterior distribution is quite time-consuming. Our hybrid MCMC algorithm, however, works very efficiently in practice while keeping all advantages of the GLM. In addition, the hierarchical structure allows a more flexible choice of prior to fit the situation according to the requirements. Our numerical experiments support this proposition.

# 2 Bayesian inference

In regression analysis of counts, our major objective is to explain the relationship between the count data and covariates. Based on the estimated regression coefficients, the future counts can be replicated (or rather predicted) from the explanatory covariates. As explained before, the negative binomial GLM via the link function $g(\cdot) = \log(\cdot) = \exp^{-1}(\cdot)$ has been chosen as the regression model. For simplicity purposes, we explain our Bayesian context on the example of regional crash counts.

## 2.1 Modeling

**Statistical model** Let $Y_{ij}$ be the variable of accident counts occurring during period $i$ in given region $j$ with $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Let $\mathbf{x}_{ij} = [1, x_{ij}^1, \ldots, x_{ij}^P]^T$ be the corresponding covariate vector, $\beta_j = [\beta_j^0, \beta_j^1, \ldots, \beta_j^P]^T$ be the unknown regression coefficients and $\psi_j$ be the dispersion parameter of the underlying NB distribution. Given $(\mathbf{x}_{ij}, \beta_j, \psi_j)$, $Y_{ij}$ is assumed to follow the NB distribution, expressed as

$$Y_{ij}|\mathbf{x}_{ij}, \beta_j, \psi_j \sim \mathcal{NB}(\lambda_{ij}, \psi_j), \ \lambda_{ij} = \exp(\mathbf{x}_{ij}^T\beta_j),$$

with the probability density

$$f(y_{ij}|\mathbf{x}_{ij}, \beta_j, \psi_j) = \frac{\Gamma(y_{ij}+\psi_j)}{y_{ij}!\,\Gamma(\psi_j)}\left[\frac{\lambda_{ij}}{\lambda_{ij}+\psi_j}\right]^{y_{ij}}\left[\frac{\psi_j}{\lambda_{ij}+\psi_j}\right]^{\psi_j}.$$

We have $\mathbb{E}(Y_{ij}|\mathbf{x}_{ij}, \beta_j, \psi_j) = \lambda_{ij}$ and $\mathrm{Var}(Y_{ij}|\mathbf{x}_{ij}, \beta_j, \psi_j) = \lambda_{ij}(1 + \lambda_{ij}/\psi_j)$. It is coherent with the GLM rule that $\mathbb{E}(Y_{ij}|\mathbf{x}_{ij}, \beta_j, \psi_j) = \exp(\mathbf{x}_{ij}^T\beta_j) = g^{-1}(\mathbf{x}_{ij}^T\beta_j)$ and the variance exceeds the mean since $\psi_j > 0$. Decreasing the dispersion parameter $\psi_j$ towards 0 corresponds to increasing the overdispersion effect; increasing $\psi_j$ towards $+\infty$ leads to the convergence towards the Poisson distribution $\mathcal{P}(\lambda_{ij})$. This NB distribution is thus widely applied to the regression analysis of counts which appear overdispersed. Our statistical objective is to estimate the parameters $\theta_j = (\beta_j, \psi_j)$.

**Hierarchical prior choices** Assuming $J = 4$, the hierarchical Bayesian model can be described by the directed acyclic graph (DAG) shown in Figure 1. It illustrates the three probabilistic levels: the data model, the process model and the parameter model. The data model corresponds to observations $\mathbf{y}_j = \{y_{1j}, \ldots, y_{nj}\}$, the process model regards the parameters $\theta_j$ of the NB distribution and the parameter model focuses on setting a probabilistic distribution on the hyperparameters $\gamma_j$ of the parameters $\theta_j$, namely $\pi(\gamma_j|\rho)$. Each region has its own observations and random variables, but they share the common hyperparameters $\rho$ to reflect the dependence among them.

The first task is to choose a prior distribution $\pi(\cdot)$ for the process vector $\theta_j = (\beta_j, \psi_j)$. The Normal distribution seems a natural prior choice for $\beta_j$, the embedded linear regression coefficients (see Gelman et *al.*, 2003; Carlin and Louis 2008). $\psi_j$ is typically assumed to follow a Gamma distribution because of its positive nature and one can easily adjust the dispersion degree through the shape and scale parameters of

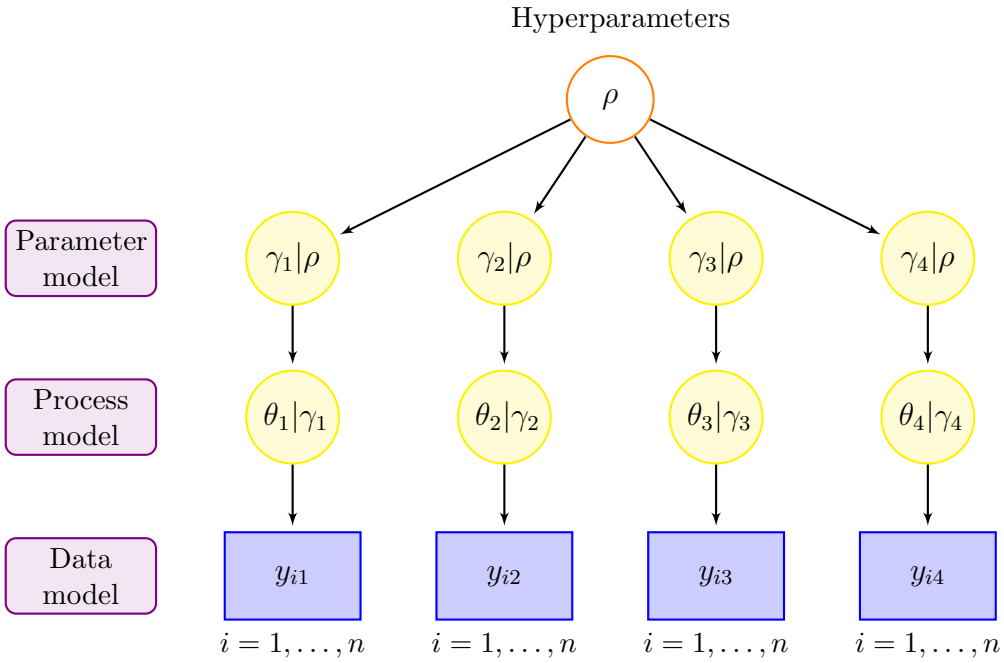


**Figure 1:** Directed acyclic graph (DAG) of the hierarchical Bayesian model ($J = 4$).

the Gamma distribution. Thus, we express their prior distributions in the following way:

$$\begin{aligned} \beta_j|m_j, \Sigma_j &\sim \mathcal{N}_{P+1}(m_j, \Sigma_j), && (2) \\ \psi_j|a_j, b_j &\sim \mathcal{G}(a_j, b_j), && (3) \end{aligned}$$

where $\mathcal{N}_{P+1}$ denotes a multivariate Normal distribution of dimension $P+1$ and $\mathcal{G}$ denotes a Gamma distribution. A conjugate hyperprior distribution has been selected for each parameter of $\gamma_j = \{m_j, \Sigma_j, a_j, b_j\}$

$$\begin{aligned} m_j|\mu, \Sigma_j, \alpha &\sim \mathcal{N}_{P+1}(\mu, \Sigma_j/\alpha) && (4) \\ \Sigma_j|\Lambda, \nu &\sim \mathcal{IW}_{P+1}(\Lambda, \nu) && (5) \\ a_j|l, s &\propto s^{a-1}/\gamma(a)^l \\ b_j|p, q &\sim \mathcal{G}(p, q) \end{aligned}$$

where $\mathcal{IW}_{P+1}$ denotes the Inverse-Wishart distribution of dimension $P+1$. The prior choice of $a_j$ is conjugate with the Gamma distribution of $\psi_j$ (Eq. (3)), proposed in Fink (1995). The hyperparameters $\rho = \{\mu, \alpha, \Lambda, \nu, s, l, p, q\}$ are assumed to be specified by the practitioner, typically from the expert knowledge. For instance, the hypermean $\mu$ is the predictive prior mean of $\beta_j$, as

$$\mathbb{E}(\beta_j) = \mathbb{E}\left(\mathbb{E}(\beta_j|m_j)\right) = \mathbb{E}(m_j) = \mu. \quad (6)$$

In Section 2.3, we provide more details about the choice of the hyperparameters. For a review of the dedicated methods, please refer to Kennedy and O'Hagan, 2001. In our regression model, the constant coefficient $\beta_j^0$ stands as a random effect on the mean estimate of $Y_{ij}$. More precisely,

$$\begin{aligned} \mathbb{E}(Y_{ij}|\mathbf{x}_{ij}, \beta_j, \psi_j) &= \exp(\mathbf{x}_{ij}^T \beta_j) \\ &= \exp(\beta_j^0) \exp(\mathbf{x}_{ij}^1 \beta_j^1 + \cdots + \mathbf{x}_{ij}^P \beta_j^P) \\ &= \epsilon_j \exp(\mathbf{x}_{ij}^1 \beta_j^1 + \cdots + \mathbf{x}_{ij}^P \beta_j^P), \end{aligned}$$

where $\epsilon_j = \exp(\beta_j^0)$ is a nonnegative multiplicative random effect term.

