RESEARCH

Bayesian Shrinkage Priors in Zero-Inflated and Negative Binomial Regression models with Real World Data Applications of COVID-19 Vaccine, and RNA-Seq

Arinjita Bhattacharyya[1], Riten Mitra[1], Shesh Rai[1,2,3,4,5*] and Subhadip Pal[1]

*Correspondence:
shesh.rai@louisville.edu
[1]Department of Bioinformatics & Biostatistics, University of Louisville, KY, USA
Full list of author information is available at the end of the article

Abstract

Background: Count data regression modeling has received much attention in several science fields in which the Poisson, Negative binomial, and Zero-Inflated models are some of the primary regression techniques. Negative binomial regression is applied to modeling count variables, usually when they are over-dispersed. A Poisson distribution is also utilized for counting data where the mean is equal to the variance. This situation is often unrealistic since the distribution of counts will usually have a variance that is not equal to its mean. Modeling it as Poisson distributed leads to ignoring under- or overdispersion, depending on if the variance is smaller or larger than the mean. Also, situations with outcomes having a larger number of zeros such as RNASeq data require Zero-inflated models. Variable selection through shrinkage priors has been a popular method to address the curse of dimensionality and achieve the identification of significant variables.

Methods: We present a unified Bayesian hierarchical framework that implements and compares shrinkage priors in negative-binomial and zero-inflated negative binomial regression models. The key feature is the representation of the likelihood by a Polya-Gamma data augmentation, which admits a natural integration with a family of shrinkage priors. We specifically focus on the Horseshoe, Dirichlet Laplace, and Double Pareto priors. Extensive simulation studies address the efficiency of the model and mean square errors are reported. Further, the models are applied to data sets such as the Covid-19 vaccine, and Covid-19 RNA-Seq data among others.

Results: The models are robust enough to address variable selection, and MSE decreases as the sample size increases, having lower errors in $p > n$ cases. The noteworthy results showed that the adverse events of Covid-19 vaccines were dependent on age, recovery, medical history, and prior vaccination with a remarkable reduction in MSE of the fitted values. No. of publications of Ph.D. students were dependent on the no. of children, and the no. of articles in the last three years.

Conclusions: The models are robust enough to conduct both variable selections and produce effective fit because of their high shrinkage property and applicability to a broad range of biometric and public health high dimensional problems.

Keywords: shrinkage priors; negative binomial regression; horseshoe; Dirichlet Laplace; MCMC; Polya-Gamma; vaccine; RNASeq; Covid-19 vaccine; data augmentation

1 Introduction

The extension of linear models to generalized linear models (GLMs) introduced by [1] framework to handle data that are not typically modeled using a normal distribution (e.g., binary, count data) is a moment of immense success in the history of statistics. Most of the real-life datasets have number of explanatory variables (p) greater than the number of observations (n).The methodology for analyzing RNA sequencing data is rapidly expanding. Methods having application of shrinkage, flexibility of the designs, are highly in demand [2]. High-dimensional predictor selection and sparse signal recovery are routine statistical and machine learning practices.Sparsity relies on the property of a few large signals among many (nearly) zero noisy observations. A common goal in high-dimensional inference is to recover the low-dimensional signals observed in noisy observations[3]. The idea of global-local shrinkage hierarchies[4] has become the foremost research areas in Bayesian literature that incorporates heavy tailed prior distributions for coefficients in generalized linear regression models. In the exponential rise in the development of research in shrinkage priors, works that have gained mass popularity are [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] among many. An overview of several shrinkage priors with several data applications is given in [15]. Here we discuss the posterior simulation for negative binomial (NB) regression and Zero-inflated Negative binomial (ZINB) for count data. Our main focus is the utilization of the Polya-Gamma (PG) data augmentation strategy [16] which utilizes Polya-Gamma random variables to enhance posterior simulations. The performance of three different priors Horseshoe (H)[17],Dirichlet Laplace (DL) [14], Double Pareto (DP)[13] are measured and also the method is applied to benchmark data sets. Bayesian global-local (GL) shrinkage estimation is the state-of the art for Gaussian regression models, extension to non-standard regression techniques such as Poisson and NB are the ones that we concentrate in our methodology. Here two extensions of the global-local shrinkage framework is implemented. Firstly, the utilization of the PG data augmentation technique to generate simple algorithms for sampling with NB and ZINB regression likelihoods. Results show that the priors are highly competitive on the basis of mean square errors (MSE). Extensive simulation studies and real data applications are conducted to evaluate the performance of the these priors with respect to prediction accuracy and MSE for variable selection.

The rest of the section follows as section 2 describes in detail the tradition methods such as Poisson and NB and their Bayesian counterpart such as BNB and BZINB.Section3 explains the simulation details and parameters, followed by results 4. Real data scenarios are explained in Section 5, finally summarizing manuscript with a discussion 6.

2 Method

2.1 Poisson Regression

In modeling the number of times an event occurs a generalized linear model (GLM) such as Poisson or negative binomial regression is commonly applied.Let $y = (y_1, y_2, \ldots, y_n)$ denote the vector of n count measurements of a dependent variable of interest, and $x_i = (x_{i,1}, \ldots, x_{i,p})$ denote the vector of predictors (explanatory variables, covariates) associated with the response y_i. Let $X = (x_1, \ldots, x_n)$

be the $n \times p$ matrix of explanatory variables. The GLM contains Poisson distribution which has a probability function $f(y_i; \lambda_i) = \frac{\lambda_i^y e^- \lambda_i}{y_i!}$ where the mean and variance are equal. $E(y_i) = var(y_i) = \lambda_i$. The log-likelihood is given by $l(\lambda \mid y) = \sum_{i=1}^{n}(y_i log\lambda_i - \lambda_i - log(y_i!))$. In GLM , a non-linear transformation or a link function of the mean response λ_i is applied which is a linear function of the covariates X [18],[19]. The link function for Poisson regression is $log(\lambda_i) = \beta_0 + x_i^T\beta$, where β_0 is the intercept and β is the vector regression coefficients. Poisson distribution depending on single parameter λ is equi-dispersed. If the Poisson mean is assumed to have a random intercept term and this term enters the conditional mean function in a multiplicative manner, the following relationship is achieved [20]:

$$\begin{aligned}\lambda_i &= exp(\beta_0 + x_i^T\beta + \epsilon_i) \\ &= e^{(x_i^T\beta)}e^{(\beta_0+\epsilon_i)} \\ &= e^{(\beta_0+x_i^T\beta)}e^{\epsilon_i} \\ \lambda_i &= \mu_i\nu_i \end{aligned} \tag{1}$$

Here, $exp(\beta_0 + \epsilon_i)$ is defined as the random intercept; $mu_i = exp(\beta_0 + x_i^T\beta)$ is the log-link function between the mean $E(y_i)$ and the independent fixed covariate matrix X. The real data, however, often shows that the variance is larger than the mean, which is called overdispersion. The two-parameter negative binomial distribution has more flexible and can efficiently model overdispersed count data leading to correct standard errors and inferences [21]. Many parametric models for count data are obtained by additionally introducing a heterogeneity term in the Poisson model. Unobserved heterogeneity is usually included as a multiple of the Poisson mean. $y \mid \mu, \nu \sim Poi(\mu\nu)$ and the random heterogeneity term $\nu \geq 0$ is integrated out to obtain the distribution of $y \mid \mu$. These model structures are well-known as doubly stochastic Poisson by Cox [22] and a Cox process by Kingman [23]. In general, $E(\nu) = 1$ is the setting for several leading models. Different distributions of v leads to various generalizations of Poisson and here the Poisson-Gamma mixture is explained which leads to Negative Binomial distribution.

