Normal-Compound Gamma priors with Count Data

Ahmed Alhamzawi\textsuperscript{a*}, Gorgees Shaheed Mohammad\textsuperscript{b}

\textsuperscript{a}University of Al-Qadisiyah, College of Science, Department of Mathematics, Iraq. Email: ahmed.alhamzawi@qu.edu.iq

\textsuperscript{b}University of Al-Qadisiyah, College of Education, Department of Mathematics, Iraq. Email: gorgees.alsalamy@qu.edu.iq

\section*{1. Introduction}

Consider the linear model

\[ y = X\beta + \epsilon, \]

where \( y = (y_1, \ldots, y_n)^T \), \( X = (x_{11}, \ldots, x_{1p}), \beta = (\beta_1, \ldots, \beta_p)^T \) and \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \) defined by \( \epsilon_i \sim N(0, \sigma^2) \), are the vector of observations, the design matrix of covariates, the vector of unknown regression coefficient and the vector of errors, respectively. In this paper we assume that this model is concerned with count data. The study of Bayesian regression for count data has become an important area in this subject. Several studies have been proposed such as the study of the crash frequency and its influencing factors \cite{17}, the study of university credits and its relationship to pre-enrolment assessment tests \cite{7}, the Malaysian motor insurance claim \cite{5}, socioeconomic factors and number of tuberculosis \cite{16} and the application of count data to environmental epidemiology \cite{11}. These varies examples show the importance and the need to develop...
The development of count data models in the framework of Bayesian regression was develop in several stages beginning in the 1970s to the modern day [14,6,13]. The most notable of these stages is the work by [9] involving the use of a Poisson process to analysis regression models. Some of modern advances in this arena have involved the introduction of quantiles through conditional quantile functions [10] which requires some particular assumptions about the model while allowing the researcher to analyze the effect of the covariates on each quantile of the distribution [12]. Since the dependent variable \( y \) defined by (1) is a generated using a Poisson countable process

\[
y_i \sim \text{Poisson}(\mu_i) \\
\mu_i \sim \exp \{x_i^T \beta\} 
\]

With only nonnegative integers then it is necessary to convert it to a continuous variable. In order to do this we use the jittering process presented in [12] by adding a uniform variable \( u_i \) to our dependent variable \( y_i \) then taking the log to produce the desired continuous variable

\[
y_i^* = \ln \{y_i + u_i\} \\
u_i \sim \text{unif}(0,1)
\]

In this paper, we will analyze the Bayesian regression framework in the present of count data with a normal scale-mixture combined with a normal-compound gamma prior of the form

\[
\beta_j | \sigma^2, z_1 \sim N(0, \sigma^2 z_1) \\
z_1 \sim CG(c_1, ..., c_N, \phi) \\
\sigma^2 \sim IG(c_0, d_0)
\]

where \( CG(c_1, ..., c_N, \phi) \) is a compound-gamma distribution of order \( N \) and \( IG(c_0, d_0) \) is the inverse gamma distribution with shape parameter \( c_0 \) and scale parameter \( d_0 \).

The paper will be structured as follows: in section 2 the compound gamma prior will be introduced, in section 3 our sampler will be derived, in sections 4 and 5 we will use simulated and real data, respectively, to compare the accuracy of our model with other models.

1. The normal-compound gamma prior

The compound gamma prior can be written as

\[
\pi(x) = \int_0^\infty ... \int_0^\infty G(z_1, c_1, z_2)G(z_2, c_2, z_3) ... G(z_N, c_N, z_{N+1})dz_2 ... dz_N
\]

Or alternatively [1]

\[
\pi(x) = \int_0^\infty ... \int_0^\infty \left[ \prod_{i=1}^N \frac{z_i^{c_i} z_1^{c_i-1} [-z_1 z_{i+1}]}{\Gamma(c_i)} \right] dz_2 ... dz_N
\]

where \( z_1 = x \) and \( z_{N+1} = \phi \) is a constant. A more efficient way is to write this a product of multiple scale mixtures [1] with

\[
z_1 \sim G(c_1, 1), z_2 \sim IG(c_2, 1), ..., z_N \sim (G(c_N, \phi) \text{ odd} \ N \ IG(c_N, \phi) \text{ even} \ N
\]

where \( G(a, b) \) is the gamma distribution with shape parameter \( a \) and inverse scale (rate) parameter \( b \).

