CHAPTER 1

Scalable Bayesian Variable Selection Regression Models for Count Data

Yinsen Miao[*,**], Jeong Hwan Kook[*], Yadong Lu[**], Michele Guindani[**] and Marina Vannucci[*]

[*] Rice University, Department of Statistics, 6100 Main St, Houston, TX 77005

[**] University of California, Irvine, Department of Statistics, Brent Hall 2241, Irvine, CA 92697

Abstract

Variable selection, also known as feature selection in the machine learning literature, plays an indispensable role in scientific studies. In many research areas with massive data, finding a subset of representative features that best explain the outcome of interest has become a critical component in any researcher's workflow. In this chapter, we focus on Bayesian variable selection regression models for count data, and specifically on the negative binomial linear regression model and on the Dirichlet-multinomial regression model. We address the variable selection problem via *spike-and-slab* priors. For posterior inference, we review standard MCMC methods and also investigate computationally more efficient variational inference approaches that use data augmentation techniques and concrete relaxation methods. We investigate performance of the methods via simulation studies and benchmark datasets.

Keywords: Bayesian Variable Selection, Count Data, Data Augmentation, Dirichlet-Multinomial Regression, Negative Binomial Regression, Spike-and-Slab Priors, Tensorflow, Variational Inference

Chapter points

- We consider linear regression models for count data, specifically negative Binomial regression models and Dirichlet-multinomial regression models. We address variable selection via the use of *spike-and-slab* priors on the regression coefficients.
- We develop efficient variational methods for scalability in the number of covariates that are based on augmentation techniques and concrete relaxation methods.
- We provide C/C++ code at https://github.com/marinavannucci/snbvbs, for the negative binomial case, and Python code at https://github.com/mguindanigroup/vbmultdir, for the Dirichlet-multinomial case.

1. Introduction

Variable selection, also known as feature selection in the machine learning literature, plays an indispensable role in scientific studies: in cancer research, biomedical scientists seek to find connections between cancer phenotypes and a parsimonious set of genes; in finance, economists look for a small portfolio that can accurately track the performance of stock market indices such as the S&P 500. In many research areas with massive data, finding a subset of representative features that best explain the outcome of interest has become a critical component in any researcher's workflow.

As evidenced by numerous research papers published in either theory or practice, variable selection for linear regression models has been an important topic in the statistical literature for the past several decades. Variable selection methods can be categorized into roughly three groups: criteria-based methods including traditional approaches such as AIC/BIC [6, 43], penalized regression methods [47, 12, 14, 58] and Bayesian approaches [30, 16, 5]. In this chapter, we focus primarily on Bayesian approaches for variable selection that use *spike-and-slab* priors. An obvious advantage when using these priors is that, in addition to the sparse estimation of the regression coefficients, these methods produce posterior probabilities of inclusion (PPIs) for each covariate. Moreover, Bayesian approaches have the advantages of being able to aggregate multiple sub-models from a class of possible ones, based on their corresponding posterior probabilities. This approach is known as Bayesian model averaging (BMA) and can lead to improved prediction accuracy over single models [18].

Despite the great features offered by *spike-and-slab* priors, computational issues remain a challenge. The posterior distribution for a candidate model usually does not have a closed-form expression, and its inference may be computationally intractable even for a moderate number of predictors. To address the problem, approximate methods that use Markov Chain Monte Carlo (MCMC) stochastic searches have been extensively used [16, 5]. Recently, variational inference (VI) methods [7, 20, 34, 53, 41] have attracted attention as a faster and more scalable alternative. These methods have also been used for model selection in different applied modeling contexts, particularly in bioinformatics [19] and neuroimaging [32, 54].

In this chapter, we focus primarily on regression models for count data, and specifically on negative binomial linear regression models and on Dirichlet-multinomial regression models. In both settings, we formulate a Bayesian hierarchical model with variable selection using *spike-and-slab* priors. For posterior inference, we review standard MCMC methods and also investigate computationally more efficient variational inference approaches that use data augmentation techniques and concrete relaxation methods. We investigate performance of the methods via simulation studies and benchmark datasets.

2. Bayesian Variable Selection via Spike-and-Slab Priors

In ordinary linear regression, a response y_i is modeled as

$$y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma^2), \tag{1.1}$$

for $i = 1, \ldots, n$, with $\boldsymbol{x}_i \in \mathbb{R}^p$ a vector of p known covariates, $\boldsymbol{\beta} = \left[\beta_1, \ldots, \beta_p\right]^T$ a vector of regression coefficients and β_0 the baseline or intercept. A Bayesian approach to variable selection in linear regression models formulates the selection problem via hierarchical priors on the unknown coefficients β_k, $k = 1, \ldots, p$. In this chapter we examine one of the most widely used sparsity-inducing priors, known as the *spike-and-slab* prior [30]. This prior can be written as

$$\beta_k \mid \gamma_k \sim \gamma_k \text{Normal}\left(0, \sigma_\beta^2\right) + (1 - \gamma_k)\, \delta_0, \quad k = 1, \ldots, p, \tag{1.2}$$

with γ_k a latent indicator variable of whether the k-th covariate has a nonzero effect on the outcome, δ_0 a point mass distribution at 0, and σ_β^2 the variance of the prior effect size. Typically, independent Bernoulli priors are imposed on the γ_k's, i.e. $\gamma_k \sim \text{Bernoulli}(\pi)$. For reviews on the general topic of Bayesian variable selection for regression models with continuous responses we refer interested readers to [33, 13]. Alternatively, shrinkage priors, that do not impose a spike at zero, can be considered, such as the normal-gamma [17], the horseshoe [36], and the LASSO [35] priors.

Recently, non-local prior densities have been used in Bayesian hypothesis testing and variable selection, as an attempt to balance the rates of convergence of Bayes factors under the null and alternative hypotheses [23]. The large sample properties of Bayes factors obtained by local alternative priors imply that, as the sample size increases, evidence accumulates much more rapidly in favor of true alternative models than the true null models. Suppose the null hypothesis H_0 is $\beta \in \Theta_0$ and the alternative hypothesis H_1 is $\beta \in \Theta_1$. Here, we define a non-local density if $p(\beta \mid H_1) = 0$ for all $\beta \in \Theta_0$ and $p(\beta \mid H_1) > 0$ for all $\beta \in \Theta_1$. In the variable selection settings considered in this chapter, the hypotheses relate to the significance of the coefficients, i.e. H_0: $\beta = 0$ versus H_1: $\beta \neq 0$. Therefore, a non-local selection prior is defined as a mixture of a point mass at zero and a continuous non-local alternative distribution,

$$\beta_k \mid \gamma_k \sim \gamma_k p\left(\beta_k; \sigma_\beta^2\right) + (1 - \gamma_k)\, \delta_0, \quad k = 1, \ldots, p, \tag{1.3}$$

where $p\left(\beta_k; \sigma_\beta^2\right)$ is a non-local density characterizing the prior distribution of β_k under the alternative hypothesis. Similarly as in the traditional *spike-and-slab* prior formulation, a non-local selection prior models the sparsity explicitly by assigning a positive mass at the origin. However, unlike a flat Gaussian distribution, the density $p\left(\beta_k; \sigma_\beta^2\right)$ does not place a significant amount of probability mass near the null value zero, thus properly reflecting the prior belief that the parameter is away from zero under H_1. In

this chapter, we use the product second moment (pMOM) prior [23, 44] and assume that the β_k's are independent of each other and are drawn from

$$p(\boldsymbol{\beta}; \sigma_\beta^2) = \prod_{k=1}^{p} \frac{\beta_k^2}{\sigma_\beta^2} \text{Normal}\left(0, \sigma_\beta^2\right). \tag{1.4}$$