2.2 Negative Binomial as Poisson-Gamma Mixture

The NB model is derived from a Poisson-gamma mixture distribution [18]. The interpretation and derivation of NB from Poisson-Gamma is detailed in [24]. The heterogeneity parameter is assumed to have a Gamma distribution. $\nu_i \sim Gamma(d, b)$. The two-parameter Gamma distribution is represented as

$$k(\nu_i; d, b) = \frac{b^d}{\Gamma(d)} e^{-b\nu_i} {\nu_i}^{d-1} \tag{2}$$

$E(\nu_i) = \frac{d}{b}, Var(\nu_i) = \frac{d}{b^2}$, setting $E(\nu_i) = 1$ we get $d = b$, leading to a one-parameter Gamma distribution with $E(\nu_i) = 1, Var(\nu_i) = \frac{d}{d^2} = \frac{1}{d}$. Now the Poisson model $Poi(\mu\nu)$ can have easier interpretation if worked with the transformation $\lambda = \mu\nu$, i.e. $\nu = \frac{\lambda}{\mu}$. The Jacobian is obtained as $\frac{d\nu}{d\lambda} = \frac{1}{\mu}$. The probability density

function (p.d.f) of λ is then given as

$$\begin{aligned} g(\lambda \mid d, \mu) &= \frac{1}{\mu}\frac{d^d}{\Gamma(d)}\frac{\lambda}{\mu}^{d-1} e^{-\frac{\lambda}{\mu}d} \\ &= \frac{(\frac{d}{\mu})^d}{\Gamma d}\lambda^{d-1}e^{-\frac{\lambda}{\mu}d} \end{aligned} \tag{3}$$

The Poisson-gamma mixture is

$$\begin{aligned} h(y \mid \mu, d) &= \int f(y \mid \lambda)g(\lambda \mid d, \mu)d\lambda \\ &= \int \frac{e^{-\lambda}\lambda^y}{y!} \times \frac{(\frac{d}{\mu})^d}{\Gamma(d)}\lambda^{d-1}e^{-\frac{\lambda}{\mu}d} \\ &= \frac{(\frac{d}{\mu})^d}{\Gamma(d)y!}\int \lambda^{y+d-1}exp(-(1+\frac{d}{\mu})\lambda)d\lambda \\ &= \frac{(\frac{d}{\mu})^d}{\Gamma(d)\Gamma(y+1)}(1+\frac{d}{\mu})^{-(d+y)}\Gamma(d+y) \\ &= \frac{\Gamma(y+d)}{\Gamma(y+1)\Gamma(d)}(\frac{d}{\mu+d})^d(\frac{\mu}{\mu+d})^y \end{aligned} \tag{4}$$

The property of the gamma function is utilized in getting the above form: $\Gamma(m) = \int_0^\infty t^{m-1}e^{-t}dt$ for any $m \geq 0$ $\Gamma(m-1) = m!, c^{-m}\Gamma(m) = \int_0^\infty t^{m-1}e^{-ct}dt$ for any $c \geq 0$.
The equation (4) can also be represented as

$$\binom{y+d-1}{d-1}(\frac{d}{\mu+d})^d(\frac{\mu}{\mu+d})^y \tag{5}$$

Taking $\alpha = \frac{1}{d}$, we get the probability mass function (p.m.f) of the NB distribution as

$$f(y_i; \mu, \alpha) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})}(\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}}(1-\frac{1}{1+\alpha\mu_i})^{y_i} \tag{6}$$

Here, $E(y \mid \mu, \alpha) = \mu, Var(y \mid \mu, \alpha) = \mu(1+\alpha\mu) \geq \mu$, since $\alpha \geq 0$. The variance can also be represented as $\mu + \frac{\mu^2}{\phi}$, $\phi = d$ is the dispersion parameter. Most often it is expressed as $\alpha = \frac{1}{\phi} = \frac{1}{d}$, α is the parameter responsible for heterogeneity and models the over-dispersion amount in the data An alternative parametrization of NB used in many references as well as algorithms is

$$Pr(Y = y) = \binom{y+h-1}{h-1}(1-p)^h p^y, y = 0, 1, 2, \ldots, \tag{7}$$

where $0 < p < 1$ and $h \geq 0$. Then $E(y) = \frac{ph}{1-p}$ and $Var(Y) = \frac{ph}{(1-p)^2}$. Letting $h = \frac{1}{\alpha}, p = \frac{\alpha\mu}{1+\alpha\mu}$, yields the same parametrization of NB distribution in (6) and $h = d$ in (5).

The log-likelihood of NB can be expressed as

$$l(\mu; y, \alpha) = \Pi_{1=1}^{n} exp(log\Gamma(y_i + \frac{1}{\alpha}) - log\Gamma(y_i + 1) - log\Gamma(\frac{1}{\alpha}) + \frac{1}{\alpha} log(\frac{1}{1+\alpha\mu_i}) + y_i log(1 - \frac{1}{1+\alpha\mu_i})) \quad (8)$$

The mean of (7) is

$$E(y) = \frac{ph}{1-p}$$
$$E(\frac{y}{h}) = \frac{p}{1-p} \quad (9)$$

The link function in this parametrization of (7) is log-odds, i.e. $log(E(\frac{y_i}{h})) = log(\frac{p}{1-p}) = \beta_0 + x_i^T\beta$. So, the likelihood for the Negative Binomial distribution ignoring the constant term is

$$l(\beta; y, h) \propto \prod_{i=1}^{n} (1-p)^h p_i^y = \prod_{i=1}^{n} \frac{\exp(x_i^T \beta y_i)}{(1+\exp(x_i^T\beta))^{y_i+h}} \quad (10)$$

, where $p_i = \frac{exp(x_i^T\beta)}{1+exp(x_i^T\beta)}$. This β vector is assumed to include the intercept term, i.e. $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$.

2.3 Bayesian Negative Binomial Regression with Hierarchical Prior Structures (BNB)

The likelihood of NB (10) is not in a closed form and will require Polya-Gamma (PG) data augmentation. The auxillary variables will be sampled from Polya-gamma distribution with parameter $y_i + h$ and $x_i^T\beta$. The Polya-Gamma distribution is the exact distribution needed to augment this posterior for simulation thus obtaining closed form posterior distributions that can be easily handled via Gibbs sampling. The method is useful when modeling proportions on the log-odds scale. Binary logistic regression [**?**] and negative binomial regression (NB) are the two fore frontiers that meets the criteria. To facilitate posterior sampling, we introduce a set of auxillary variables that follow Polya-Gamma distribution that are represented as scale mixtures of normals. Conditional on the latent variables, inference proceeds via straightforward Gibbs sampling. A Polya-Gamma variable, $w \sim PG(b, c)$ with $b > 0$ and $c \in \mathbb{R}$, can be defined as follows.

$$w = \frac{1}{2\pi^2} \sum_{i=1}^{\infty} \frac{Z_i}{(k-\frac{1}{2})^2 + \frac{c^2}{4\pi^2}}; Z_i \overset{i.i.d}{\sim} Gamma(b, 1) \quad (11)$$