## 2.2 The hybrid MCMC algorithm

In our hierarchical Bayesian approach, the hybrid MCMC algorithm, namely the Metropolis-Hastings-within-Gibbs algorithm (see Robert and Casella, 2004) is applied. The Gibbs sampling requires both the process variables $\theta_j$ and the parameter variables $\gamma_j$ to be treated random, which can be iteratively simulated from their full conditional posterior distributions. According to the Bayes formula

$$\pi(\sigma|\mathbf{Z}) \quad \propto \quad \pi(\mathbf{Z}|\sigma) \cdot \pi(\sigma)$$

with $\mathbf{Z}$ the observations as well as the currently simulated parameters different from $\sigma$, the full conditional posterior distributions of $\beta_j$, $\psi_j$, $m_j$, $\Sigma_j$, $a_j$ and $b_j$ can be computed. The conjugate priors of $m_j$, $\Sigma_j$, $a_j$ and $b_j$ lead to a closed-form full conditional posterior distribution. Below the $(r+1)$-th iteration, the Gibbs sampler can be described as follows.

**Gibbs sampler (at the $(r+1)$-th iteration)**

`Given` $(\beta_j^{[r]}, \psi_j^{[r]}, m_j^{[r]}, \Sigma_j^{[r]}, a_j^{[r]}, b_j^{[r]})$ `for` $r = 0, 1, 2, \ldots,$ `generate`

1. $\Sigma_j^{[r+1]}|\cdots \sim \mathcal{IW}_{P+1}\left(\Lambda + (m_j^{[r]} - \beta_j^{[r]})(m_j^{[r]} - \beta_j^{[r]})^T + \alpha(m_j^{[r]} - \mu)(m_j^{[r]} - \mu)^T,\ \ \nu + 2\right)$
2. $m_j^{[r+1]}|\cdots \sim \mathcal{N}_{P+1}\left(\frac{\alpha}{1+\alpha}\mu + \frac{1}{1+\alpha}\beta_j^{[r]},\ \ \frac{\Sigma_j^{[r+1]}}{1+\alpha}\right)$
3. $\beta_j^{[r+1]}|\ldots \sim \prod_{i=1}^{n} \mathcal{NB}(y_{ij}|\beta_j^{[r+1]}, \psi_j^{[r]}) \cdot \mathcal{N}(\beta_j^{[r+1]}|m_j^{[r+1]}, \Sigma_j^{[r+1]})$
4. $a_j^{[r+1]}|\cdots \propto \left(\psi_j^{[r]} s\right)^{a_j^{[r+1]}-1} / \Gamma\left(a_j^{[r+1]}\right)^{l+1}$
5. $b_j^{[r+1]}|\cdots \sim \mathcal{G}\left(a_j^{[r+1]} + p, \psi_j^{[r]} + q\right)$
6. $\psi_j^{[r+1]}|\ldots \sim \prod_{i=1}^{n} \mathcal{NB}(y_{ij}|\beta_j^{[r+1]}, \psi_j^{[r+1]}) \cdot \mathcal{G}(\psi_j^{[r+1]}|a_j^{[r+1]}, b_j^{[r+1]})$

The full conditional posterior distributions of $a_j$, $\beta_j$ and $\psi_j$ are not belonging to any known family of distributions. Numerical methods are thus necessary for their simulation. With help of the Exponential distribution $\mathcal{E}(-\log(\psi s))$ as an instrumental distribution, $a$ can be easily drawn through an Acceptance-Rejection sampling (see Neal, 2003) or other related numerical algorithms.

**Metropolis-Hastings algorithm** The process variables $\beta_j$ and $\psi_j$ follow a more complicated full conditional posterior distribution. Numerical methods, for instance the slice sampler (Neal, 2003), the elliptical slice sampler (Bishop, 2006) and the Metropolis-Hastings (MH) algorithm (see for instance Tierney, 1995), should be considered. As mentioned in Dittmar (2013), in higher dimension the convergence gets significantly worse for slice samplers while with a convenient variance, the MH algorithm can still get efficient convergence. The MH algorithm has thus been chosen, which is based on an instrumental distribution $\mathcal{J}$ from which it is possible to sample. The choice of $\mathcal{J}$ is thus a critical issue. Although the convergence of the MH algorithm is ensured under some generic conditions on the target distribution (see Tierney, 1995), the rate of convergence depends strongly on $\mathcal{J}$ and could be very slow. Ideally, an efficient instrumental distribution should be close to the target distribution or permit enough exploration to find the area where the target density is high. At the $s$-th iteration of MH algorithm, three kinds of instrumental distributions can be considered to draw $\beta_{j,s}^{[r+1]}$:

$\mathcal{J}_1$ : $\mathcal{N}\left(m_j^{[r+1]},\, \Sigma_j^{[r+1]}\right)$, the inherited prior distribution leading to an independent MH algorithm;

$\mathcal{J}_2$ : $\mathcal{N}\left(\beta_{j,s-1}^{[r+1]},\, \Phi_j^{[r+1]}\right)$, a Normal distribution centered on the current value with random variance

$$\Phi_j^{[r+1]} \quad = \quad \left( \frac{\mathbf{x}_j^T \mathbf{x}_j}{\mathrm{Var}(\log \epsilon_j)} + \mathrm{diag}(\Sigma_j^{[r+1]})^{-1} \right)^{-1}, \tag{7}$$

where $\epsilon_j \sim \mathcal{G}\left(\psi_j^{[r]}, \psi_j^{[r]}\right)$, $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T \in \mathcal{M}_{n\times(P+1)}$, and $\mathrm{diag}(\Sigma_j^{[r+1]})^{-1}$ denotes the diagonal matrix holding the diagonal of the inverse of the current variance matrix $\Sigma_j^{[r+1]}$. The construction of $\Phi_j^{[r+1]}$ is original which accounts for the variability of the covariates $\mathbf{x}_j$, the current variance matrix and the current dispersion information;

$\mathcal{J}_3$ : $\mathcal{N}\left(\beta_{j,s-1}^{[r+1]},\, \kappa \cdot \Sigma_j^{[r+1]}\right)$, a Normal distribution centered on the current value with variance adjusted by the parameter $\kappa$, leading to a symmetric MH algorithm.