3. Negative Binomial Regression Models

For $i = 1, \ldots, n$, let now y_i indicate observed counts on an outcome variable. Count data can be modeled via a negative binomial distribution, obtaining the regression model

$$y_i \mid r, \psi_i \sim \text{NB}\left(r, \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}\right), \tag{1.5}$$

with $\psi_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}$ and with r the overdispersion parameter. Given the law of total expectation and variance, the expectation and variance of y_i can be calculated as

$$\begin{aligned} \mathbb{E}\left[y_i \mid \boldsymbol{x}_i\right] &= \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0 + \log r\right), \\ \text{Var}\left[y_i \mid \boldsymbol{x}_i\right] &= \mathbb{E}\left[y_i \mid \boldsymbol{x}_i\right] + \frac{1}{r}\mathbb{E}^2\left[y_i \mid \boldsymbol{x}_i\right], \end{aligned} \tag{1.6}$$

showing that $\text{Var}\left[y_i \mid \boldsymbol{x}_i\right] > \mathbb{E}\left[y_i \mid \boldsymbol{x}_i\right]$ and thus that the negative binomial model can account for overdispersion. Later on we will introduce auxiliary variables to facilitate the use of data augmentation techniques that allow conjugate inference on the parameters $\boldsymbol{\beta}$ and r. We write the prior model as follows:

$$\begin{aligned} \beta_k \mid \gamma_k &\sim \gamma_k \text{Normal}\left(0, \sigma_\beta^2\right) + (1 - \gamma_k)\, \delta_0, \\ \gamma_k &\sim \text{Bernoulli}\left(\pi\right), \\ \beta_0 &\sim \text{Normal}\left(0, \sigma_{\beta_0}^2\right), \\ r &\sim \text{Gamma}\left(a_r, b_r\right), \\ \sigma_\beta^2 &\sim \text{Scaled-Inv-}\chi^2\left(\nu_0, \sigma_0^2\right). \end{aligned} \tag{1.7}$$

Typically, a flat normal prior is imposed on the intercept term β_0, since there is usually no reason to shrink it towards zero. Parameters σ_β^2 and π control the sparsity of the model. Performance of variable selection can be sensitive to these parameter settings. Two popular prior choices for π are the beta distribution $\pi \sim \text{Beta}\left(a_\pi, b_\pi\right)$ and the uniform distribution on the log scale $\log\left(\pi\right) \sim \text{Uniform}\left(\pi_{\min}, \pi_{\max}\right)$ [57]. When π is marginalized, the obtained prior distributions on $\boldsymbol{\gamma}$ are a beta binomial distribution and a truncated beta distribution, respectively. We impose a convenient heavy-tail conjugate prior called scaled inverse chi-square distribution on the slab variance pa-

rameter σ_β^2 where ν_0 is the degree of freedom for the scale parameter σ_0^2. For stability purpose, it is recommended to use a large ν_0 for sparse models [7].

For posterior inference, with variable selection as the main focus, we are interested in recovering a small subset of covariates with significant association to the outcome. In the proposed Bayesian model, the relative importance of the k-th covariate can be assessed by computing its marginal posterior probability of inclusion (PPI) as

$$\text{PPI}\,(k) \equiv p\,(\gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{\sum_{\boldsymbol{\gamma}_{-k}} p\,(\boldsymbol{\gamma}_{-k}, \gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{X})}{\sum_{\boldsymbol{\gamma}_{-k}} p\,(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{X})}, \tag{1.8}$$

which involves a sum over 2^p possible models marginalized over the other model parameters. Classical MCMC algorithms can be used to compute this analytically intractable term. Approaches that use data augmentation schemes have proven particularly efficient.

3.1. Data Augmentation

Here we employ the Pólya-Gamma augmentation approach of Polson et al. [37] to sample $\boldsymbol{\beta}$ and an additional data augmentation scheme to obtain a closed-form, tractable update rule for the overdispersion parameter r, which we adapt from Zhou et al. [56].

A random variable ω following a Pólya-Gamma distribution with parameters $b \in \mathbb{R}_+$, $c \in \mathbb{R}$ is defined as

$$\omega \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/\left(4\pi^2\right)}, \tag{1.9}$$

where the $g_k \sim \text{Gamma}\,(b, 1)$ are independent gamma random variables and $\overset{D}{=}$ indicates equality in distribution. The main result from Polson et al. [37] is that given a random variable ω with density $\omega \sim \text{PG}\,(b, 0)$, $b \in \mathbb{R}_+$ the following integral identity holds for all $a \in \mathbb{R}$:

$$\frac{\exp\,(\psi)^a}{(1 + \exp\,(\psi))^b} = 2^{-b} \exp\,(\kappa\psi)\, \mathbb{E}_\omega \left[\exp\left(-\omega\psi^2/2\right)\right], \tag{1.10}$$

where $\kappa = a - b/2$. Additionally, the conditional distribution $p\,(\omega \mid \psi)$, arising from treating the above integrand as the unnormalized joint density of (ω, ψ), is

$$p\,(\omega \mid \psi) = \frac{\exp\left(-\psi^2\omega/2\right)}{\mathbb{E}_\omega\left[\exp\left(-\psi^2\omega/2\right)\right]} p\,(\omega \mid b, 0)\,, \tag{1.11}$$

which is also in the Pólya-Gamma class, i.e., $\omega \mid \psi \sim \text{PG}(b, \psi)$. For more details regarding the derivation of the result, we refer interested readers to Polson et al. [37]. Comparing Equation (1.10) with the negative binomial regression likelihood given in

Equation (1.5) we can define $a = y_i$ and $b = y_i + r$ and therefore write out the likelihood function as

$$\mathcal{L}(y_i \mid \psi_i, r) = \frac{\Gamma(y_i + r)}{\Gamma(y_i + 1)\Gamma(r)} \frac{\exp(\psi_i)^{y_i}}{(1 + \exp(\psi_i))^{y_i + r}}, \tag{1.12}$$

where $\Gamma(\cdot)$ is the gamma function. We are ready to appeal to the above Pólya-Gamma augmentation and write the likelihood function of the i-th observation conditioned on the augmented variable $\omega_i \sim \text{PG}(y_i + r, 0)$ as

$$\mathcal{L}(y_i \mid \psi_i, r) \propto \exp(\kappa_i \psi_i) \, \mathbb{E}_{\omega_i}\left[\exp\left(-\omega_i \psi_i^2 / 2\right)\right], \tag{1.13}$$

with $\kappa_i = (y_i - r)/2$ and $[\omega_i \mid \psi_i] \sim \text{PG}(y_i + r, \psi_i)$.