**Proof.** The proof of this equivalence is giving in [1].
The properties of this prior were extensively studied in [1]. Most interestingly, it was shown that this prior is a generalization of various popular models (See [2,15,3] for $N = 2$ and [4] for $N = 4$). In particular, it was shown that this model works for data with different degrees of sparsity. This property is farther demonstrated in Figure 1. The above prior has been investigated in several scenarios and types of data. In this paper, it is our aim to derive and analyze the posterior inference in the area of count data. We notice that for small values of $c_1$, both in the cases of compounding two and four gamma distributions for there is singularity at zero and distribution mass is concentrated near zero. Thus, this shows how the model works for sparse data. This is shown by the thin lines in the graph. On the other hand, the thick lines show the prior for the non-sparse framework where more distribution has at both tails of the distribution. In [2], it is shown that for $N = 2$ the EM algorithm can accurately sample the true sparsity (or density) of our data to get more insight to the value of our hyperparameters.

2. Posterior Inference

The full conditionals for $N = 2$ were calculated by [2] to obtain

$$
\beta | \text{rest} \sim N(\mu_\beta, \Sigma^{-1}\sigma^2) \quad z_1 | \text{rest} \sim \text{GIG} \left( \frac{\beta^T \beta}{\sigma^2 z_2}, 2, c_1 - \frac{1}{2} \right) \quad z_2 | \text{rest} \sim \text{IG} \left( c_2 + \frac{1}{2} \frac{\beta^T \beta}{2\sigma^2 z_1} + 1 \right) \sigma^2 | \text{rest} \\
\sim \text{IG} \left( \frac{n + p + 2c_0}{2}, \frac{(y - X\beta)^T (y - X\beta) + \beta^T Z^{-1} \beta + 2d_0}{2} \right)
$$

(8)

More generally for $N \geq 2$, the full conditionals are given by [1]

$$
\beta | \text{rest} \sim N(\mu_\beta, \Sigma^{-1}\sigma^2) \quad z_{\text{odd}} | \text{rest} \sim \text{GIG} \left( \frac{\beta^T Z_{-k}^{-1} \beta}{\sigma^2}, 2, c_k - \frac{1}{2} \right) \quad z_{\text{even}} | \text{rest} \sim \text{IG} \left( c_k + \frac{1}{2} \frac{\beta^T Z^{-1} \beta}{2\sigma^2} + 1 \right) \sigma^2 | \text{rest} \\
\sim \text{IG} \left( \frac{n + p + 2c_0}{2}, \frac{(y - X\beta)^T (y - X\beta) + \beta^T Z^{-1} \beta + 2d_0}{2} \right)
$$

(9)

where $\mu_\beta = \xi_2^{-2} \Sigma^{-1} X^T V^{-1} (y - \xi_1 V), V = \text{diag}(v_1, ..., v_n), \quad Z = \text{diag}(\prod_{i=1}^{N_k} z_{k_1}, ..., \prod_{i=1}^{N_k} z_{k_p}), Z_{-k} = \text{diag}(\prod_{i=1,i \neq k}^{N_k} z_{k_1}, ..., \prod_{i=1,i \neq k}^{N_k} z_{k_p})$ and $\Sigma = \xi_2^{-2} X^T V^{-1} X + Z^{-1}$. Therefore, we have the normal distribution $N(\mu_\beta, \Sigma^{-1}\sigma^2)$. For updating the hyperparameters, again we follow [1] using

![Figure 1. Plot of the NCG prior with the solid, dashed represent $N = 2$ and $N = 4$, respectively, while the thick and thin lines represent larger and smaller values of $c_1$, respectively.](image-url)
\[
Q(\xi, \xi^{old}) = \sum_{k=1}^{N} \sum_{i=1}^{p} (-1)^{k+1} c_k E_{c_k^{old}}[(z_{ki})|y^*] + c_N(\phi) - \sum_{k=1}^{N} (\gamma(c_k)) \tag{10}
\]

where \(\xi^{old} = (c_1^{old}, \ldots, c_N^{old})\) and \(C\) all the terms not containing \(c_1, c_2, \ldots, c_N\). Then, we have

\[
\gamma'(c_k) = \sum_{i=1}^{p} (-1)^{k+1} E_{c_k^{old}}[(z_{ki})|y^*] + c_N(\phi)I(k = N) \tag{11}
\]

We will set \(c_0 = \sigma_0 = 10^{-5}\) in the prior of \(\sigma^2\) for estimating the hyperparameters \(c_k\) from the EM algorithm. The EM algorithm for finding the values of \(c_k\) will be used every few iterations of the MCMC algorithm. We will implement the self-adaptive normal compound model Monte Carlo EM (MCEM) algorithm.