Similarly, three different instrumental distributions can be tried for $\psi_{j,s}^{[r+1]}$:

$\tilde{\mathcal{J}}_1$ : $\mathcal{G}\left(a_j^{[r+1]},\, b_j^{[r+1]}\right)$, the inherited prior distribution leading to an independent MH algorithm;

$\tilde{\mathcal{J}}_2$ : $\mathcal{G}\left(n\,\psi_{j,s-1}^{[r+1]} + a_j^{[r+1]},\, n + b_j^{[r+1]}\right)$, a Gamma distribution with both the shape and rate parameters fixed close to those of the target distribution, a reasonable mean value being kept, as

$$\begin{aligned} \mathbb{E}(\widetilde{\psi}_{j,s}^{[r+1]} | \ldots) \quad &= \quad \frac{n\,\psi_{j,s-1}^{[r+1]} + a_j^{[r+1]}}{n + b_j^{[r+1]}} \;=\; \frac{n}{n + b_j^{[r+1]}}\,\psi_{j,s-1}^{[r+1]} \;+\; \frac{b_j^{[r+1]}}{n + b_j^{[r+1]}}\,\widehat{\psi}_j \\ &\simeq \quad \psi_{j,s-1}^{[r+1]}, \end{aligned}$$

where $\widehat{\psi}_j = a_j^{[r+1]}/b_j^{[r+1]}$ denotes an unbiased estimator of $\psi$ which should not be far from the current simulator $\psi_{j,s-1}^{[r+1]}$. It is coherent with $\mathbb{E}(\psi_j) = a_j/b_j$ deriving from the prior choice $\psi_j \sim \mathcal{G}(a_j, b_j)$;

$\tilde{\mathcal{J}}_3$ : $\mathcal{N}\left(\psi_{j,s-1}^{[r+1]},\, \kappa \cdot {\psi_{j,s-1}^{[r+1]}}^2\right)$, a Normal distribution centered on the current value with $\kappa$ the squared coefficient of variation.

Numerical experiments have been carried out to find the well fitting instrumental distributions, detailed in Section 3.2. The MH algorithm itself is detailed in Appendix A. 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. Once the convergence has been accepted for each parameter of interest, we could start collecting the drawn variables and further statistical tests could be carried out on these stocked samples.

## 2.3 Calibration of conjugate priors

**Conjugate prior of $m_j$** We aim to calibrate the hyperparameters $\mu$ and $\alpha$ in the prior distribution $m_j|\mu, \Sigma_j, \alpha \sim \mathcal{N}_{P+1}(\mu, \Sigma_j/\alpha)$. The prior mean $\mu$ can be chosen to be $\beta_{\mathrm{Exp}}$, the expert's chosen regression coefficients as indicated in Eq. (6). We calibrate $\alpha$ from the full conditional posterior distribution of $m_j$, which possesses $\frac{1}{1+\alpha}\beta_j + \frac{\alpha}{1+\alpha}\mu$

as mean. It mixes the data information $\beta_j$ and the prior knowledge $\mu$ with respective importance weights 1 and $\alpha$. Since $\beta_j$ is based on $n$ observations $\{y_{1j}, \ldots, y_{nj}\}$ and $\mu$ is the prior mean of a virtual sample $\{m_j\}$, $\alpha$ can thus be regarded as the ratio of the virtual sample size to the observation sample size $n$. $\alpha$ can be adjusted with respect to our knowledge or belief on the prior information. When $\alpha$ is close to 0, the impact of the prior distribution disappears. When $\alpha$ is large, there is no more impact of data. A default choice is $\alpha = 0.01$, which means that the prior information is as important as the information brought by *one* data among 100 observations.

**Conjugate prior of $\Sigma_j$** Here we calibrate the hyperparameters $\Lambda$ and $\nu$ in $\Sigma_j|\Lambda, \nu \sim \mathcal{IW}_{P+1}(\Lambda, \nu)$. This is more challenging as $\Lambda$ and $\nu$ are more difficult to interpret. We first fix the inverse scale matrix $\Lambda = t \cdot \Sigma_{\text{Exp}}$, where $\Sigma_{\text{Exp}}$ denotes the expert's chosen variance matrix and $t$ is the related hyperparameter to be specified. This formulation is natural since from the prior choice of $\Sigma_j$ we have

$$\mathbb{E}(\Sigma_j) \;=\; \frac{\Lambda}{\nu - (P+1) - 1} \;=\; \frac{t}{\nu - P - 2}\Sigma_{\text{Exp}} \;=\; \Sigma_{\text{Exp}},$$

by fixing $\nu = t + P + 2$. In the following, we only need to calibrate $t$. We choose to analyse the full conditional posterior distribution of $\Sigma_j$, which is an Inverse-Wishart distribution as given in Gibbs sampler. The inverse scale matrix contains three terms, where the second one $(m_j - \beta_j)(m_j - \beta_j)^T$ and the third one $\alpha(m_j - \mu)(m_j - \mu)^T$ correspond to the total squared derivation (sd.) within the sample $\{\beta_j\}$ of relative size 1 and the total sd. within the virtual sample $\{m_j\}$ of relative size $\alpha$, respectively. They indicate an unbiased estimator of the sample variance

$$\widehat{\Sigma}_j \;=\; (m_j - \beta_j)(m_j - \beta_j)^T \;=\; \alpha(m_j - \mu)(m_j - \mu)^T.$$

The full conditional posterior distribution of $\Sigma_j$ can then be written as

$$\Sigma_j|\ldots \;\sim\; \mathcal{IW}_{P+1}\left(t\,\Sigma_{\text{Exp}} \;+\; 2\,\widehat{\Sigma}_j,\;\; \nu + 2\right).$$

Under the assumption that $\nu = t + P + 2$, the posterior mean $\mathbb{E}(\Sigma_j|\ldots)$ equals $\frac{t}{t+2}\Sigma_{\text{Exp}} + \frac{2}{t+2}\widehat{\Sigma}_j$. This is an elegant expression which allows us to tune the importance of the prior information through $t$ with respect to the data information. $t$ is accordingly homogeneous to 2. Recalling that $\alpha$ is interpreted as the relative size of a virtual sample which is homogeneous to 1, we take thus $t = \alpha + 1$.

# 3 Numerical experiments

In our experiments, the Bayesian approach has been applied upon the simulated data as well as the real data for a GLM regression problem. The classical MLE has been chosen as a benchmark and was provided by the function *glm.nb* in the MASS package of R.

## 3.1 Test statistics description

The following criteria have been called in our experiments.

**Mean squared error (MSE)** The MSE criterion measures the accuracy of the prediction on the test set, defined as $\text{MSE}(\widehat{\mathbf{y}}) = \mathbb{E}\left[(\widehat{\mathbf{y}} - \mathbf{y}^{\text{F}})^2\right]$ the expected value of the squared difference between the fitted values $\widehat{\mathbf{y}}$ and the true future observations $\mathbf{y}^{\text{F}}$. It can be easily estimated as

$$\widehat{\text{MSE}}(\widehat{y}) \;=\; \frac{1}{m}\sum_{i=1}^{m}(\widehat{y}_i - y_i^{\text{F}})^2, \;\; \text{with } \widehat{y}_i \;=\; \exp\left((x_i^{\text{F}})^T\widehat{\beta}\right).$$

In this expression, $m$ is the test sample size, $x_i^{\mathrm{F}}$ is the $i$-th future covariates and $\widehat{\beta}$ denotes the regression coefficients estimated from the previous observations $\mathbf{y}$ and $\mathbf{x}$. $\widehat{\beta}$ is typically derived from maximizing the log likelihood in the frequentist approach. In the Bayesian approach,

$$\widehat{\beta} \;=\; \frac{1}{G}\sum_{g=1}^{G}\beta^{(g)},$$

with $\beta^{(g)} \sim \pi(\cdot|\mathbf{y},\mathbf{x})$ a posterior sample resulting from the hybrid MCMC algorithm.