We adopt an additional data augmentation scheme to obtain a closed-form, tractable update rule for the overdispersion parameter r. We note that $y_i \sim \text{NB}(r, p_i)$ can be expressed as a compound Poisson distribution [38]

$$\begin{aligned} y_i &= \sum_{l=1}^{L_i} u_{il} \text{ where } i \in \{1, \ldots, n\} \text{ and } l \in \{1, \ldots, L_i\}, \\ L_i &\sim \text{Poisson}(-r \log(1 - p_i)), \\ u_{il} &\overset{\text{iid}}{\sim} \text{Logarithmic}(p_i), \end{aligned} \tag{1.14}$$

where L_i can be interpreted as the number of groups, u_{il} is number of individuals within l-th group and y_i the number of total individuals for the i-th observation. Therefore, exploiting conjugacy between the Gamma and Poisson distributions, a Gamma(a_r, b_r) prior on r leads to the conditional posterior

$$[r \mid \ldots] \sim \text{Gamma}\left(a_r + \sum_{i=1}^{n} L_i, b_r - \sum_{i=1}^{n} \log(1 - p_i)\right). \tag{1.15}$$

The remaining question is how to obtain the conditional posterior of L_i. Zhou et al. [56] show that the probability mass function (PMF) of L_i is the Antoniak equation

$$P(L_i = l_i \mid y_i, r) \overset{\text{def}}{=} f_L(l_i \mid y_i, r) = |s(y_i, l_i)| \frac{r^{l_i} \Gamma(r)}{\Gamma(r + y_i)}, \tag{1.16}$$

where $0 \leq l_i \leq y_i$ and $s(y_i, l_i)$ is the Stirling number of the first kind [3, 45]. By definition, $|s(0, 0)| = 1$, $|s(0, l)| = 0$ for $l > 0$, $|s(y_i, l_i)| = 0$ for $l_i > y_i$, and the other values are given by the recursion as $|s(y_i + 1, l)| = |s(y_i, l - 1)| + y_i |s(y_i, l)|$. The Antoniak equation (1.16) can also be interpreted as the probability that y_i samples from a Dirichlet process with concentration parameter r will return l_i distinct groups, which follows a Chinese restaurant table (CRT) distribution. Consider a Chinese restaurant with an infinite number of tables, each with infinite capacity. Given a concentration parameter

r, we would like to sit y_i customers in this restaurant using the following rule: a customer w, $w = 1, \ldots, y_i$, will either choose a new empty table (group) with probability $r/(r + w - 1)$ or decide to sit at an occupied table otherwise. Hence we can treat the event of creating new tables (groups) as an independent Bernoulli trial and count the number of successful events. The expected mean and variance of table counts, given y_i seated customers, are

$$\begin{aligned} \mathbb{E}[L_i] &= \sum_{w=1}^{y_i} \frac{r}{r+w-1} = r\left(\Psi(r+y_i) - \Psi(r)\right), \\ \mathrm{Var}[L_i] &= r\left(\Psi(r+y_i) - \Psi(r)\right) + r^2\left(\Psi'(r+y_i) - \Psi'(r)\right), \end{aligned} \tag{1.17}$$

where $\Psi(\cdot)$ is the digamma function. Using those analytical moments, we apply the central limit theorem (CLT) [11] and utilize the following asymptotic approximations

$$\begin{aligned} L_i &\asymp \mathrm{Normal}\left(\mathbb{E}[L_i], \mathrm{Var}[L_i]\right), \\ L_i &\asymp \mathrm{Poisson}\left(\mathbb{E}[L_i]\right), \end{aligned} \tag{1.18}$$

to sample L_i when y_i is large.

3.2. MCMC Algorithm

We integrate out the sparsity prior parameter π. Additionally, to gain further computational speed, in our implementation we use the Pólya-Gamma augmentation to marginalize over β_0 and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ when updating the variable selection indicators $\boldsymbol{\gamma}$ and then perform the remaining updates conditional upon a sufficient estimate of those parameters. A generic iteration of our MCMC therefore consists of two Metropolis-Hasting steps on $\boldsymbol{\gamma}$ and $\tau_\beta = \sigma_\beta^{-2}$ within two Gibbs updates on $\boldsymbol{\omega}$ and r:

- To sample the model selection parameter $\boldsymbol{\gamma}$, we follow the modified add-delete-swap algorithm proposed by [7] which selects variable at a frequency which is proportional to the likelihood. Specifically, we propose an add move with a probability proportional to $p\left(\boldsymbol{y} \mid \boldsymbol{X}, \gamma_k = 1, \boldsymbol{\gamma}_{-k}, \boldsymbol{\omega}, \tau_\beta, r\right)$, and a delete move with probability proportional to $p\left(\boldsymbol{y} \mid \boldsymbol{X}, \gamma_k = 0, \boldsymbol{\gamma}_{-k}, \boldsymbol{\omega}, \tau_\beta, r\right)$. Let us denote the marginal likelihood of model $M_{\boldsymbol{\gamma}}$ with the abbreviated notation $\ell(\boldsymbol{\gamma})$ as

$$\ell(\boldsymbol{\gamma}) \equiv p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \tau_\beta, r\right) \propto \frac{\tau_\beta^{\frac{m}{2}}}{\sqrt{\bar{\omega}}} \left|\boldsymbol{S}_{\boldsymbol{\gamma}}^{\tau_\beta}\right|^{\frac{1}{2}} \exp\left(\frac{1}{2}\left(SSR_{\boldsymbol{\gamma}}^{\tau_\beta} + \frac{\bar{\kappa}^2}{\bar{\omega}}\right)\right), \tag{1.19}$$

  where $\boldsymbol{S}_{\boldsymbol{\gamma}}^{\tau_\beta}$ and $\mathrm{SSR}_{\boldsymbol{\gamma}}^{\tau_\beta}$ are $\left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \hat{\boldsymbol{\Omega}} \boldsymbol{X}_{\boldsymbol{\gamma}} + \tau_\beta \mathbb{I}_m\right)^{-1}$ and $\hat{\boldsymbol{\kappa}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{S}_{\boldsymbol{\gamma}}^{\tau_\beta} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \hat{\boldsymbol{\kappa}}$, respectively. We define $\boldsymbol{\Omega} = \mathrm{diag}(\boldsymbol{\omega})$, $\bar{\kappa} = \sum_{i=1}^n \kappa_i$, $\bar{\omega} = \sum_{i=1}^n \omega_i$, $\hat{\boldsymbol{\kappa}} = \boldsymbol{\kappa} - \frac{\bar{\kappa}}{\bar{\omega}}\boldsymbol{\omega}$, and $\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega} - \frac{\boldsymbol{\omega}\boldsymbol{\omega}^T}{\bar{\omega}}$. $\mathbb{I}_m$ is an identity matrix of dimension $m \times m$ and $\mathrm{SSR}_{\boldsymbol{\gamma}}^{\tau_\beta}$ is often referred to as the