### 3. Simulation Studies

To demonstrate the advantages of the proposed model, we will use simulated data to analyze the predictive ability of our prior and compare it to other published models. Specifically, the comparison will be applied with the Beta Prime model \((N = 2)\) proposed by [3], our prior \(NCG10\) with \((N = 10)\), the Bayesian Lasso, the Bayesian adaptive Lasso (aLasso) and the elastic net (Enet). The simulated data will be compared with the mean squared error (MSE), the false positive rate (FPR) and the false negative rate (FNR).

**Simulation 1**

In our first simulation, we will study the simulated data generated with a very sparse model by setting \(\beta = (7, 0, 0, 0, 0, 0, 0, 0, 0)\). From the results presented in Table 1 that averaged from \(n = 100\) repeated simulations with 15000 iterations each, we can see that the results show that our prior produces better than other presented methods. Furthermore, we can see that the compound gamma prior gives the smallest MSE compared to all the other models presented. The hyperparameters are updated each 100 iterations using equation (11). Additionally, our model performs very well in terms FPRs and FNRs which are necessary for selecting the best model for variable selection. From the trace plots and histograms in Figure 2, we notice how well our prior converges compared to the stationary distribution.

**Table 1. Results for Simulation 1.**

<table>
<thead>
<tr>
<th></th>
<th>MSE (sd)</th>
<th>FPR (sd)</th>
<th>FNR (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCG2</td>
<td>0.0056 (0.0050)</td>
<td>0.0000 (0.0000)</td>
<td>0.0000 (0.0000)</td>
</tr>
<tr>
<td>NCG10</td>
<td><strong>0.0036 (0.0047)</strong></td>
<td>0.0000 (0.0000)</td>
<td>0.0000 (0.0000)</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.0203 (0.0177)</td>
<td>0.0000 (0.1000)</td>
<td>0.0000 (0.2000)</td>
</tr>
<tr>
<td>aLasso</td>
<td>0.0063 (0.0064)</td>
<td>0.0000 (0.0000)</td>
<td>0.0000 (0.4000)</td>
</tr>
<tr>
<td>Enet</td>
<td>0.0261 (0.0243)</td>
<td>0.0000 (0.3000)</td>
<td>0.0000 (0.0000)</td>
</tr>
</tbody>
</table>

**Simulation 2**

To study our model farther and get a deeper view of its behavior with different types of data will a similar size of covariates as the simulation above but we will decrease the sparsity of the model by setting \(\beta = (4.5, 2, 0, 3, 0, 0, 0, 8)\) with \(n = 100\) repeated simulations with 10000 iterations each. The results are showing in Table 2. We see that the results farther prove what Figure 1 shows, namely, that as the sparsity increase the model performs better higher values of \(N\) and conversely, models with smaller values of \(N\) are better candidates for data with less sparsity. Similar
to simulation 1, to the trace plots and histograms in Figure 3 show that our mode converges better than the stationary distribution. Similarly, the hyperparameters are updated each 100 iterations.

**Table 2.** Results for Simulation 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE (sd)</th>
<th>FPR (sd)</th>
<th>FNR (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCG2</td>
<td><strong>0.0206 (0.0122)</strong></td>
<td>0.2100 (0.2638)</td>
<td>0.4100 (0.4312)</td>
</tr>
<tr>
<td>NCG10</td>
<td>0.0304 (0.0134)</td>
<td>0.2300 (0.2723)</td>
<td>0.4600 (0.5037)</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.0662 (0.0156)</td>
<td>0.2900 (0.4793)</td>
<td>0.4900 (0.5126)</td>
</tr>
<tr>
<td>aLasso</td>
<td>0.0401 (0.0160)</td>
<td>0.2100 (0.4320)</td>
<td>0.4300 (0.5252)</td>
</tr>
<tr>
<td>Enet</td>
<td>0.0699 (0.0208)</td>
<td>0.2600 (0.5912)</td>
<td>0.4200 (0.5523)</td>
</tr>
</tbody>
</table>

**Figure 2.** Trace plots (top) and histograms (bottom) for simulation 1.
4. Conclusion

We have shown that our proposed method for count data using normal-compound gamma prior for the scale mixture in the framework of normal-scale mixture perform very well compared to other existing models such as the Beta Prime prior (NCG2) [3], The Bayesian lasso, the Bayesian adaptive lasso and the Bayesian elastic net. We aim to study this method further in the future with different types of data such as censored and quantile data.

Figure 3. Trace plots (top) and histograms (bottom) for simulation 2.
References