**Logarithmic score (LS)** The logarithmic score is a strictly proper scoring rule, which measures the prediction accuracy in a probabilistic level. It can be expressed as $\mathrm{LS}(p,i) = \log p(i)$, where $p(i)$ denotes the forecast probability that the event $i$ is realized. In the Bayesian approach, let $\{\theta^{(1)},\ldots,\theta^{(G)}\}$ be $G$ quasi-i.i.d. samples of $\theta = (\beta,\psi)$ derived from the Gibbs sampler. Thus $\theta^{(g)} \sim \pi(\cdot|\mathbf{y},\mathbf{x})$. Given $\theta^{(g)} = (\beta^{(g)},\psi^{(g)})$, we replicate the $m$ future data $\mathbf{y}^{\mathrm{F}}$ from the $m$ future covariates $\mathbf{x}^{\mathrm{F}}$, denoted by $\mathbf{y}^{\mathrm{rep}}$, as

$$\mathbf{y}^{\mathrm{rep}}|\theta^{(g)} \;\sim\; \mathcal{NB}\left(\exp\left((\mathbf{x}^{\mathrm{F}})^T\beta^{(g)}\right),\ \psi^{(g)}\right), \tag{8}$$

which investigate in the following posterior predictive distribution

$$\pi(\mathbf{y}^{\mathrm{rep}}|\mathbf{y},\mathbf{x}) \;=\; \int \pi(\mathbf{y}^{\mathrm{rep}}|\theta)\,\pi(\theta|\mathbf{y},\mathbf{x})\,d\theta.$$

Thus, the predictive logarithmic score would be

$$\mathrm{LS} \;=\; \int \log\mathbb{P}(\mathbf{y}^{\mathrm{rep}}=\mathbf{y}^{\mathrm{F}}|\theta)\,\pi(\theta|\mathbf{y},\mathbf{x})\,d\theta \;\simeq\; \frac{1}{mG}\sum_{g=1}^{G}\sum_{i=1}^{m}\log\mathbb{P}(\mathbf{y}_i^{\mathrm{rep}}=y_i^{\mathrm{F}}|\theta^{(g)}),$$

where $\mathbb{P}(\mathbf{y}^{\mathrm{rep}}=\mathbf{y}^{\mathrm{F}}|\theta^{(g)})$ is the probability of the NB distribution at point $\mathbf{y}^{\mathrm{F}}$ as described in Eq. (8). On the other hand, in the frequentist approach, LS can be easily expressed as $\frac{1}{m}\sum_{i=1}^{m}\log\mathbb{P}(\mathbf{y}_i^{\mathrm{rep}}=y_i^{\mathrm{F}}|\widehat{\theta})$, with $\widehat{\theta}=(\widehat{\beta},\widehat{\psi})$ the maximum likelihood estimator.

***p*-value for the average statistic T** The average statistic T is defined on the test set $\mathbf{y}^{\mathrm{F}}$, as $\mathrm{T}(\mathbf{y}^{\mathrm{F}}) = \bar{\mathbf{y}}^{\mathrm{F}} = \frac{1}{m}\sum_{i=1}^{m} y_i^{\mathrm{F}}$. In the Bayesian context, given $\theta^{(g)}$ derived from the Gibbs sampler ($g=1,\ldots,G$) and future covariates $\mathbf{x}^{\mathrm{F}}$, we get $M$ repetitions of replicated data as $\mathbf{y}_M^{(g)} = \{\mathbf{y}_1^{(g)},\ldots,\mathbf{y}_M^{(g)}\}$, with each $\mathbf{y}_j^{(g)} = \{y_{j1}^{(g)},\ldots,y_{jm}^{(g)}\}$ replicating the $m$ future data $\mathbf{y}^{\mathrm{F}} = \{y_1^{\mathrm{F}},\ldots,y_m^{\mathrm{F}}\}$ derived from the NB distribution. The whole replicated sample is denoted by $\mathbf{y}_M^{\mathrm{rep}} = \{\mathbf{y}_M^{(1)},\ldots,\mathbf{y}_M^{(G)}\}$. T can then be calculated in the posterior sense as

$$\begin{aligned}
\mathrm{T}(\mathbf{y}_M^{\mathrm{rep}}|\mathbf{y},\mathbf{x}) = \mathbb{E}(\bar{\mathbf{y}}_M^{\mathrm{rep}}|\mathbf{y},\mathbf{x}) \;&=\; \int \bar{\mathbf{y}}_M^{\mathrm{rep}}\,\pi(\mathbf{y}_M^{\mathrm{rep}}|\mathbf{y},\mathbf{x})\,d\mathbf{y}_M^{\mathrm{rep}} \\
&=\; \iint \bar{\mathbf{y}}_M^{\mathrm{rep}}\,\pi(\mathbf{y}_M^{\mathrm{rep}}|\theta)\,\pi(\theta|\mathbf{y},\mathbf{x})\,d\theta\,d\mathbf{y}_M^{\mathrm{rep}} \\
&\simeq\; \frac{1}{GMm}\sum_{g=1}^{G}\sum_{j=1}^{M}\sum_{i=1}^{m} y_{ji}^{(g)}.
\end{aligned}$$

In the frequentist approach, with the MLE $\widehat{\theta}=(\widehat{\beta},\widehat{\psi})$ and the future covariates $\mathbf{x}^{\mathrm{F}}$, $M$ repetitions of replicated data $\widehat{\mathbf{y}}_M^{\mathrm{rep}} = \{\widehat{\mathbf{y}}_1,\ldots,\widehat{\mathbf{y}}_M\}$ can be obtained, with each $\widehat{\mathbf{y}}_j = \{\widehat{\mathbf{y}}_{j1},\ldots,\widehat{\mathbf{y}}_{jm}\}$ replicating the $m$ future data $\mathbf{y}^{\mathrm{F}}$. In this case, T can be calculated as

$$\mathrm{T}(\widehat{\mathbf{y}}_M^{\mathrm{rep}}|\widehat{\theta}) = \mathbb{E}(\bar{\widehat{\mathbf{y}}}_M^{\mathrm{rep}}|\widehat{\theta}) \;=\; \int \bar{\widehat{\mathbf{y}}}_M^{\mathrm{rep}}\,\pi(\widehat{\mathbf{y}}_M^{\mathrm{rep}}|\widehat{\theta})\,d\widehat{\mathbf{y}}_M^{\mathrm{rep}} \;\simeq\; \frac{1}{Mm}\sum_{j=1}^{M}\sum_{i=1}^{m}\widehat{y}_{ji}.$$

In both cases, we can compute its confidence interval and $p$-value from the replicated data. The $p$-value is defined as the probability that the replicated value is more extreme than the observed data. The classical form (in frequentist inference) is as follows.

$$\begin{aligned} p_C &= \mathbb{P}\left(\mathrm{T}(\widehat{\mathbf{y}}_M^{\mathrm{rep}}) \geq \mathrm{T}(\mathbf{y}^{\mathrm{F}})|\widehat{\theta}\right) = \int \mathbf{1}_{\{\mathrm{T}(\widehat{\mathbf{y}}_M^{\mathrm{rep}})\geq \mathrm{T}(\mathbf{y}^{\mathrm{F}})\}}\, \pi(\widehat{\mathbf{y}}_M^{\mathrm{rep}}|\widehat{\theta})\, d\widehat{\mathbf{y}}_M^{\mathrm{rep}} \\ &\simeq \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}_{\{\bar{\widehat{\mathbf{y}}}_j \geq \bar{\mathbf{y}}^{\mathrm{F}}\}}. \end{aligned}$$

The Bayesian $p$-value for T can be estimated as follows.

$$\begin{aligned} p_B &= \mathbb{P}\left(\mathrm{T}(\mathbf{y}_M^{\mathrm{rep}}) \geq \mathrm{T}(\mathbf{y}^{\mathrm{F}})|\mathbf{y},\mathbf{x}\right) = \int \mathbf{1}_{\{\mathrm{T}(\mathbf{y}_M^{\mathrm{rep}})\geq \mathrm{T}(\mathbf{y}^{\mathrm{F}})\}}\, \pi(\mathbf{y}_M^{\mathrm{rep}}|\mathbf{y},\mathbf{x})\, d\mathbf{y}_M^{\mathrm{rep}} \\ &= \iint \mathbf{1}_{\{\mathrm{T}(\mathbf{y}_M^{\mathrm{rep}})\geq \mathrm{T}(\mathbf{y}^{\mathrm{F}})\}}\, \pi(\mathbf{y}_M^{\mathrm{rep}}|\theta)\, \pi(\theta|\mathbf{y},\mathbf{x})\, d\theta\, d\mathbf{y}_M^{\mathrm{rep}} \\ &\simeq \frac{1}{MG}\sum_{j=1}^{M}\sum_{g=1}^{G} \mathbf{1}_{\{\bar{\mathbf{y}}_j^{(g)} \geq \bar{\mathbf{y}}^{\mathrm{F}}\}}. \end{aligned}$$