sum of squares due to regression (SSR). We write the acceptance probability for the add and delete move as

$$\mathcal{A}(\gamma_k = 0, \hat{\gamma}_k = 1) = \min\left\{1, \frac{a_\pi + m}{b_\pi + p - m - 1} \frac{\ell(\gamma_k = 1, \boldsymbol{\gamma}_{-k})}{\ell(\gamma_k = 0, \boldsymbol{\gamma}_{-k})} \frac{\sum_{j:\gamma_j=0} \frac{\ell(\gamma_j=1,\boldsymbol{\gamma}_{-j})}{\ell(\gamma_j=0,\boldsymbol{\gamma}_{-j})}}{\sum_{j:\hat{\gamma}_j=1} \frac{\ell(\gamma_j=0,\hat{\boldsymbol{\gamma}}_{-j})}{\ell(\gamma_j=1,\hat{\boldsymbol{\gamma}}_{-j})}}\right\},$$

$$\mathcal{A}(\gamma_k = 1, \hat{\gamma}_k = 0) = \min\left\{1, \frac{b_\pi + p - m}{a_\pi + m - 1} \frac{\ell(\gamma_k = 0, \boldsymbol{\gamma}_{-k})}{\ell(\gamma_k = 1, \boldsymbol{\gamma}_{-k})} \frac{\sum_{j:\gamma_j=1} \frac{\ell(\gamma_j=0,\boldsymbol{\gamma}_{-j})}{\ell(\gamma_j=1,\boldsymbol{\gamma}_{-j})}}{\sum_{j:\hat{\gamma}_j=0} \frac{\ell(\gamma_j=1,\hat{\boldsymbol{\gamma}}_{-j})}{\ell(\gamma_j=0,\hat{\boldsymbol{\gamma}}_{-j})}}\right\},$$

where $\boldsymbol{\gamma}_{-k}$ is the set of all indicator variables excluding the k-th one. Computations can be made more efficient by using Cholesky decompositions. See [40] for details.

- We perform a Metropolis-Hasting (MH) update on the log of the slab precision τ_β

$$\log \hat{\tau}_\beta = \log \tau_\beta + u, \tag{1.20}$$

where u is a random draw from a Normal $\left(0, \sigma_\epsilon^2\right)$ and σ_ϵ^2 is the MH step size variance. Then we admit this candidate $\hat{\tau}_\beta$ with acceptance probability

$$\mathcal{A}\left(\tau_\beta, \hat{\tau}_\beta\right) = \min\left\{1, \exp\left(\frac{1}{2}\left(SSR_\gamma^{\hat{\tau}_\beta} - SSR_\gamma^{\tau_\beta}\right)\right)\left(\frac{\left|\hat{\tau}_\beta \boldsymbol{S}_\gamma^{\hat{\tau}_\beta}\right|}{\left|\tau_\beta \boldsymbol{S}_\gamma^{\tau_\beta}\right|}\right)^{1/2}\right\}. \tag{1.21}$$

- Using the compound Poisson distribution [55] representation of the negative binomial distribution, we show the conditional posterior of r as

$$\begin{aligned}[r \mid \ldots] &\sim \text{Gamma}\left(a_r + \sum_{i=1}^{n} L_i, b_r + \sum_{i=1}^{n} \log\left(1 + \exp\left(\psi_i\right)\right)\right), \\ [L_i \mid \ldots] &\sim \text{CRT}\left(y_i, r\right),\end{aligned} \tag{1.22}$$

where CRT is Chinese restaurant table distribution.

- Polson et al. [37] showed that the posterior of ω_i given the linear term ψ_i and the other remaining parameters follows a Pólya-Gamma distribution. Therefore, the conditional update for each ω_i for $i = 1, \ldots, n$ is given by

$$[\omega_i \mid \ldots] \propto \exp\left(-\omega_i \psi_i^2 / 2\right) \text{PG}\left(\omega_i; y_i + r, 0\right) \propto \text{PG}\left(y_i + r, \psi_i\right). \tag{1.23}$$

3.3. Variational Inference Algorithm

Unlike MCMC methods, variational inference (VI) is based on an optimization problem [4]. Let us consider the set of parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the conditional posterior distribution $f(\boldsymbol{\beta}, \boldsymbol{\gamma})$ given $r, \omega, \pi, \sigma_\beta^2$. The underlying idea of VI is to pick a family of distributions $q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \boldsymbol{Q}$, with free variational parameter $\boldsymbol{\theta}$, and then use the gradient

descent algorithm on $\boldsymbol{\theta}$ to minimize the Kullback-Leibler (KL) divergence between the variational approximation q and the posterior distribution $f(\boldsymbol{\beta}, \boldsymbol{\gamma})$ as

$$\begin{aligned} q^* &= \arg\min_{q \in Q} KL(q \parallel f) = \int\int q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\beta}, \boldsymbol{\gamma})}{f(\boldsymbol{\beta}, \boldsymbol{\gamma})} d\boldsymbol{\beta} d\boldsymbol{\gamma} \\ &= \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega}, \boldsymbol{\vartheta}, r) - \left\{ \mathbb{E}^Q \left[\log p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{X}, \boldsymbol{\omega}, \boldsymbol{\vartheta}, r) \right] + \mathbb{H}\left[q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \right] \right\} \\ &= \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega}, \boldsymbol{\vartheta}, r) - \text{ELBO}, \end{aligned} \tag{1.24}$$

with $\boldsymbol{\vartheta}$ the set of hyperparameters or $\boldsymbol{\vartheta} = \left(\pi, \sigma_\beta^2\right)$, and where $\mathbb{H}\left[q(\boldsymbol{\beta}, \boldsymbol{\gamma})\right]$ denotes the entropy of the variational distribution. Given that the conditional marginal likelihood $\log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega}, \boldsymbol{\vartheta}, r)$ does not depend on $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, one can maximize the remaining term on the right hand side, often referred to as the evidence lower bound (ELBO).

For practical reasons the variational family $\boldsymbol{Q}$ is chosen to be a set of parametric distributions from the exponential family. In particular, in order to reduce the computational complexity of the optimization, a common approach is to assume that the latent variables are mutually independent and each governed by a distinct factor in the variational density. This class of variational family $\boldsymbol{Q}$ is known as the mean-field variational family. In particular, in the negative binomial regression model case introduced in this chapter we assume

$$q(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta}) = \prod_{k=1}^{p} q(\beta_k, \gamma_k; \theta_k), \tag{1.25}$$

with

$$q(\beta_k, \gamma_k; \theta_k) = \begin{cases} \alpha_k \text{Normal}(\beta_k \mid \mu_k, s_k^2) & \text{if } \gamma_k = 1 \\ (1 - \alpha_k)\, \delta_0(\beta_k) & \text{otherwise,} \end{cases} \tag{1.26}$$

and variational parameters $\theta_k = \left(\alpha_k, \mu_k, s_k^2\right)$. This factorized approximation is widely used for VI with *spike-and-slab* priors [26, 49, 7, 54, 20]. The closed form of the