The variable distribution is similar to Gamma distribution and, as b increases, it becomes approximately normal [25]. So here $Y_i \sim NB(h, \frac{exp(x_i^T\beta)}{1+exp(x_i^T\beta)})$ and $W_i \sim PG(h, |x_i^T\beta)$. Similar, to logistic regression case, the joint posterior density of the parameter β and W with prior $\pi(\beta)$ is obtained as

$$\pi(\beta, w \mid Y) = \frac{\pi(\beta)}{c(Y)} f(w \mid \beta) \prod_{i=1}^{n} \frac{\exp(x_i^T \beta y_i)}{1+\exp(x_i^T\beta)} \quad (12)$$

Here $c(Y)$ is the marginal distribution of Y. We can see that

$$\int \pi(\beta, w \mid Y)dw = \pi(\beta \mid Y)$$

which is our targeted density. The conditional independence of Y_i and W_i implies that $\pi(w \mid \beta, y) = f(w \mid \beta)$. Thus, we can draw from $\pi(w \mid \beta, y)$ by making n independent draws from the Polya-Gamma distribution. The other conditional density $\pi(\beta \mid w, y)$ is multivariate normal.

A hierarchical representation of the Horseshoe prior [9] is stated as

$$\begin{aligned}
\beta_j \mid \Lambda_j^2\tau^2 \sim N_p(0, \Lambda_j^2\tau^2), j = 1, 2, ..., p \\
\Lambda_j^2 \mid \gamma_j \sim IG\left(\frac{1}{2}, \frac{1}{\gamma_j}\right) \\
\tau^2 \mid \xi \sim IG\left(\frac{1}{2}, \frac{1}{\xi}\right) \\
\gamma_1, \gamma_2, ..., \gamma_p, \xi \sim IG\left(\frac{1}{2}, 1\right)
\end{aligned} \tag{13}$$

Here, Σ of the distribution of β is a diagonal matrix with elements $\left(\Lambda_1\tau^2, \Lambda_2\tau^2, ..., \Lambda_p\tau^2\right)$. From equations (12), and the above hierarchical prior structure, the full posterior distribution is given by:

$$\begin{aligned}
\pi\left(\beta, \Lambda_j^2, w_i, \gamma_j, \tau^2, \xi \mid Y\right) \propto \prod_{i=1}^{n} \frac{\exp(x_i^T\beta y_i)}{(1+\exp(x_i^T\beta))^h} (1+\exp(x_i^T\beta))^h h\left(w_i\right) \\
\exp(-\frac{\left(x_i^T\beta\right)^2 w_i}{2}) \frac{\exp(-\frac{1}{2}(\beta^T\Sigma^{-1}\beta))}{\sqrt{2\pi\Sigma}} \\
\prod_{j=1}^{p} \frac{(\Lambda_j^2)^{-(\frac{1}{2}+1)} \exp(\frac{-1}{\gamma_j\Lambda_j^2})}{\gamma_j^{\frac{1}{2}}} \frac{(\tau^2)^{-(\frac{1}{2}+1)} \exp(\frac{-1}{\tau^2\xi})}{\xi^{\frac{1}{2}}} \\
\gamma_j^{-(\frac{1}{2}+1)} \exp(\frac{-1}{\gamma_j}) \; \xi^{-(\frac{1}{2}+1)} \exp(\frac{-1}{\xi})
\end{aligned} \tag{14}$$

where $h(w_i)$ is obtained from

$$h(w) = \sum_{k=0}^{\infty} (-1)^k \frac{2k+1}{\sqrt{2\pi w}} \exp(-\frac{(2k+1)^2}{8w}), 0 < w < \infty \tag{15}$$

The conditional distributions required for our analysis follows:
The conditional density of β given y, w is

$$\pi\left(\beta \mid \Sigma, W_D, Y\right) \sim N_p\left(\left(X^T W_D X + \Sigma^{-1}\right)^{-1} X^T y^*, \left(X^T W_D X + \Sigma^{-1}\right)^{-1}\right) \tag{16}$$

where, W_D and Σ are diagonal matrices where the elements are $(w_1, w_2, ..., w_n)$, $(\Lambda_1^2\tau^2, ..., \Lambda_p^2\tau^2)$ respectively and, $y^* = \left(\frac{y_1-h}{2}, ..., \frac{y_n-h}{2}\right)$.

The conditional density of w_i given x_i, β is

$$\pi\left(w_i \mid \beta\right) \sim PG\left(y_i + h, x_i^T \beta\right) \tag{17}$$

The conditional density of the hyper-parameters are as follows

$$\begin{aligned}
\pi(\Lambda_j^2 \mid \gamma_j, \beta_j, \xi, \Lambda_j^2) &\sim IG\left(1, \frac{1}{\gamma_j} + \frac{\beta_j^2}{2\tau^2}\right) \\
\pi(\gamma_j \mid \Lambda_j^2, \beta_j, \xi, \tau^2) &\sim IG\left(1, 1 + \frac{1}{\Lambda_j^2}\right) \\
\pi(\tau^2 \mid \gamma_j, \beta_j, \xi, \Lambda_j^2) &\sim IG\left(\frac{p+1}{2}, \frac{1}{\xi} + \sum_{j=1}^{p} \frac{\beta_j^2}{2\Lambda_j^2}\right) \\
\pi(\xi \mid \tau^2) &\sim IG\left(1, 1 + \frac{1}{\tau^2}\right)
\end{aligned} \tag{18}$$

Again, here all the posterior densities are in the closed form, and follow simple densities like Normal, Polya-Gamma and Inverse-Gamma making sampling from them trivial. Exploiting the scale-mixture representation of the global-local shrinkage priors, it is straightforward to formulate the Gibbs sampler.
The hierarchical structure of the Dirichlet Laplace prior Bhattacharya:2015 is

$$\begin{aligned}
\beta_j &\sim N_p\left(0, \psi_j \phi_j^2 \tau^2\right), \\
\psi_j &\sim exp\left(\frac{1}{2}\right) \\
\phi &\sim Dir\left(a, a, ..., a\right) \\
\tau &\sim G\left(pa, \frac{1}{2}\right)
\end{aligned} \tag{19}$$

The conditional posterior distributions remain same for $\beta \mid y_i$ and $w_i \mid \beta$ is similar to that of equations (16) and (17).
The conditional density of the hyper-parameters as obtained similar to Theorem 2.2 in [14] are as follows:

$$\begin{aligned}
\pi(\psi \mid \phi, \tau, \beta) &\sim IG\left(\frac{\phi_j \tau}{|\beta_j|}, 1\right) \\
\pi(\tau \mid \phi, \beta) &\sim GIG\left(pa - p, 1, 2\sum_{j=1}^{p} \frac{|\beta_j|}{\phi_j}\right)
\end{aligned} \tag{20}$$

To sample $\pi(\phi \mid \beta_j)$ sample $T_j \sim GIG\left(a - 1, 1, 2|\beta_j|\right)$, set $\phi_j = \frac{T_j}{T}, T = \sum_{j=1}^{p} T_j$. where $GIG(a, b, c)$ is the Generalized Inverse Gaussian distribution with density $f(x; a, b, c) \propto x^{(c-1)} e^{\frac{-1}{2}(ax + \frac{b}{x})}$.