**Marginal effect (ME) of standardized covariates** We first define the $i$-th standardized covariate $k$ ($i = 1, \ldots, m$; $k = 0, 1, \ldots, P$) as

$$\widetilde{x}_i^k = \frac{x_i^k - \bar{x}^k}{\sqrt{\mathrm{Var}(\mathbf{x}^k)}} := \frac{x_i^k - \bar{x}^k}{\sigma^k}$$

with $\bar{x}^k = \frac{1}{m}\sum_{i=1}^{m} x_i^k$ and $\mathbf{x}^k = \{x_1^k, \ldots, x_m^k\}$. By noting $\mathbf{x}_i = (x_i^0, x_i^1, \ldots, x_i^P)$, $\widetilde{\mathbf{x}}_i = (\widetilde{x}_i^0, \widetilde{x}_i^1, \ldots, \widetilde{x}_i^P)$, $\bar{\mathbf{x}} = (\bar{x}^0, \bar{x}^1, \ldots, \bar{x}^P)$ and $\sigma = (\sigma^0, \sigma^1, \ldots, \sigma^P)$, we have $\mathbf{x}_i = \bar{\mathbf{x}} + \sigma\widetilde{\mathbf{x}}_i$. The marginal effect of $\widetilde{x}_i^k$ on the expected count data can be expressed as

$$\begin{aligned} \frac{\partial\, \mathbb{E}(\mathbf{y}_i|\mathbf{x}_i)}{\partial\, \widetilde{x}_i^k} &= \frac{\partial\, \exp(\mathbf{x}_i^T\beta)}{\partial\, \widetilde{x}_i^k} = \frac{\partial\, \exp((\bar{\mathbf{x}} + \sigma\widetilde{\mathbf{x}}_i)^T\beta)}{\partial\, \widetilde{x}_i^k} \\ &= \sigma^k\beta^k \exp(\mathbf{x}_i^T\beta). \end{aligned}$$

Thus, the mean marginal effect of the standardized covariate $k$ can be estimated by

$$\widehat{\mathrm{ME}}(k) = \sigma^k\widehat{\beta}^k \exp(\bar{\mathbf{x}}^T\widehat{\beta}), \quad \text{with } \bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i,$$

which measures the marginal influence on the mean count data brought by one unit change of the covariate $k$. The marginal effect can be considered as an importance index to determine the dominant covariate and it possesses significant features for understanding and interpreting the count prediction.

## 3.2 A simulated case study

This simulation study can be considered as an exploring research.

**Data generation** Assuming there were $J = 4$ regions, in each region we constructed a training set $(\mathbf{y}, \mathbf{x})$ with $n = 24$ data ($n$ was increased to 100 in the final experiment) and a test set $(\mathbf{y}^{\mathrm{F}}, \mathbf{x}^{\mathrm{F}})$ with $m = 120$ data. The former is for the modeling procedure and the latter is for the prediction validation. The count data $(\mathbf{y}, \mathbf{y}^{\mathrm{F}})$ were generated from the NB distribution:

$$y_{ij} \sim \mathcal{NB}\left(\exp(x_{ij}^T\beta_j),\, \psi_j\right),\ i = 1, \ldots, n+m,\ j = 1, \ldots, J,$$

where $\beta_j = [\beta_j^0, \beta_j^1, \ldots, \beta_j^P]^T$ with each $\beta_j^k \sim \mathcal{U}(0, 1)$ a Uniform distribution between 0 and 1; $\mathbf{x}_{ij} = [1, x_{ij}^1, \ldots, x_{ij}^P]^T$ with each $x_{ij}^k \sim \mathcal{N}(0, 0.3^2)$ and $\psi_j \sim \mathcal{U}(5, 40)$. The

number $P$ of covariates was assumed to vary among 4, 6 and 8 ($P$ was increased to 20 in the final experiment). In each case, the generation was repeated 50 times to obtain 50 independent datasets. In the GLM regression problem, we compared the performances of the frequentist and Bayesian approaches, with help of the criteria MSE and LS.

**Choosing prior distributions** In the Gibbs sampler, we assumed that no expert knowledge was available in order to set up a more challenging problem. A non-informative prior distribution has been chosen on the hyperparameters $\rho = \{\mu, \alpha, \Lambda, \nu, s, l, p, q\}$ as

$$\begin{aligned} \mu &= \mathbf{0}_{P+1} \qquad (9) \\ \alpha &= 0.01 \\ \Sigma_{\text{Exp}} &= \mathbf{I}_{P+1} \\ \Lambda &= (\alpha+1)\cdot\Sigma_{\text{Exp}} \\ \nu &= (\alpha+1)+(P+1)+1 = \alpha+P+3 \\ s &= 0.001 \\ l &= 1 \\ p &= 1 \\ q &= 1, \qquad (10) \end{aligned}$$

where $\mathbf{I}_{P+1}$ denotes an identity matrix of dimension $(P+1)\times(P+1)$. $\alpha = 0.01$ tunes the prior knowledge as important as 1 observation in 100 samples. $\Lambda$ and $\nu$ have been chosen following the idea in Section 2.3. See more details on the Bayesian diagnostics of prior-data agreement in Fu et *al.*, 2012 and Bousquet, 2008.

**Checking instrumental distributions** As explained before, an important issue of the MH algorithm is to find out well fitting instrumental distributions for $\beta_j$ and $\psi_j$. Following the proposition in Section 2.2, we tested four possible instrumental distributions for each variable of interest. 16 combinations have thus been created and the related MSE could be compared. We show in Figure 2 that with the following choices we reached the minimal MSE averaged on 50 independent datasets.

$$\begin{aligned} \widetilde{\beta}_{j,s}^{[r+1]} &\sim \mathcal{N}\left(\beta_{j,s-1}^{[r+1]},\, \Phi_j^{[r+1]}\right); \\ \widetilde{\psi}_{j,s}^{[r+1]} &\sim \mathcal{N}\left(\psi_{j,s-1}^{[r+1]},\, \left(0.05\cdot\psi_{j,s-1}^{[r+1]}\right)^2\right). \end{aligned}$$

The construction of $\Phi_j^{[r+1]}$ was described in Eq. (7). We emphasize that these 50 repeated datasets were totally different from what we used in the following.

Three parallel Markov chains were simulated within 60,000 Gibbs sampling iterations, with each Gibbs iteration embedding one MH iteration for $\beta_j$ and $\psi_j$. The first 10,000 burn-in period was discarded and once the convergence had been verified with the BG statistic, every 50-th sample was collected to generate 1,000 quasi-i.i.d. samples. The lag 50 was chosen from an ACF (Auto Correlation Function) test that the autocorrelation of each marginal sample of $\beta_j$ and $\psi_j$ remained below 0.2 when the lag exceeded 50. On a Intel® Core™ i7 processor 2.60GHz computer, it took about 1h to finish 60,000 iterations of the hybrid MCMC in this simulated case.

**Result interpretation** By increasing the number of covariates $P$ from 4 to 8, Figure 3 displays the mean predictive MSE and LS as well as their boxplots on 50 independent repetitions in 4 regions. We can see that the Bayesian approach overperforms the frequentist method in terms of both MSE and LS, regardless of number