ELBO, which we denote as $F(\boldsymbol{\vartheta};\boldsymbol{\theta})$ can be derived as

$$\begin{aligned} F(\boldsymbol{\vartheta};\boldsymbol{\theta}) &\stackrel{\text{def}}{=} \text{ELBO} = \log p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega}, \pi, \sigma_\beta^2, r\right) \geq -\frac{1}{2}\log\bar{\omega} + \frac{\bar{\kappa}^2}{2\bar{\omega}} + \tilde{\boldsymbol{\kappa}}^T \boldsymbol{X}\boldsymbol{A}\boldsymbol{\mu} \\ &+ \sum_{i=1}^{n}\left\{\log\Gamma(y_i + r) - \log\Gamma(y_i + 1) - \log\Gamma(r) - (y_i + r)\log 2\right\} \\ &- \frac{1}{2}\left\{\sum_{k=1}^{p}\left(\boldsymbol{X}^T\hat{\boldsymbol{\Omega}}\boldsymbol{X}\right)_{kk}\left(\left(\mu_k^2 + s_k^2\right)\alpha_k - \mu_k^2\alpha_k^2\right) + \boldsymbol{\mu}^T\boldsymbol{A}\left(\boldsymbol{X}^T\hat{\boldsymbol{\Omega}}\boldsymbol{X}\right)\boldsymbol{A}\boldsymbol{\mu}\right\} \\ &+ \sum_{k=1}^{p}\frac{\alpha_k}{2}\left[1 + \log\left(\frac{s_k^2}{\sigma_\beta^2}\right) - \frac{s_k^2 + \mu_k^2}{\sigma_\beta^2}\right] - \sum_{k=1}^{p}\alpha_k\log\left(\frac{\alpha_k}{\pi}\right) - \sum_{k=1}^{p}(1-\alpha_k)\log\left(\frac{1-\alpha_k}{1-\pi}\right), \end{aligned} \tag{1.27}$$

with $\boldsymbol{A} = \text{diag}\left(\alpha_1, \alpha_2, \ldots, \alpha_p\right)$. By taking partial derivatives of the variational parameters and setting them to zero, we obtain the updating rules for α_k, μ_k and s_k^2:

$$s_k^2 = \frac{1}{\left(\boldsymbol{X}^T\hat{\boldsymbol{\Omega}}\boldsymbol{X}\right)_{kk} + \tau_\beta}, \tag{1.28}$$

$$\mu_k = s_k^2\left(\left(\boldsymbol{X}^T\tilde{\boldsymbol{\kappa}}\right)_k - \sum_{i\neq k}^{p}\alpha_i\mu_i\left(\boldsymbol{X}^T\hat{\boldsymbol{\Omega}}\boldsymbol{X}\right)_{ik}\right), \tag{1.29}$$

$$\text{Logit}(\alpha_k) = \frac{\mu_k^2}{2s_k^2} + \log\left(\frac{s_k}{\sigma_\beta}\right) + \text{Logit}(\pi). \tag{1.30}$$

In order to maximize the ELBO, we devise two variational inference expectation-maximization (VIEM) schemes. The first scheme is described in Figure 1.1 and comprises of a VI-step, an E-step and a M-step. In the VI-step, we use coordinate gradient descent which iteratively updates the variational approximation (1.25). In the E-step, we treat the augmentation variable $\boldsymbol{\omega}$ and overdispersion parameter r as missing latent variables and use the results from Polson et al. [37] and Zhou et al. [56] to update them via the corresponding posterior expected values. In the M-step, we solve for the maximum *a posteriori* (MAP) estimates of σ_β^2 and π. The posterior of σ_β^2 is a scaled inverse chi-square distribution with mode (i.e., the MAP estimator) given by $\hat{\sigma}_\beta^2 = \left(\sum_{k=1}^{p}\alpha_k(\mu_k^2 + s_k^2) + \nu_0\sigma_0^2\right)/(\tilde{\nu}_0 + 2)$. The posterior for π is a beta distribution whose posterior MAP is $\hat{\pi} = \left(\sum_{k=1}^{p}\alpha_k + a_\pi - 1\right)/(p + a_\pi + b_\pi - 2)$. Furthermore, the posterior for r is a gamma distribution with expectation $\mathbb{E}[r] = \tilde{a}_r/\tilde{b}_r$, with $\tilde{a}_r = a_r + \sum_{i=1}^{n}\mathbb{E}[L_i]$ and $\tilde{b}_r = b_r + \sum_{i=1}^{n}\log(1 + \exp(\psi_i))$, where each expectation of L_i is given by Equation (1.17) and the posterior distribution of ω_i is given in Equation

Algorithm 1: VIEM algorithm for Negative Binomial Regression

initialize: $\left(\mu_k, \alpha_k, s_k^2\right)$ for $k = \{1, \ldots, p\}, \hat{\omega}, \hat{\sigma}_\beta^2, \hat{\pi}, \hat{r}$

repeat

Update the Variational parameters via coordinate gradient descent:

for $k = 1 : p$ **do**

1. Update s_k^2 according to Equation (1.28).

2. Update μ_k according to Equation (1.29).

3. Update α_k according to Equation (1.30).

end

Update selection hyperparameters $\hat{\sigma}_\beta^2$ and $\hat{\pi}$ via their MAP estimates.

Update latent variables $\hat{r}$ and $\hat{\omega}_i$, $i \in \{1, \ldots, n\}$ via posterior expectations.

until *ELBO Converges*

Figure 1.1: Variational inference expectation-maximization (VIEM) scheme.

(1.23), with expectation $\mathbb{E}[\omega_i] = (y_i + \mathbb{E}[r])\left\{\frac{\tanh(\psi_i/2)}{2\psi_i}\right\}$. With this scheme, we iterate the three steps until some convergence criterion is met. A commonly used stopping rule is to terminate the algorithm when changes of the ELBO between iterations are less than some pre-specified threshold. An alternative criterion is to use the entropy of the selection parameter $\boldsymbol{\gamma}$ defined as

$$H(\boldsymbol{\gamma}) = -\sum_{k=1}^{p} \{\alpha_k \log_2(\alpha_k) + (1-\alpha_k)\log_2(1-\alpha_k)\}. \tag{1.31}$$

In the second variational scheme, described in Figure 1.2, we integrate out the parameters in $\boldsymbol{\vartheta}$ via importance sampling [7] and estimate the PPIs, defined as in Equation (1.8), as

$$\text{PPI}(k) \approx \frac{\sum_{s=1}^{N} p\left(\gamma_k = 1 \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{\vartheta}^{(s)}, r\right) w\left(\boldsymbol{\vartheta}^{(s)}\right)}{\sum_{s=1}^{N} w\left(\boldsymbol{\vartheta}^{(s)}\right)}, \tag{1.32}$$

with $w(\boldsymbol{\vartheta})$ the unnormalized importance sampling weight for $\boldsymbol{\vartheta}$, calculated by substituting the unknown marginal likelihood $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega}, \boldsymbol{\vartheta}, r)$ with its ELBO. Importance sampling can improve the estimates of the PPIs as it averages over $\boldsymbol{\vartheta}$. Furthermore, since importance samples are independent from each other, one can employ parallel computing framework such as OpenMP [8] to take advantage of multi-core computers.