The hierarchical structure of Double Pareto prior[26] is

$$\begin{aligned} \beta \mid \Lambda, \tau &\sim N_p(0, D_\tau), \\ w_i \mid x_i, \beta &\sim PG\left(1, x_i^T \beta\right), \\ \tau_j \mid \Lambda_j &\sim exp(\frac{\Lambda^2}{2}) \\ \Lambda_j &\sim G(\zeta, \eta) \end{aligned} \tag{21}$$

Again, the conditional densities of $\beta \mid y_i$ and $w_i \mid \beta$ remains same as (**??**). Here $\Sigma = D_\tau$ is a diagonal matrix with elements $(\tau_1, \tau_2, ..., \tau_p)$.

The conditional density of rest of the hyper-parameters are as follows:

$$\begin{aligned} \pi\left(\tau_j \mid \beta, \Lambda, y\right) &\sim GIG\left(\frac{1}{2}, \Lambda_j^2, \beta_j^2\right) \\ \pi\left(\Lambda \mid \beta, y\right) &\sim Gamma\left(\zeta + 1, \eta + |\beta_j|\right) \end{aligned} \tag{22}$$

2.4 Bayesian Zero-Inflated Model with Hierarchical Prior Structures (BZINB)

Zero-inflated models represents the excess zeros, and a count distribution for the remaining values, thus forming a mixture of zeros. The model is very useful when there is an excess number of two types of zeros in the concerned response variable. By construction, zero-inflated models partition zeros into two types. The first type, typically referred to as a "structural" zero, corresponds to individuals who are not at risk for an event, and therefore have no opportunity for a positive count. The second type, termed the "at-risk" or "chance" zero, applies to a latent class of individuals who are at risk for an event but nevertheless have an observed response of zero[27]. For example, in our application with Covid-19 vaccine data set, examining the number of adverse events, the structural zeros might represent patients who had no adverse event thus have no recorded adverse event. In contrast, the at-risk zeros might correspond to patients with a single occurrence of adverse event which has been determined not clinically significant , thus contributes to at-risk zero. Similarly for RNA Seq datasets, where the genes are counts containing a high proportion of zeros, the zero-inflated models can be viewed as latent class models in which the classes are formed by the two types of zeros. The zero-inflated model has two parts that models consisting of negative binomial distribution, and the logit distribution.

$$f(y_i) = p_i I_{(w_i=1)} NB(\mu_i, r) + (1 - p_i) I_{(w_i=0)} g(y_i = 0) \tag{23}$$

$$f(y_i | r, \beta, w_i = 1) = \frac{\Gamma(y_i + r)}{\Gamma(r) y_i!} (1 - \phi_i)^r \phi_i^{y_i} \tag{24}$$

,where $\phi_i = \frac{exp(x_i^T \beta)}{1 + exp(x_i^T \beta)}$ where $NB(\mu_i, r)$ takes the form of equation 6. Here y_i is the count response for the ith individual. The latent indicator variable wi that takes values 1 and 0 with probabilities p_i and $1 - pi$ where $y_i \sim NB(\mu_i, r)$ and $y_i = 0$. The indicatorwi variable is binary, thus modelled with logistic regression as follows:

$$logit(p_i) = logit[P(w_i = 1 | \alpha)] = x_i^T \alpha \tag{25}$$

The negative binomial distribution has likelihood similar to equation 10 and the posterior distribution of the coefficients β are obtained from equation 12. With the hierarchical prior structures on the coefficients β from section 2.3, The posterior distribution for both *alpha* and β are modelled similarly as 2.3 citebhattacharyya2021applications, Neelon:2019.

3 Simulation

Here each of the simulation structure along with their parameters are defined.

3.1 BNB

The data is generated utilizing the Poisson-gamma mixture representation of NB as expressed in (4). In the data generation process, the true values of β, are defined. The data is generated as follows:

1 Calculate $\mu_i = exp(\beta_0 + x_i^T \beta)$
2 Get $\lambda_i = \frac{\mu_i}{h}\nu_i$, where $\nu_i \sim Gamma(h, 1)$
3 Generate $y_i \sim Poi(\lambda_i)$

The covariates are generated from multivariate normal distribution with mean vector 0 and covariance matrix Σ. 80% of the data set is reserved for the training set, and 20% for the test data set.

- S1: $n = 200, p = 10$ with a correlation of about $\rho = 0.5$ among the covariates. The coefficient vector is given by $\beta = (0.5, 0.5, -0.5, 0.5, -0.6, \underbrace{0, \ldots, 0}_{5})^T$ with 5 non-zero coefficients.
- S2: $n = 120, p = 10, \rho = 0.2$ Five non-zero coefficients. $\beta = (0.5, 0.5, -0.5, 0.5, -0.6, \underbrace{0, \ldots, 0}_{5})^T$.
- S3: $n = 50, p = 10$, coefficients similar to S2.
- S4: $n = 100, p = 20$ $\rho = 0.1, \beta = (\underbrace{0, \ldots, 0}_{5}, \underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0.3, \ldots, 0.3}_{5})$.
- S5: $n = 50, p = 500, \beta = (log(1.75), log(1.75), -log(1.75), -log(1.75), -log(1.75), \underbrace{0, \ldots, 0}_{495})^T$

Here we have considered two broad settings of design matrices comprising $p = 200$ and $p = 500$ covariates. The sampling of β for $p > n$ with $p = 500$ is conducted from a Gaussian distribution and follows the fast sampling algorithm of [28]. For prediction accuracy, the mean squared error (MSE) is used as a prediction accuracy criteria, which is defined as $\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}$. In terms of variable selection performance, the number of the truly nonzero coefficients which are incorrectly set to zero (FP), and the number of the true zero coefficients which are correctly set to zero (FN). The higher the values of FP, and the lower the values of FN, the better the variable selection performance is. The variable selection is determined by posterior credible intervals. For each simulation setting, 100 datasets were generated, and MSE, results were calculated by averaging over these 100 datasets. MCMC is used for sampling which relies on block-updating the sets of parameters. The no. of simulation runs is 10000 with 6000 as burn-in. Trace plots, auto-correlation plots, Geweke z-statistics [29] were some of the diagonistic criterion. R pacakge coda [30] was used to conduct the posterior sample diagnosis and pgdraw [31] [32] was used

to generate the Polya-Gamma random variates. All the computations have been carried out in RStudio. The R package coda [30] was used to conduct the posterior sample diagnosis and pgdraw [31] [32] was used to generate the Polya-Gamma random variates.

3.2 BZINB

The following simulation scenarios were set up to understand the behaviour of the model. We followed similar data generation process as of [27].