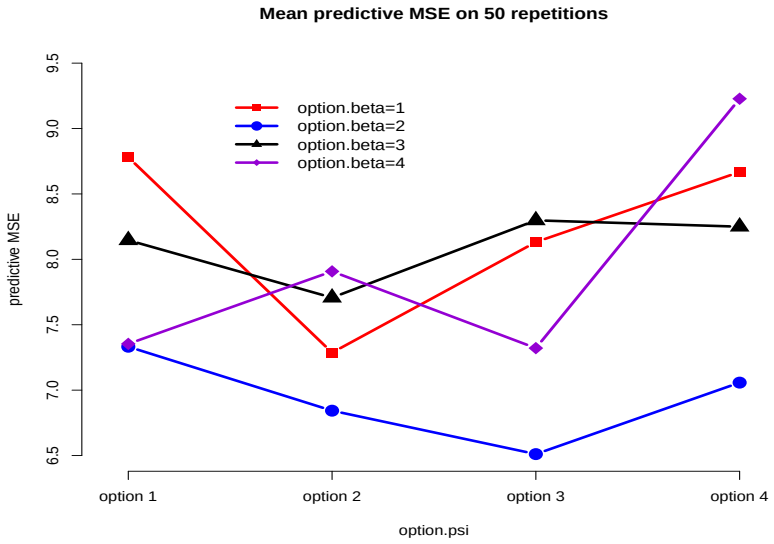


**Figure 2:** The predictive averaged MSE on 50 datasets with variate instrumental distributions of $\beta$: $\mathcal{J}_1 : \mathcal{N}(m_j^{[r+1]}, \Sigma_j^{[r+1]})$, $\mathcal{J}_2 : \mathcal{N}(\beta_{j,s-1}^{[r+1]}, \Phi_j^{[r+1]})$, $\mathcal{J}_3 : \mathcal{N}(\beta_{j,s-1}^{[r+1]}, 0.05^2 \cdot \Sigma_j^{[r+1]})$, $\mathcal{J}_4 : \mathcal{N}(\beta_{j,s-1}^{[r+1]}, 0.25^2 \cdot \Sigma_j^{[r+1]})$ and variate instrumental distributions of $\psi$: $\tilde{\mathcal{J}}_1 : \mathcal{G}(a_j^{[r+1]}, b_j^{[r+1]})$, $\tilde{\mathcal{J}}_2 : \mathcal{G}(n\,\psi_{j,s-1}^{[r+1]} + a_j^{[r+1]}, n + b_j^{[r+1]})$, $\tilde{\mathcal{J}}_3 : \mathcal{N}(\psi_{j,s-1}^{[r+1]}, (0.05 \cdot \psi_{j,s-1}^{[r+1]})^2)$, $\tilde{\mathcal{J}}_4 : \mathcal{N}(\psi_{j,s-1}^{[r+1]}, (0.25 \cdot \psi_{j,s-1}^{[r+1]})^2)$.

of covariates. The Bayesian inference stands out especially when $P$ increases, as the GLM mixes more uncertainties and becomes more complicated. We can see an amplification of the gap between the two methods.

Then we assumed the availability of expert knowledge on the regression coefficients $\beta$, which could help us to "correctly" choose the hypermean $\mu = \beta_{\text{Exp}}$. The other hyperparameters are more difficult to interpret, we thus kept the previous non-informative values for them. Moreover, $\alpha$ was fixed to 1 to emphasize that the prior information is as important as the dataset. The Metropolis-Hastings-within-Gibbs algorithm and the MLE method have been run in the most complicated case, i.e. with 8 covariates. Figure 4 displays a comparison of the predictive LS and MSE through the MLE method, the non-informative Bayesian approach and the *partially* informative Bayesian approach. The boxplot is based on 50 repetitions in 4 regions. Benefiting from good prior information, the Bayesian approach can significantly over-perform the frequentist approach and improve the non-informative Bayesian approach.

In Figure 5, we increased the number $n$ of training data from 24 to 100 and assumed there were $P = 20$ explanatory covariates. Intuitively, the MLE method takes advantage when the sample size becomes larger, whereas the Bayesian approach is favored if the model becomes more complicated with more covariates. Figure 5 confirms this guess as the difference between the two methods decreased compared with Figure 4. Still, the performance of the Bayesian method was very satisfying.

### 3.3 A real-life case study: road crash counts in four regions of Switzerland

This experiment regards 1024 daily road accident counts registered in 3 years (2007, 2008 and 2010) in $J = 4$ regions of Switzerland. The first 344 counts were used as the training set and the last 680 counts constituted the test set. In other words, we predicted the number of accidents in 2008 and 2010 by using the registered accident counts in 2007. The considerable covariates included the rainfall, the temperature,

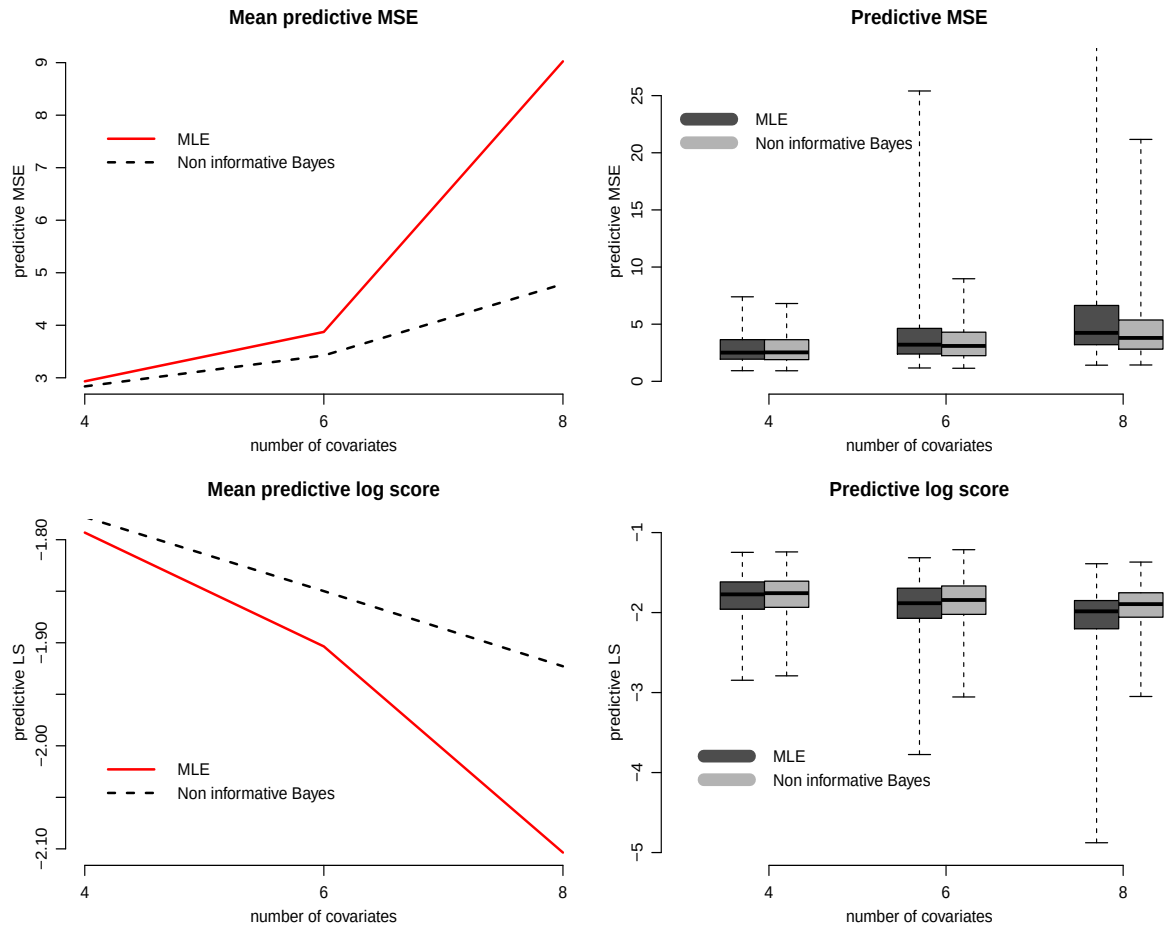


**Figure 3:** The predictive LS and MSE of accident counts as a function of the number of covariates

their interaction, the daily traffic on the main highway and the road crash counts on the previous day. The summary statistics about the main covariates on the test set were provided in Table 1. As we can see, the major traffic owned extremely high average and high standard deviation, which widely varied from region to region. We thus chose the logarithmic form to make it comparable with other covariates such as the temperature and the rainfall.

The GLM regression coefficients were estimated through the MLE and Bayesian methods. In the Bayesian context, we assumed that no expert knowledge was available to avoid any unbalanced comparison. The non-informative prior distribution has thus been taken, as expressed in Eq. (9)-(10) in Section 3.2. Once again, 60,000 Gibbs sampling iterations have been carried out with the first 10,000 burn-in period discarded. The convergence was verified with the BG statistic applied on three parallel Markov chains. On a Intel® Core™ i7 processor 2.60GHz computer, it took about 3h to finish 60,000 iterations of the hybrid MCMC for all the four regions. Once the convergence has 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 test. From these 1,000 samples, the predictive MSE, LS, ME on main covariates (i.e. traffic, temperature and rainfall), the 95% confidence interval of the predictive average crash as well as its $p$-value were calculated, as shown in Table 2.