Algorithm 2: VIEM-IS algorithm for Negative Binomial Regression

initialize: $\left(\mu_k, \alpha_k, s_k^2\right)$, for $k = \{1, \ldots, p\}, \hat{\omega}, \hat{\boldsymbol{\vartheta}} = \left(\hat{\sigma}_\beta^2, \hat{\pi}\right), \hat{r}$
given : Sample $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(ns)}$ from importance distribution $\tilde{p}(\boldsymbol{\vartheta})$
for $s = 1 : ns$ **do**
  repeat
    Update the Variational parameters via coordinate gradient descent:
    for $k = 1 : p$ **do**
      1. Update s_k^2 according to Equation (1.28).
      2. Update μ_k according to Equation (1.29).
      3. Update α_k according to Equation (1.30).
    end
    Update the latent variables $\hat{r}$ and $\hat{\omega}_i$ for $i \in \{1, \ldots, n\}$ via their posterior expectation.
  until *ELBO Converges*
  Compute the unnormalized importance weights $w(\boldsymbol{\vartheta})$
  Set $\boldsymbol{\alpha}^{(i)} = \boldsymbol{\alpha}$ and $\boldsymbol{\mu}^{(i)} = \boldsymbol{\mu}$.
end

Compute the normalized importance weights $\hat{w}(\boldsymbol{\vartheta})$.
Compute the weighted average of $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\beta}^{(s)} = \boldsymbol{\alpha}^{(s)} \cdot \boldsymbol{\mu}^{(s)}$ using $\hat{w}(\boldsymbol{\vartheta})$.

Figure 1.2: Variational inference expectation-maximization via importance sampling (VIEM-IS) scheme.

4. Dirichlet-multinomial Regression Models

The second model for count data that we consider is the Dirichlet-multinomial log-linear regression model. Here, for each observaton i, $i = 1, \ldots, n$, we assume multivariate count data and write $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})$ to indicate the vector of counts on J outcome variables, for $j = 1, \ldots, J$. As in the previous model, we let $\boldsymbol{x}_i$ indicate the vector of measurements on p covariates. We start by modeling the multivariate count data $\boldsymbol{y}_i$ using a multinomial distribution

$$\boldsymbol{y}_i \mid \boldsymbol{\phi}_i \sim \text{Multinomial}\left(y_{i+}, \boldsymbol{\phi}_i\right), \tag{1.33}$$

with $y_{i+} = \sum_{j=1}^{J} y_{ij}$ the summation of all counts in the vector, and where the parameter $\boldsymbol{\phi}_i$ is defined on the J dimensional simplex

$$\mathcal{S}^{J-1} = \left\{ (\phi_{i1}, \ldots, \phi_{iJ}) : \ \phi_{ij} \geq 0, \ \sum_{j=1}^{J} \phi_{ij} = 1 \right\}.$$

We further impose a conjugate Dirichlet prior on $\boldsymbol{\phi}_i$, that is $\boldsymbol{\phi}_i \sim \text{Dirichlet}(\boldsymbol{\xi}_i)$, where $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iJ})$ indicates a J-dimensional vector of strictly positive parameters. An advantage of our hierarchical formulation is that conjugacy can be exploited to integrate $\boldsymbol{\phi}_i$ out, obtaining the Dirichlet-multinomial model, $\boldsymbol{y}_i \sim \text{DM}(\boldsymbol{\xi}_i)$, with probability mass function

$$f(\boldsymbol{y}_i|\boldsymbol{\xi}_i) = \binom{y_{i+}}{\boldsymbol{y}_i} \frac{\Gamma(y_{i+}+1)\Gamma(\xi_{i+})}{\Gamma(y_{i+}+\xi_{i+})} \prod_{j=1}^{J} \frac{\Gamma(y_{ij}+\xi_{ij})}{\Gamma(\xi_{ij})\Gamma(y_{ij}+1)}, \tag{1.34}$$

and $\xi_{i+} = \sum_j^J \xi_{ij}$. First described in [31] as the compound multinomial, the DM model allows more flexibility than the multinomial when encountering overdispersion, as it induces an increase in variance by a factor $(y_{i+} + \xi_{i+}) / (1 + \xi_{i+})$.

Next, we incorporate the covariates into the modeling via a log-linear regression framework where the DM parameters depend on the available covariates. More specifically, we define $\zeta_{ij} = \log(\xi_{ij})$ and assume

$$\zeta_{ij} = \alpha_j + \sum_{k=1}^{p} \beta_{kj}\, x_{ik}. \tag{1.35}$$

In this formulation, the intercept term α_j corresponds to the log baseline parameter for outcome j, whereas the regression parameter β_{kj} captures the association between the k-th covariate and the j-th outcome. Identifying the significant associations is then equivalent to determining the non-zero β_{kj} parameters, a task we can achieve via *spike-and-slab* priors. Here, we use the formulation of [46] and introduce a set of latent binary indicators of the type $\boldsymbol{\gamma}_j = \left(\gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{pj}\right)$ such that $\gamma_{kj} = 1$ if the k-th covariate influences the j-th outcome and $\gamma_{kj} = 0$ otherwise, and write the prior on β_{kj} as

$$\beta_{kj} \mid \gamma_{kj} \sim \gamma_{kj} p\left(\beta_{kj}; \sigma_\beta^2\right) + (1 - \gamma_{kj})\, \delta_0, \tag{1.36}$$

$$\gamma_{kj} \sim \text{Bernoulli}\,(\pi)\,, \tag{1.37}$$

where $p(\beta_{kj}; \sigma_\beta^2)$ is the non-local prior and π again controls the sparsity of the model. For the non-local prior, we consider the product second moment prior described in (1.4). Finally we assume normal priors on the baseline α_j's i.e. $\alpha_j \sim \text{Normal}\left(0, \sigma_\alpha^2\right)$ and use large σ_α^2 to encode a diffuse prior on each α_j.

4.1. MCMC Algorithm

We refer readers to [51] for a MCMC stochastic search method for the Dirichlet-multinomial regression model. Here, instead, we formulate an alternative, scalable variational Bayes algorithm.

4.2. Variational Inference with Reparameterization

Unlike the NB regression model, the DM regression model does not have any known data augmentation schemes that can be paired with a parametric variational family to exploit conditional conjugacy. This is often the case for Bayesian hierarchical models where the corresponding ELBO objective is a function of intractable expectations with respect to the variational distributions. In such settings, the optimal variational parameters can be found via gradient descent algorithm and the ELBO can be approximated by Monte Carlo samples from the variational distributions. Reducing the variance of the gradient estimators plays a significant role in improving model accuracy and scalability of these methods. Below we review the generalized reparameterization (G-REP) gradient method proposed in [42] to obtain low-variance gradient in the case of continuous latent variables.