- Sim1: $n = 1000, p = 4, \alpha = (0.5, 0.5, -0.25, 0.25), \beta = (0.5, -1.00, 0.75, -0.25), r = 1$
- Sim2: $n = 500, p = 10, \alpha = (0.5, 0.5, -0.25, 0.25, 0.5, 0.5, -0.25, 0.5, 0.5, -0.25), \beta = (0.5, 0.5, 0.75, -0.25, 0.5, 0.75, 0.25, 0.75, -0.25), r = 1$
- Sim3: $n = 400, p = 500, \alpha = \beta = (\underbrace{0.25, \ldots, 0.25}_{100}, \underbrace{-0.25, \ldots, -0.25}_{400}), r = 4$

4 Simulation Results

Table 1 shows the variable selection performance and MSE along with their standard deviations for these simulation scenarios for BNB model. Comparing S1, S2, and S3 accuracy and sensitivity were similar with the decrease in sample size for all the three priors. The MSE for β and their standard errors increases for all the three priors as the sample size decreases. From S1 and S2 and S3 are three scenarios having similar settings with the N/P ratio changing, there is a clear trend that with the decrease in sample size increases the MSE, S1 and S2 has 100% sensitivity implying all the non-zero β can be identified. There is not much of an impact of correlation on the results. S5 has 1% non-zero β with low effect sizes but almost 67% of the time they are identified by the 2 priors DL and Horseshoe. DP fails to identify the non-zeros, though keeping the MSEs low.

In summary, it is obvious that the simulation results has demonstrated the use of these three priors unanimously with the low standard errors across datasets. Trace plots, Geweke diagnostics and Monte Carlo standard errors were indicative of convergence and showed reasonable mixing fora range of model parameters. We tested the sampler for various settings of n, p and coefficients. Comparing S2, and S3 the MSE for β and their standard errors increases for all the three priors and drastically reduces the MSE of the fitted values as well. The increase in correlation in S4 than S3 doesn't seem have much effect. As *nandp* increases from S1 to S2 there is a substantial decrease in MSE of β but not for the fitted values. S5, S7 and S12 compare different sample size under similar coefficient and simulation settings. MSE and their standard error decreases as their sample size increases. S9, S10 and S11 again compare sample sizes $n = 200, 100, 8$ for a different set of coefficient settings. Here all the βs were non-zero. Again the decrease in MSE with the increase of sample size is noticeable and also in the MSE of fitted values. In the all the above $n > p$ scenarios we see that Horseshoe, and Double Pareto perform comparatively better than Dirichlet Laplace. For $p > n$ scenarios S13 and S6, with the increase in the n/p ratio increases MSE increases. In summary, it is obvious that the simulation results has demonstrated the use of these three priors unanimously. We build on the R codes from this package and extend it with the three different prior structures.

In BZINB model, Horseshoe prior clearly outperforms the other two priors across simulation scenarios. The Sim3 $p > n$ case has similar performance across priors. The variable selection performances are also measured with the MSEs and all the three priors performed reasonably Table 2.

5 Real Data Application

Here three real data applications are considered. In all the applications the count data and the excess amounts of zeros in the outcome made BNB and BZINB regression a best fit for the data. The variables selected by the three priors for BNB and BZINB , Poisson regression and NB models for each of the datasets is in the table 3. For the three priors, the starting values are chosen as the maximum likelihood estimates from NB Regression. The MSE for the coefficients for the tree priors are obtained by $\sum_{i=1}^{p} \frac{\left(\beta_i - \hat{\beta}_i\right)^2}{p}$, p is the number of coefficients, β_i: estimates obtained from NB regression and $\hat{(\beta_i)}$ are the posterior mean estimates. This MSE measure can also be interpreted as a departure from the frequentist estimates.

5.1 No. of PhD publications

The first one deals with the number of publications produced by Ph.D. biochemists of [33]. It is available in the R package "pscl"[34].The response variable is the number of articles in the last three years of Ph.D. Five explanatory variables were used. They are: the gender (x_1), the marital status (x_2), the number of children under age six (x_3), prestige of Ph.D. program (x_4), and the number of articles by the mentor in last three years (x_5). In this application, the response variable is following NB distribution. (x_3) and (x_5) were selected by the Bayesian methods. The NB, poisson followed by BNB methods were suitable for the data set, as per the MSE of fitted values.

5.2 Nuts data

In the second real data application, we considered the nuts dataset [18]. Here, $n = 52$ and $p = 7$. The nuts dataset defines the squirrel behavior and several features of the forest across different plots in Scotland's Abernathy Forest. It is available in the R package "COUNT"[35].The response variable is the number of cones stripped follows negative binomial distribution. The explanatory variables are: the number of trees per plot (x_1), the number of DBH per plot (x_2), mean tree height per plot (x_3), canopy closure (as a percentage) (x_4), standardized number of trees per plot (x_5), standardized mean tree height per plot (x_6), standardized canopy closure (as a percentage) (x_7). Here we use x_5, x_6, x_7.

The variables chosen by all the methods except NB were x_5, x_6, x_7. So all the models seem reasonable except NB. The variables selected in the first two datasets (Biochemists, and NUTS) are included withing the set of the variables selected either by the traditional Poisson or NB regression.

5.3 US National Medical Expenditure Survey

The third data set originated from the US National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. The NMES is based upon a sample of the civilian non-institutionalized population and individuals admitted to long-term care

facilities during 1987. The data are a sub-sample of individuals ages 66 and over all of whom are covered by Medicare (a public insurance program providing substantial protection against health-care costs). It is available in the R package "AER"[36]. It is a data frame containing 4,406 observations on 19 variables. The response variable considered is the number of physician office visits among other type of count variables present (emergency visits, no. of non-physician hospital outpatient visits etc.). We have considered 14 dependent variables hospital: number of hospital stays x_1; health:self-perceived health status x_2, levels are "poor", "average", "excellent"; chronic:number of chronic conditions x_3; adl: indicator of whether individual having limits in activities of daily living x_4 levels: "limited","normal"; region: indication region of the individual x_5 levels northeast, midwest, west, other; age x_6; afam (race): If African-American x_7; gender:male/female x_8; married: marital status x_9; school:number of years of education x_{10}; income (USD) x_{11}; employment x_{12}; insurance: whether the individual is covered by private insurance yes/No x_{13}; medicaid x_{14}. The MSE for the fitted values are also given. Figure 1 the posterior distribution of the 14 variables and their respective confidence intervals. For the NMES data, the three priors selects 10 important out of 14 variables, Poisson selects 13 and NB 8 of them. It seems that the three priors perform better than NB where they do include the relevant variables such as limited activity level x_4 and race x_7 but doesn't do over-fitting such as Poisson.

5.4 PBMC RNA-Seq data

The data set that is analyzed here is taken from https://satijalab.org/seurat/articles/pbmc3ktutorial.html. The RNASeq data is analyzed as a per gene model where each gene is the outcome and the cell type as the covariate. The top 10 genes selected by BNB and BZINB are in Table 3. The genes common between H and DL are 2742, 6369, 9338, 3665,7203,11458, 2217, 4360,4368.The genes common between DL and DP are 2742,13706,6369,9338,11458,7203, 2217, and 4368. The genes common between DP and H are 2742,6369,9338,3665,11458,7203,2217,3665,4368,4368. The genes common between three priors are 2742, 6369, 9338,7203,11458,3665, 2217, 4368. The common genes between H and DL are 74, 267, 1804,3417,6049,8146. The genes that belong to the intersection of DL and DP are 74, 267,1804, 3417, and 6049. The genes common between DP and H are 74, 267, 1804, 3417, and 6049. The genes common between the three priors are 74, 267, 1804, 3417, and 6049. The genes selected by both the models are selected by ranking them by the least MSE.