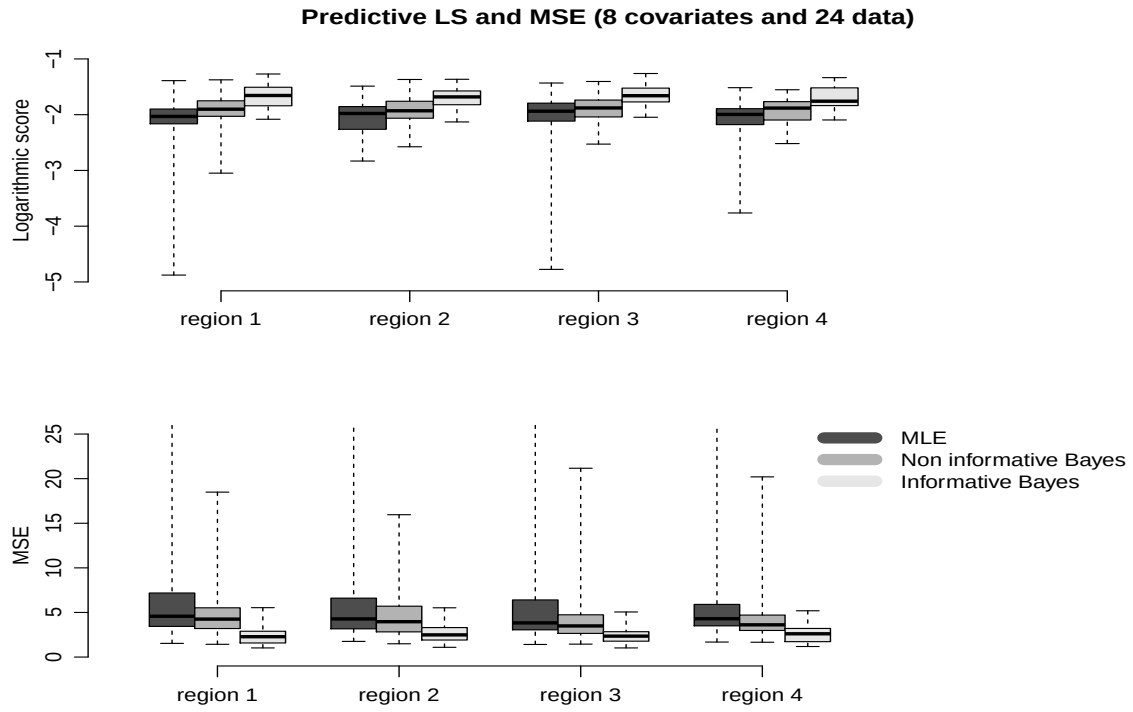


**Figure 4:** The predictive LS and MSE with 8 covariates and 24 observations in 4 regions

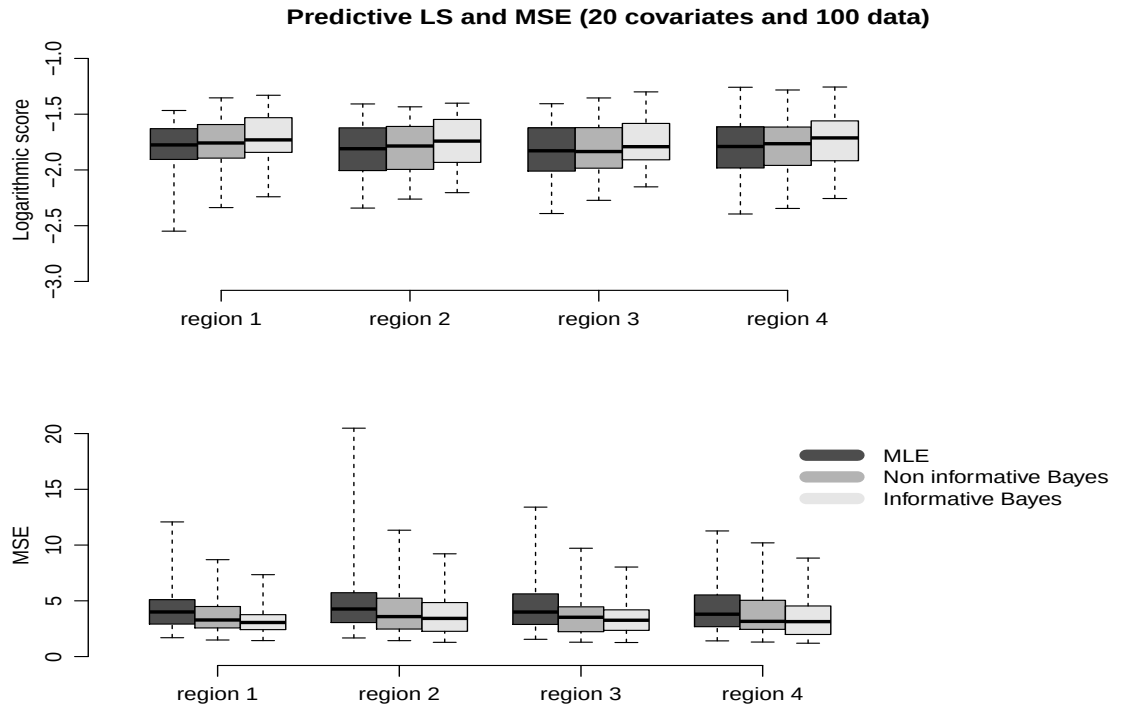


**Figure 5:** The predictive LS and MSE with 20 covariates and 100 observations in 4 regions

**Table 1:** Summary statistics in the test set of road crash counts

| Regions | Counts & Covariates | Min. | Max. | Average | St. dev. |
|---|---|---|---|---|---|
| 1 | crash counts | 1 | 22 | 8.61 | 3.58 |
| | rainfall | 0 | 102.0 | 5.37 | 12.16 |
| | temperature | -5.0 | 23.4 | 8.99 | 6.54 |
| | major traffic | 5615 | 17094 | 12904 | 2944.26 |
| | log(major traffic) | 8.63 | 9.74 | 9.43 | 0.26 |
| 2 | crash counts | 0 | 4 | 0.41 | 0.67 |
| | rainfall | 0 | 104.5 | 7.12 | 19.53 |
| | temperature | -7.7 | 20.9 | 6.49 | 6.70 |
| | major traffic | 2676 | 13603 | 7948 | 1854.34 |
| | log(major traffic) | 7.89 | 9.52 | 8.95 | 0.24 |
| 3 | crash counts | 0 | 15 | 5.36 | 2.57 |
| | rainfall | 0 | 123.0 | 5.35 | 13.26 |
| | temperature | -11.4 | 21.1 | 6.41 | 7.39 |
| | major traffic | 15305 | 45380 | 33681 | 6060.46 |
| | log(major traffic) | 9.63 | 10.72 | 10.40 | 0.20 |
| 4 | crash counts | 0 | 8 | 1.69 | 1.59 |
| | rainfall | 0 | 83.2 | 4.25 | 10.33 |
| | temperature | -9.2 | 19.6 | 5.94 | 6.42 |
| | major traffic | 9071 | 61251 | 28202 | 8154.88 |
| | log(major traffic) | 9.11 | 11.02 | 10.21 | 0.27 |

In terms of goodness of fit, the Gibbs algorithm largely improved the MLE's performance. A lower MSE, higher LS, narrower confidence interval centered on the posterior predictor of crash counts which was close to the true observation, and a lower probability that those replicated values exceeded the true ones (the so-called $p$-value) were exhibited. The improvement especially resided in regions 2 and 4, where the crash number was lower and even close to 0 (see Table 1). This means that in these two regions, we had more 0's as observation which raised the difficulty of estimation. The Bayesian approach can especially improve the MLE method in this tricky case.

ME helps to evaluate the significance of covariates in the prediction of crash counts, namely the marginal influence on the predicted crash counts from one unit change of the covariate, as shown in Table 2. The rainfall and temperature seemed the dominant covariates and their influence varied among the regions. The frequentist and Bayesian approaches gave coherent results. Attention should be paid to compute the marginal effect of the major traffic since we applied its logarithmic form in the modeling procedure. Hence we could use $\frac{\partial f(\log x)}{\partial x} = \frac{1}{x}\frac{\partial f(\log x)}{\partial (\log x)}$ to calculate the partial derivative on the original traffic value, where $\frac{\partial f(\log x)}{\partial (\log x)}$ followed the same computation as for other covariates.