4.2.1. Reparameterization of the Gradient

Given data x and a continuous latent variable z such that $p(x, z)$ is differentiable with respect to z, a reparameterization transforms z into a new random variable ϵ defined by an invertible transformation $\epsilon = \mathcal{T}^{-1}(z; \theta)$ and $z = \mathcal{T}(\epsilon; \theta)$, where $\epsilon = \mathcal{T}^{-1}(z; \theta)$ can be considered as a standardization procedure that makes the distribution of ϵ weakly dependent on z. By change of variable, the reparameterized model can be written as $p(x, \epsilon; \theta) = p\left(x, \mathcal{T}\left(\epsilon; \theta\right)\right) \times J\left(\epsilon; \theta\right)$, where $J(\epsilon; \theta) = |\det \nabla_{\epsilon} \mathcal{T}(\epsilon; \theta)|$ denotes the determinant of the Jacobian of the transformation. A noticeable property of a valid reparameterization is the marginal likelihood invariance property

$$p(x) = \int p(x, z)dz = \int p(x, \epsilon)d\epsilon = \int p(x, \mathcal{T}(\epsilon; \theta))J(\epsilon; \theta). \tag{1.38}$$

Thus, while θ enters into the above equation as a new model parameter, the marginal probability $p(x)$ remains unchanged. However, the reparameterized posterior distribution $p(\epsilon|x, \theta)$ is dependent on θ and this dependence of the posterior on θ can be exploited to improve accuracy and computational efficiency [48]. In the variational inference context, we can consider $p(\epsilon|x, \theta)$ to be the first part of the ELBO objective corresponding to the expectation of the log likelihood with respect to the variational distributions parameterized by θ. When updating θ via stochastic gradient descent, one can now take advantage of the information provided from the model likelihood. This will generally lead to a faster convergence of θ, and fewer samples of ϵ to estimate a low-variance gradient [42].

4.2.2. Concrete Relaxation

While G-REP can be used to optimize the variational parameters for the regression coefficients β, this approach cannot be used for the discrete model selection variable γ.

Recently, Maddison et al. [28] and Jang et al. [21] have proposed a reparameterization for discrete random variables using the Concrete distribution, which is a continuous relaxation of discrete random variables. The Concrete distribution is a parametric family of continuous distributions on the simplex with closed form densities, parameterized by a location $a > 0$ and a temperature $\lambda > 0$. A key feature of this class of distributions is that any discrete distribution can be seen as the discretization of a Concrete one. For example, the binary model selection random variable $\gamma \sim \text{Bernoulli}(\pi)$ is equivalent to $\gamma \sim \text{BinConcrete}(a, \lambda)$, and γ can be sampled as

$$\gamma = \frac{1}{1 + \exp\left(-\left(\log(a) + L\right)/\lambda\right)}, \quad L = \frac{u}{1-u}, \tag{1.39}$$

where $u \sim \text{Uniform}(0, 1)$. When λ approaches zero, the concrete distribution $q_{a,\lambda}(\gamma)$ converges to $\text{Bernoulli}(\pi = \frac{a}{1+a})$. Because the discretization procedure of the Concrete distribution allows for the optimization of parameter a via gradient-based methods, we can use this reparameterization scheme to optimize π with respect to the ELBO.

4.2.3. Hard Concrete Distribution

A drawback of the Binary Concrete distribution is that a realization from the distribution may not be exactly zero and may be susceptible to the temperature value λ. To resolve this problem, Louizos et al. [27] extended the work of Maddison et al. [28] and Jang et al. [21] and introduced the Hard Concrete Distribution. Let s be a random variable with probability density $q(s) = \text{BinConcrete}(a, \lambda)$ and cumulative density $Q(s)$. After sampling s, we can "stretch" the value to the (c_0, c_1) interval, with $c_0 < 0$ and $c_1 > 1$, and apply a hard-sigmoid

$$s \sim \text{BinConcrete}(a, \lambda), \quad \bar{s} = s(c_1 - c_0) + c_0, \quad z = \min\left(1, \max\left(0, \bar{s}\right)\right). \tag{1.40}$$

This induces a distribution where the mass of $q(\bar{s})$ on the negative domain is "folded" to a delta peak at zero, and mass larger than one is "folded" to a delta peak at one, such that $q(\bar{s})$ is truncated to the (0,1) range. Then, z is a hard-sigmoid rectification of s with support $\{0, 1\}$, as desired. It can be shown that the probability of z being nonzero can be computed as

$$p(z \neq 0) = Q(\bar{s} \geq 0) = \frac{1}{1 + \exp\left(-\left(\log\left(a\right) - \lambda \log\left(-\frac{c_0}{c_1}\right)\right)\right)}. \tag{1.41}$$

With this reparameterization, we can sample a discrete Bernoulli random variable with the above probability and learn a via gradient descent. For posterior inference, we follow Louizos et al. [27] and use $c_0 = -0.1, c_1 = 1.1, \lambda = \frac{2}{3}$.

4.2.4. Variational Inference Approximation

Finally, we describe the variational distributions $q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathcal{Q}$, with free variational parameter $\boldsymbol{\theta}$, for our Dirichlet-multinomial model. For efficient computation, we again use a mean field approximation of the joint posterior of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ of the type

$$q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{k=1}^{p} \prod_{j=1}^{J} q\left(\beta_{kj}, \gamma_{kj}\right) = \prod_{k=1}^{p} \prod_{j=1}^{J} q\left(\beta_{kj} \mid \gamma_{kj}\right) q\left(\gamma_{kj}\right), \tag{1.42}$$

where $q(\beta_{kj} \mid \gamma_{kj})$ is defined as

$$q\left(\beta_{kj} \mid \gamma_{kj}; \boldsymbol{\theta}_{kj}\right) = \begin{cases} \frac{1}{2}\text{Normal}\left(\beta_{kj} \mid \mu_{1kj}, \sigma^2_{1kj}\right) + \frac{1}{2}\text{Normal}\left(\beta_{kj} \mid \mu_{2kj}, \sigma^2_{2kj}\right) & \text{if } \gamma_{kj} = 1 \\ \delta_0 & \text{if } \gamma_{kj} = 0 \end{cases} \tag{1.43}$$

with variational parameter $\boldsymbol{\theta}_{kj} = \left(\mu_{1kj}, \sigma_{1kj}, \mu_{2kj}, \sigma_{2kj}\right)$. Since the pMOM density has two modes, we propose a mixture of two normal distributions as the variational approximation when $\gamma_{kj} = 1$, while the approximation collapses to a spike at zero when $\gamma_{kj} = 0$. Samples from the above distribution can be obtained via the reparameterization $u \sim \text{Uniform}(0, 1), \epsilon \sim \text{Normal}\,(0, 1)$ and

$$\beta_{kj} \mid \gamma_{kj} = \begin{cases} \epsilon\sigma^2_{1kj} + \mu_{1kj} & \text{if } u < 0.5 \text{ and } \gamma_{kj} = 1 \\ \epsilon\sigma^2_{2kj} + \mu_{2kj} & \text{if } u \geq 0.5 \text{ and } \gamma_{kj} = 1 \\ 0 & \text{if } \gamma_{kj} = 0. \end{cases} \tag{1.44}$$