5.5 Covid-19 Vaccine Data

This data set consisted of the Covid-19 vaccine administered over the year 2021 and the symptoms (adverse events) gathered from the administration of vaccines. This data is a part of the Vaccine Adverse Event Reporting System (VAERS) which was created by the Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC) to receive reports about adverse events that may be associated with vaccines.The variables that are considered are age (x_1), sex (x_2), if there is life threat or not (x_3), if there was emergency room visit (x_4), if hospitalized (x_5), number of hospital days (x_6), no. of extended stay (x_7), disability status (x_8), recovery status (x_9), medical history (x_{10}), other medications (x_{11}), laboratory data

(x_{12}), disease during vaccination (x_{13}), prior vaccination status (x_{14}), allergy status (x_{15}), doctor's office visit (x_{16}), emergency visit (x_{17}), vaccination route (x_{18}), vaccination dose (x_{19}), vaccine manufacturer (x_{20}).The dimension of the cleaned data set post-processing was having 100 samples and 20 variables. The number of days between vaccination and onset of adverse symptoms is treated as the response variable (y) that had a median of 1 day, mean of 7 days, maximum minimum of 0, and 2.43 years respectively. The results from the BNB model, BZINB model, and both these models with respective priors are given in Table 3. We applied the region of practical equivalence (ROPE) method [37] was utilized to select the variables after obtaining the posterior samples. The variables that were significant by the ROPE method with 5% cut-off for ROPE by mainly the Horseshoe and the DP priors for both the BNB and BZINB models belonged to the super set of the set of the variables. The variables age (x_1), sex (x_2), if there is life threat or not (x_3), if there was emergency room visit (x_4), no. of extended stay (x_7), other medications (x_{11}), laboratory data (x_{12}), disease during vaccination (x_{13}), prior vaccination status (x_{14}), allergy status (x_{15}) belonged to the intersection of the two sets of variables identified by the BNB and BZINB models with Horseshoe prior, which was a significant overlap along with matching the variables identified by the traditional methods without shrinkage priors. In general, the BZINB with DP and Horseshoe priors were able to select more variables than the BNB model. Specifically, variables such as medical history (x_{10}), prior vaccination status (x_{14}), vaccination route (x_{18}), vaccine manufacturer (x_{20}) were some of the interesting features that seem to influence the results. All the shrinkage prior models surpassed the traditional models in their MSE of fitted values.

5.6 Covid-19 RNASeq Data

The data is taken from the article [38]. We selected the GSE152075 dataset from Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/gds), which contained RNA-seq data from 430 SARS-CoV-2 positive and 54 negative patients [**?**]. The data is analyzed similarly with the help of per gene model where each RNA-seq (count variable) is modelled against the covid-19 positivity status which is a binary variable. Upon pre-processing the following top 10 genes were selected by ranking the genes with the lowest MSE with the BZINB and BNB methods are in Table 3. The BZINB method was able to select about 30 genes that were not selected by the top BNB method. With the BNB method,there were 4 genes commonly selected by DL and H which are 4903, 7767, 11222, and 12202.The common genes between DL and DP are 8523,12202,13109 and between DP and H are 8998, 9016, 9856, 12202. Gene no. 12202 was selected by all three priors with BNB model.

6 Discussion

Our simulations showed that the approach performs well across a range of scenarios. The numerical results provide additional numerical and theoretical insights into the properties of global-local shrinkage priors including high-dimension case. Variable selection is a very helpful procedure for improving computational speed and prediction accuracy by identifying the most important variables that related to the response variable. The number of counts for each observation if large can make

the sampler perform poorly as the Polya-Gamma augmentation will not be efficient as it require generation of Polya-Gamma random variates equal to the number of observations[16]. The generation of such random variates is also time consuming.The Polya-Gamma procedure is in general fast, easy to implement and flexible. The rate at which one can generate Polya-Gamma random variates is a key factor in the efficiency of the Polya-Gamma scheme; hence, building fast samplers is essential. R packages such as "bayesreg"[31] deals with high-dimensional Bayesian regularised regression. Alternative models such as the BNB model, BZINB models [27], truncated models, or quantile count models provide potential future research guidelines. The use of shrinkage priors with next generation sequencing data such as RNA-Seq data with Zero-Inflated or Negative Binomial models are also areas that needs further exploration [2]. All the three priors have their own advantage and caveats. A computationally efficient Bayesian approach for variable selection is proposed here that performs quite well in simulation scenarios and provide consistent results in these different case studies. One disadvantage of data augmentation schemes is that the number of latent variables is of the order of sample size. Hence for large n, the computation can slow down, as we have seen in settings with sample sizes greater than 1000. However, this disadvantage is offset by the gains we have over the traditionally used Metropolis-Hastings, which requires choosing proposal distributions and would probably generate a considerable number of rejection steps. Additionally, the model can be extended with other shrinkage priors, and other models such as hurdle models. More generally, the proposed method can be applied in scenarios where interest lies in modeling count data within a Bayesian inferential framework and exhaustive comparison of existing shrinkage priors in the literature. We believe that the rigorous, yet simple and systematic nature of Bayesian inference coupled with the latest advances in technology in high dimensional and next generation sequencing with RNA Seq data might strongly help in contributing to expanding the research field.

7 Acknowledgements

This work was supported by the National Institute of Health grant P42 ES023716 to principal investigator: Dr S Srivastava and the National Institute of Health grant 1P20 GM113226 to principal investigator: Dr C McClain. Dr. Rai was also partially supported by Wendell Cherry Chair in Clinical Trial Research.

Funding
This work was supported by the National Institute of Health grant P42 ES023716 to principal investigator: Dr S Srivastava and the National Institute of Health grant 1P20 GM113226 to principal investigator: Dr C McClain. Dr. Shesh Rai was also partially supported by Wendell Cherry Chair in Clinical Trial Research.

Abbreviations
SSVS - Stochastic Search Variable Selection; GL- global-local; HS- Horseshoe; DL- Dirichlet Laplace; DP - Double Pareto; PG - Polya-Gamma; DA - Data-Augmentation; MCMC- Markov Chain Monte Carlo; MSE - Mean Squared Error; VS - variable selection; BZINB- Bayesian Zero-Inflated Negative Binomial; BNB - Bayesian Negative Binomial; NB- Negative Binomial; ZINB- Zero-Inflated Negative Binomial;

Availability of data and materials
The datasets used and/or analysed are publicly available and information about it is included in this article.
Not applicable

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable

Authors' contributions
A.B. has contributed to the methodology, data collection, analysis, and writing of the manuscript. R.M. and S.P. has contributed to the methodology. S.N.R. has contributed to developing ideas, analysis and valuable comments. All authors have contributed to the final preparation of the manuscript.

Author details
[1]Department of Bioinformatics & Biostatistics, University of Louisville, KY, USA. [2] Biostatistics & Bioinformatics Facility, JG Brown Cancer Center, University of Louisville, KY, USA. [3] The Christina Lee Brown Envirome Institute, University of Louisville, KY, USA. [4] University of Louisville Alcohol Research Center, University of Louisville, KY, USA. [5] University of Louisville Hepatobiology & Toxicology Center, University of Louisville, KY, USA.