Compared with the simulation study, in this case the advantage of the Bayesian inference compared with the MLE method may be limited. The reason is twofold. First, the sample size $n$ is quite large here, which may be already sufficient to get a reasonable MLE. Second, there are only 5 covariates which may reduce the modeling complexity. As shown in the previous case, we could have achieved more profit by applying the Bayesian inference if the number of covariates had increased or if the prior knowledge had been available along with few observations.

**Table 2:** The test statistics of the GLM through the frequentist and Bayesian approaches, where figures in **black** signify superior results.

| Regions | Test statistics | MLE | Non-informative Bayes |
|---|---|---|---|
| 1 | MSE | 11.041 | **11.016** |
| | LS | -2.832 | **-2.830** |
| | ME (rainfall) | 0.372 | 0.369 |
| | ME (temperature) | 0.201 | 0.208 |
| | ME (traffic) | 1.377e-04 | 1.329e-04 |
| | true average crash | 8.607 | |
| | 95% CI | [4.415, 11.747] | **[5.348, 10.579]** |
| | $p$-value | 0.650 | **0.527** |
| 2 | MSE | 0.678 | **0.460** |
| | LS | -0.875 | **-0.726** |
| | ME (rainfall) | 0.003 | 0.016 |
| | ME (temperature) | -0.041 | -0.030 |
| | ME (traffic) | 1.354e-05 | 5.171e-06 |
| | true average crash | 0.410 | |
| | 95% CI | [0.161, 0.668] | **[0.232, 0.603]** |
| | $p$-value | 0.269 | **0.212** |
| 3 | MSE | 5.604 | **5.592** |
| | LS | -2.245 | **-2.243** |
| | ME (rainfall) | 0.106 | 0.103 |
| | ME (temperature) | -0.197 | -0.146 |
| | ME (traffic) | 4.382e-05 | 3.999e-05 |
| | true average crash | 5.359 | |
| | 95% CI | [2.652, 7.535] | **[3.308, 6.877]** |
| | $p$-value | 0.671 | **0.492** |
| 4 | MSE | 3.163 | **2.683** |
| | LS | -2.108 | **-1.923** |
| | ME (rainfall) | -0.032 | -0.035 |
| | ME (temperature) | -0.152 | -0.083 |
| | ME (traffic) | 1.737e-05 | 1.435e-05 |
| | true average crash | 1.694 | |
| | 95% CI | [0.971, 3.136] | **[1.063, 1.969]** |
| | $p$-value | 0.399 | **0.149** |

# 4 Discussion

In this paper we propose a hierarchical Bayesian method to address the NB generalized regression problems. The hierarchical Bayesian inference does not focus on providing independent estimator on each subset but rather on modeling the coherence among the subsets and collecting the whole related uncertainty. Accounting for prior information may be valuable in a small sample size setting. Our numerical experiments suggest that our Bayesian methodology is a worthy competitor to the classical MLE. In the simulation study, we checked the relative merits of the methods in four different scenarios:

a) 4 covariates and 24 data – both methods do well;

b) 6 covariates and 24 data – Bayesian regression works slightly better than MLE;

c) 8 covariates and 24 data – Bayesian regression works much better than MLE and the informative Bayesian regression does even better;

d) 20 covariates and 100 data – both methods do well and Bayesian regression (with both the non-informative and informative priors) shows remarkable stability.

As shown in Table 3, the MLE works better than the Bayesian method when there are numerous data and few considerable covariates, which is in fact a trivial case. Whereas the Bayesian method gains more advantages in both accuracy and robustness in the other three cases. It is especially preferred in the quite tricky situation with few observations and a large number of covariates. It is also suggested in the intermittent series (many 0's as observation) or missing-data cases.

**Table 3:** Preference between the MLE and Bayesian approaches, where $\gg$ means high preference and $>$ means low preference ($n$ = size of dataset, $P$ = number of covariates).

| $s = \frac{n}{P}$ | small $n$ | large $n$ |
|---|---|---|
| small $P$ | Bayesian $>$ MLE | MLE $>$ Bayesian |
| large $P$ | Bayesian $\gg$ MLE | Bayesian $>$ MLE |

In terms of robustness, the MLE procedure had sometimes convergence concerns while treating with noisy data. We got warning messages returned by the mathematical software. On the other hand, the Gibbs sampling always reached its convergence to the stationary posterior distribution, even if it could take a long enough waiting time. This has been highlighted in both the generated and real-life case studies.

This NB hierarchical Bayesian approach can be applied to various research themes. For instance, one would like to predict the number of degraded rotating machines (Garnero and Montgomery, 2006), the number of patients affected by infectious or viral diseases (Gentleman et *al.*, 1994), the discrete level of pollutants propagations (Ang and Tang, 1984; Zhang and Dai, 2007), etc. If the data $\mathbf{y}$ contain missing values, one can consider accelerating the MCMC through an adaptive augmentation procedure (see Pasanisi et *al.*, 2012). In addition, if the multivariate variable $\mathbf{y}$ is highly correlated among its sub-components, 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 B. Another perspective is to apply the LASSO technique to improve the predictive performance of the Bayesian inference by shrinking the insignificant coefficients to zero. (see Tibshirani, 1996; Fu, 2015).

## Acknowledges

I gratefully thank Drs. Giorgio Corani and Nicolas Bousquet for helpful comments and advices. I would also like to acknowledge useful discussions with Dr. Alessio Benavoli. This work was supported by the PMA (*Percorsi Meteo-Aware*) project of SUPSI, Switzerland.

## References

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

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

[3] Bishop, C. M. (2006). Pattern recognition and machine learning, Springer-Verlag, New York.

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

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

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

[7] Dittmar, D. (2013). Slice sampling, Technical report.

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

[9] Fu, S., Celeux, G., Bousquet, N. and Couplet, M. (2015). *International Journal for Uncertainty Quantification*, **5**(1), 73-98.

[10] Fu, S. (2015). Hierarchical Bayesian LASSO for a negative binomial regression, *Journal of Statistical Computation and Simulation*, DOI:10.1080/00949655.2015.1106541.

[11] 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.

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

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

[14] 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.

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

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

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

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

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

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

[21] 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.

[22] 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.

[23] 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.

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

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

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

[27] Tierney, L. (1995), Introduction to general state-space Markov chain theory, *Markov Chain Monte Carlo in Practice*, 59-74, Chapman & Hall.

[28] 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.

[29] 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.

# A The Metropolis-Hastings algorithm

Given the current simulated value of other parameters as well as current $\xi_j^{[r]}$:

1. Let $\xi_{j,0} = \xi_j^{[r]}$

2. For $s = 1, \ldots, l$, updating $\xi_j^{[r]}$:

   - Generate
     $$\widetilde{\xi}_{j,s} \sim \mathcal{J}(\cdot|-, \xi_{j,s-1})$$
     where $\mathcal{J}$ is the instrumental distribution.
   - Let
     $$\alpha\left(\xi_{j,s-1}, \widetilde{\xi}_{j,s}\right) = \min\left\{\frac{\pi(\widetilde{\xi}_{j,s}|-)}{\pi(\xi_{j,s-1}|-)}\frac{\mathcal{J}(\xi_{j,s-1}|-, \widetilde{\xi}_{j,s})}{\mathcal{J}(\widetilde{\xi}_{j,s}|-, \xi_{j,s-1})}, 1\right\},$$
     take
     $$\xi_{j,s} = \begin{cases} \widetilde{\xi}_{j,s}, & \text{with probability } \alpha\left(\xi_{j,s-1}, \widetilde{\xi}_{j,s}\right); \\ \xi_{j,s-1}, & \text{otherwise.} \end{cases}$$

3. Let $\xi_j^{[r+1]} = \xi_{j,l}$

# B 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

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

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 $\psi$ denotes the common dispersion parameter. The probability density is

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

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 NM 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.