References

1. Nelder, J.A., Wedderburn, R.W.: Generalized linear models. Journal of the Royal Statistical Society: Series A (General) **135**(3), 370–384 (1972)
2. Van De Wiel, M.A., Leday, G.G., Pardo, L., Rue, H., Van Der Vaart, A.W., Van Wieringen, W.N.: Bayesian analysis of rna sequencing data by estimating multiple shrinkage priors. Biostatistics **14**(1), 113–128 (2013)
3. Bhadra, A., Datta, J., Polson, N.G., Willard, B., *et al.*: Lasso meets horseshoe: A survey. Statistical Science **34**(3), 405–427 (2019)
4. Polson, N.G., Scott, J.G.: Shrink globally, act locally: Sparse bayesian regularization and prediction. Bayesian statistics **9**, 501–538 (2010)
5. Bae, K., Mallick, B.K.: Gene selection using a two-level hierarchical bayesian model. Bioinformatics **20**(18), 3423–3430 (2004)
6. Bhadra, A., Datta, J., Polson, N.G., Willard, B., *et al.*: The horseshoe+ estimator of ultra-sparse signals. Bayesian Analysis **12**(4), 1105–1131 (2017)
7. Griffin, J., Brown, P., *et al.*: Hierarchical shrinkage priors for regression models. Bayesian Analysis **12**(1), 135–159 (2017)
8. Ishwaran, H., Rao, J.S., *et al.*: Spike and slab variable selection: frequentist and bayesian strategies. The Annals of Statistics **33**(2), 730–773 (2005)
9. Makalic, E., Schmidt, D.F.: A simple sampler for the horseshoe estimator. IEEE Signal Processing Letters **23**(1), 179–182 (2015)
10. Park, T., Casella, G.: The bayesian lasso. Journal of the American Statistical Association **103**(482), 681–686 (2008)
11. Piironen, J., Vehtari, A.: On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. arXiv preprint arXiv:1610.05559 (2016)
12. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
13. Armagan, A., Dunson, D.B., Lee, J.: Generalized double pareto shrinkage. Statistica Sinica **23**(1), 119 (2013)
14. Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B.: Dirichlet–laplace priors for optimal shrinkage. Journal of the American Statistical Association **110**(512), 1479–1490 (2015)
15. Van Erp, S., Oberski, D.L., Mulder, J.: Shrinkage priors for bayesian penalized regression. Journal of Mathematical Psychology **89**, 31–50 (2019)
16. Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using pólya–gamma latent variables. Journal of the American statistical Association **108**(504), 1339–1349 (2013)
17. Carvalho, C.M., Polson, N.G., Scott, J.G.: The horseshoe estimator for sparse signals. Biometrika **97**(2), 465–480 (2010)
18. Hilbe, J.M.: Negative Binomial Regression. Cambridge University Press, ??? (2011)
19. McCullagh, P.: Generalized Linear Models. Routledge, ??? (2018)
20. Cameron, A.C., Trivedi, P.: Regression analysis of (1998)
21. Lehman, R.R., Archer, K.J.: Penalized negative binomial models for modeling an overdispersed count outcome with a high-dimensional predictor space: Application predicting micronuclei frequency. PloS one **14**(1) (2019)
22. Cox, D.R.: Some statistical models related with series of events. Journal of the Royal Statistical Society Series B **17**, 129–164 (1955)
23. Kingman, J.: On doubly stochastic poisson processes. In: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 60, pp. 923–930 (1964). Cambridge University Press
24. Greenwood, M., Yule, G.U.: An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. Journal of the Royal statistical society **83**(2), 255–279 (1920)
25. Glynn, C., Tokdar, S.T., Howard, B., Banks, D.L., *et al.*: Bayesian analysis of dynamic linear topic models. Bayesian Analysis **14**(1), 53–80 (2019)
26. Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association **88**(422), 669–679 (1993)
27. Neelon, B., *et al.*: Bayesian zero-inflated negative binomial regression based on pólya-gamma mixtures. Bayesian Analysis **14**(3), 849–875 (2019)
28. Bhattacharya, A., Chakraborty, A., Mallick, B.K.: Fast sampling with gaussian scale mixture priors in high-dimensional regression. Biometrika, 042 (2016)
29. Geweke, J., *et al.*: Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments vol. 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, ??? (1991)
30. Plummer, M., Best, N., Cowles, K., Vines, K.: Coda: convergence diagnosis and output analysis for mcmc. R news **6**(1), 7–11 (2006)

31. Makalic, E., Schmidt, D.F.: High-dimensional bayesian regularised regression with the bayesreg package. arXiv preprint arXiv:1611.06649 (2016)
32. Makalic, E., Schmidt, D.: High-Dimensional Bayesian Regularised Regression with the BayesReg Package. arXiv:1611.06649v3
33. Long, J.S.: The origins of sex differences in science. Social forces **68**(4), 1297–1316 (1990)
34. Jackman, S., Tahk, A., Zeileis, A., Maimone, C., Fearon, J., Meers, Z., Jackman, M.S., Imports, M.: Package 'pscl'. See http://github. com/atahk/pscl (2017)
35. Hilbe, J.M.: COUNT: Functions, Data and Code for Count Data. (2016). R package version 1.3.4. https://CRAN.R-project.org/package=COUNT
36. Kleiber, C., Zeileis, A., Zeileis, M.A.: Package 'aer'. R package version 1.2 **4** (2020)
37. Kelter, R.: Analysis of bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Medical Research Methodology **20**(1), 1–18 (2020)
38. Sanchis, P., Lavignolle, R., Abbate, M., Lage-Vickers, S., Vazquez, E., Cotignola, J., Bizzotto, J., Gueron, G.: Analysis workflow of publicly available rna-sequencing datasets. STAR protocols **2**(2), 100478 (2021)

Figures

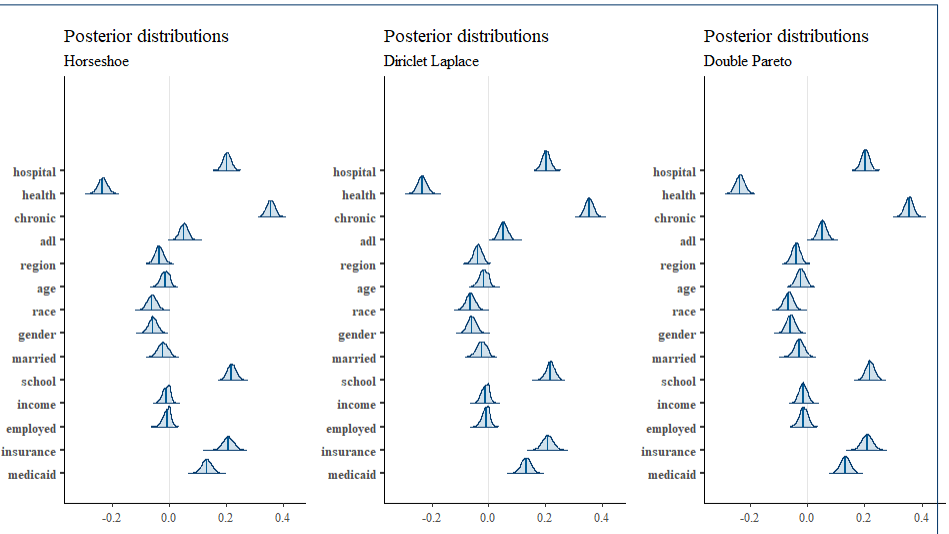

Figure 1: Posterior distribution with 95% credible intervals: NMES Data

Tables

Simulation Scenarios					
Priors	s1	s2	s3	s4	s5
N,P	200,10	120,10	50,10	100,20	50,500
Accuracy					
Horseshoe	0.993(0.026)	0.996(0.020)	0.958(0.068)	0.702(0.048)	0.827(0.066)
Dirichlet Laplace	0.991(0.029)	0.989(0.035)	0.962(0.065)	0.722(0.053)	0.829(0.064)
Double Pareto	0.992(0.027)	0.997(0.017)	0.956(0.073)	0.702(0.046)	0.832(0.071)
Sensitivity					
Horseshoe	1.000(0.000)	1.000(0.000)	0.936(0.110)	0.514(0.088)	0.662(0.123)
Dirichlet Laplace	1.000(0.000)	1.000(0.000)	0.950(0.096)	0.563(0.094)	0.668(0.114)
Double Pareto	1.000(0.000)	1.000(0.000)	0.932(0.114)	0.516(0.084)	0.012(0.003)
Specificity					
Horseshoe	0.986(0.051)	0.992(0.039)	0.980(0.072)	0.889(0.031)	0.994(0.005)
Dirichlet Laplace	0.982(0.058)	0.978(0.069)	0.989(0.031)	0.880(0.045)	0.977(0.003)
Double Pareto	0.984(0.055)	0.994(0.034)	0.988(0.033)	0.887(0.034)	0.999(0.005)
MSE					
Horseshoe	0.194(0.006)	0.201(0.022)	0.229(0.040)	0.101(0.010)	0.003(0.000)
Dirichlet Laplace	0.195(0.014)	0.201(0.022)	0.228(0.039)	0.103(0.010)	0.005(0.000)
Double Pareto	0.194(0.014)	0.201(0.022)	0.229(0.039)	0.102(0.011)	0.003(0.000)

Table 1: Variable Selection Performance among BNB Simulation Scenarios

Simulation	N,P	Horseshoe	Dirichlet Laplace	Double Pareto
MSE of β				
Sim1	1000,4	0.097(0.168)	0.134(0.392)	0.285(0.971)
Sim2	500,10	0.092(0.023)	0.156(0.132)	0.359(0.34)
Sim3	400,500	0.196(0.01)	0.196(0.01)	0.196(0.01)
MSE of α				
Sim1	1000,4	0.056(0.027)	0.053(0.032)	0.21(0.117)
Sim2	500,10	0.021(0.015)	0.176(0.208)	0.152(0.05)
Sim3	400,500	0.201(0.022)	0.201(0.022)	0.201(0.022)

Table 2: Comparison of MSE among BZINB Simulation Scenarios

Datasets			
Methods	Selected Variables	MSE.β	MSE.Y
Biochemists			
NB with Horseshoe	x_3, x_5	0.017	4.19
NB with Dirichlet Laplace	x_3, x_5	0.016	4.19
NB with Double Pareto	x_1, x_3, x_5	0.015	4.18
ZINB with Horseshoe	x_3, x_5	0.011	8.07
ZINB with Dirichlet Laplace	x_3, x_5	0.01	7.84
ZINB with Double Pareto	x_3, x_5	0.01	8.72
Poisson	x_1, x_2, x_3, x_5	0	3.75
Negative Binomial	x_1, x_3, x_5	0	3.79
Zero-Inflated Poisson	x_1, x_2, x_3, x_5	0	9.17
Zero-Inflated Negative Binomial	x_1, x_2, x_5	0	8.26
Nuts			
NB with Horseshoe	x_5, x_6, x_7	0.309	607.5
NB with Dirichlet Laplace	x_5, x_6, x_7	0.312	608.4
NB with Double Pareto	x_5, x_6, x_7	0.312	605.9
Poisson	x_5, x_6, x_7	0	184.05
Negative Binomial	x_7	0	176.23
NMES			
NB with Horseshoe	$x_1, \ldots, x_6$	0.004	64.40
NB with Dirichlet Laplace	$x_1, \ldots, x_6$	0.004	64.38
NB with Double Pareto	$x_1, \ldots, x_6$	0.004	64.39
Poisson	$x_1, \ldots, x_12, x_14 \ldots, x_16$	0	47.31
Negative Binomial	$x_1, \ldots, x_6, x_8, \ldots$	0	59.26
Covid-19 vaccine			
NB with Horseshoe	$x_1, x_2, x_3, x_4, x_6, x_7, x_{10}, \ldots, x_{19}$	6.21	390.36
NB with Dirichlet Laplace	$x_2, x_4, x_6, \ldots, x_{12}, x_{15}, x_{16}, x_{19}$	6.46	390.24
NB with Double Pareto	$x_7, x_9, x_{10}, x_{13}, x_{14}, x_{16}$	5.14	390.27
ZINB with Horseshoe	$x_1, \ldots, x_5, x_7, x_8, x_9, x_{11}, \ldots, x_{15}, x_{17}, x_{18}, x_{20}$	6.06	427.11
ZINB with Dirichlet Laplace	$x_7, x_9, x_{10}, x_{13}, x_{14}, x_{16}$	73.5	386.72
ZINB with Double Pareto	$x_1, \ldots, x_{20}$	5.88	407.19
Poisson	$x_1, x_2, x_4, x_5, x_7 \ldots, x_{14}, x_{16}, x_{17}, x_{18}, x_{20}$	2.24	1566.97
Negative Binomial	$x_1, \ldots, x_5, x_7, \ldots, x_{10}, x_{11}, x_{13} \ldots, x_{15}, x_{17}, \ldots, x_{20}$	2.42	1566.97
Zero-Inflated	$x_1, x_2, x_5, x_8, x_9, x_{10}, x_{11}, x_{13}, x_{14}, x_{16}, x_{17}, x_{18}, x_{20}$	3.5	1546.08
PBMC RNA-Seq			
NB with Horseshoe	2742, 6369, 9338, 3665, 7203, 11458, 2217, 4360, 4368	0.00	20.35
NB with Dirichlet Laplace	2742, 13706, 6369, 9338, 11458, 3665, 7203, 2217, 4360, 4368	0.00	19.92
NB with Double Pareto	2742, 13706, 6369, 9338, 11458, 7203, 2217, 3665, 4368, 11191	0.00	20.40
ZINB with Horseshoe	74, 267, 1521, 1804, 2612, 3417, 3695, 6049, 8146, 8285	0.00	20.43
ZINB with Dirichlet Laplace	74, 153, 267, 904, 1804, 3417, 6049, 6680, 7526, 8146	0.00	7.48
ZINB with Double Pareto	74, 267, 1804, 1983, 2323, 3124, 3317, 3417, 5528, 6049	0.00	6.06
Covid-19 RNA-Seq			
NB with Horseshoe	4876, 4903, 7767, 8998, 9016, 9856, 10183, 11222, 11598, 12202;	0.0	1016.338
NB with Dirichlet Laplace	4903, 4995, 7767, 7798, 8523, 10416, 11222, 12202, 13109, 16202;	0.0	3276.73
NB with Double Pareto	8523, 8998, 9016, 9301, 9856, 12202, 13109, 14012, 14476, 15265;	0.0	4971.96
ZINB with Horseshoe	22, 58, 202, 231, 272, 280, 297, 314, 404, 523;	0.0	7.11
ZINB with Dirichlet Laplace	193, 219, 293, 391, 487, 514, 589, 625, 697, 724;	0.0	1.99
ZINB with Double Pareto	76, 113, 163, 179, 202, 279, 356, 366, 428, 446;	0.0	3.97

Table 3: Variable Selection in Real World Data