For each γ_{kj} we use a Hard Concrete distribution as the approximation, i.e., $q(\gamma_{kj}) \sim \text{HardBinConcrete}(a_{kj}; \lambda = \frac{2}{3}, c_0 = -0.1, c_1 = 1.1)$. Thus we can learn $q_{a_{kj}}(\gamma_{kj} = 1)$ by performing gradient descent on a_{kj}. For the baseline terms $\alpha'_j s$, we use MAP estimates, since we are mainly interested in performing variable selection on $\beta'_{kj} s$. In summary, the ELBO objective of the DM model can be written as

$$\begin{aligned} \text{ELBO} &= \mathbb{E}^{Q}\left[\log f\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right)\right] - \text{KL}\left(q\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) \parallel p\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right)\right) \\ &= \mathbb{E}^{Q}\left[\log f\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right)\right] - \sum_{k=1}^{p} \sum_{j=1}^{J} \text{KL}\left(q\left(\gamma_{kj}\right) \parallel p\left(\gamma_{kj}\right)\right) \\ &\quad - \sum_{k=1}^{p} \sum_{j=1}^{J} q_{\pi}\left(\gamma_{kj} = 1\right) \text{KL}\left(q_{\boldsymbol{\theta}_{kj}}\left(\beta_{kj} \mid \gamma_{kj} = 1\right) \parallel p\left(\beta_{kj} \mid \gamma_{kj} = 1\right)\right). \end{aligned} \tag{1.45}$$

Since our prior on each γ_{kj} is Bernoulli (π), the KL term for γ_{kj} is

$$\begin{aligned}&\mathrm{KL}\left(q_{a_{kj}}\left(\gamma_{kj}\right) \| \, p\left(\gamma_{kj}\right)\right)\\ =&q_{a_{kj}}\left(\gamma_{kj}=1\right)\log\left(\frac{q_{a_{kj}}\left(\gamma_{kj}=1\right)}{\pi}\right)+q_{a_{kj}}\left(\gamma_{kj}=0\right)\log\left(\frac{q_{a_{kj}}\left(\gamma_{kj}=0\right)}{1-\pi}\right).\end{aligned}$$

The KL term corresponding to the pMOM prior can be expressed as

$$\begin{aligned}&\mathrm{KL}\left(q_{\boldsymbol{\theta}_{kj}}\left(\beta_{kj} \mid \gamma_{kj}=1\right) \| \, p\left(\beta_{kj} \mid \gamma_{kj}=1\right)\right)\\ =&-\mathbb{H}\left(\beta_{kj} \mid \gamma_{kj}=1\right)-\mathbb{E}_{q_{\boldsymbol{\theta}_{kj}}\left(\beta_{kj} \mid \gamma_{kj}=1\right)}\left(\log p\left(\beta_{kj} \mid \gamma_{kj}=1\right)\right),\end{aligned} \tag{1.46}$$

where $\mathbb{H}(\beta_{kj} \mid \gamma_{kj}=1)$ denotes the entropy under $q_{\boldsymbol{\theta}_{kj}}(\beta_{kj} \mid \gamma_{kj}=1)$. Both terms on the right hand side of (1.46) can be computed using Monte Carlo approximations. Furthermore, to reduce the variance of the gradient, we can express the expectation in (1.46) analytically as

$$\frac{1}{S}\sum_{s=1}^{S}\left[\log\left(\left(\beta_{kj}^{(s)}\right)^2\right)-\frac{\mu_{1kj}^2+\sigma_{1kj}^2}{4\sigma_\beta^2}-\frac{\mu_{2kj}^2+\sigma_{2kj}^2}{4\sigma_\beta^2}-\log\left(\sqrt{2\pi}\right)-\frac{3}{2}\log\left(\sigma_\beta^2\right)\right], \tag{1.47}$$

where S is the number of Monte Carlo samples used in the approximation.

4.2.5. Posterior Inference using Tensorflow

The optimization procedure to perform posterior inference using the proposed reparameterization within the variational framework is implemented in TensorFlow [1] and uses the Adam optimizer proposed by Kingma and Ba [24] for gradient optimization. The actual computation for the gradients is handled using Tensorflow's API for automatic differentiation. In order to reduce the complexity of the optimization scheme, we standardized the data, both in simulations and real data analyses, and fixed the variance variational parameters $\sigma_{1kj}^2, \sigma_{2kj}^2$ of β_{kj}, for $k = 1, \ldots, p$ and $j = 1, \ldots, J$, to 1. Given that we are interested in the selection of the relevant variable, these parameters are not the prime interest of our inference and, also, they tend to be underestimated by variational inference schemes. In case it is of interest to learn these parameters, it is advised to perform a log transform $\log(\sigma_{kj}^2) = \tilde{\sigma}_{kj}^2$ so that the parameters remain in the positive domain during gradient updates. More details regarding the implementation can be found at https://github.com/mguindanigroup/vbmultdir.

5. Simulation Study

In this section, we conduct several simulation studies and compare selection performances among different methods. For comparisons, we calculate accuracy (ACC), precision, recall, F1 score, and Matthews correlation coefficient (MCC). Given the num-

ber of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), the accuracy is calculated as (TP+TN)/(P+N), the precision as (TP)/(TP+FP), the recall as (TP)/(TP+FN), the F1 score as the geometric mean between precision and recall and the Matthews correlation coefficient as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}}. \tag{1.48}$$

The Matthews correlation coefficient takes all values of the confusion matrix into account and is generally regarded as a balanced measure that can be used even if the true classes are imbalanced. We further compute and plot the receiving operating characteristic (ROC) curve and the area under the ROC curve (AUC) to show the selection performance of each method using different thresholds on the PPIs.

Negative Binomial - small p large n example

We first simulated synthetic data with $n = 100$ samples and $p = 50$ features. The design matrix $\boldsymbol{X}$ was simulated according to a multivariate Normal$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where each μ_k, $k = 1, \ldots, p$, was drawn from a Normal $(0, 0.1)$, and where the (l, m)-th entry of the covariance matrix was set to be $\boldsymbol{\Sigma}_{lm} = \rho^{|l-m|}$ for $l \neq m$, with $\rho = 0, 0.3, 0.6, 0.9$. We sampled the marginal indicators of inclusion γ_k independently from a Bernoulli (π) with $\pi \in$ Uniform $(0.1, 0.2)$ and the corresponding non-zero β_k uniformly from the intervals $\pm\,[0.5, 2.0]$. Finally, we sampled the count data from a gamma-Poisson mixture model of the type

$$\begin{aligned} y_i &\sim \text{Poisson}\,(\lambda_i)\,, \\ \lambda_i &\sim \text{Gamma}\,(r, 1/\exp(\psi_i))\,, \\ \psi_i &= (\boldsymbol{x}_i)^T_{\boldsymbol{\gamma}}\,\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \beta_0, \end{aligned} \tag{1.49}$$

where we set $r = 1$ and $\beta_0 = 2$. Integrating λ_i out, we have that y_i follows a negative binomial distribution of the type

$$y_i \sim \text{NB}\left(r, \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}\right) \stackrel{\text{def}}{=} \text{NB}\,(r, p_i)\,. \tag{1.50}$$

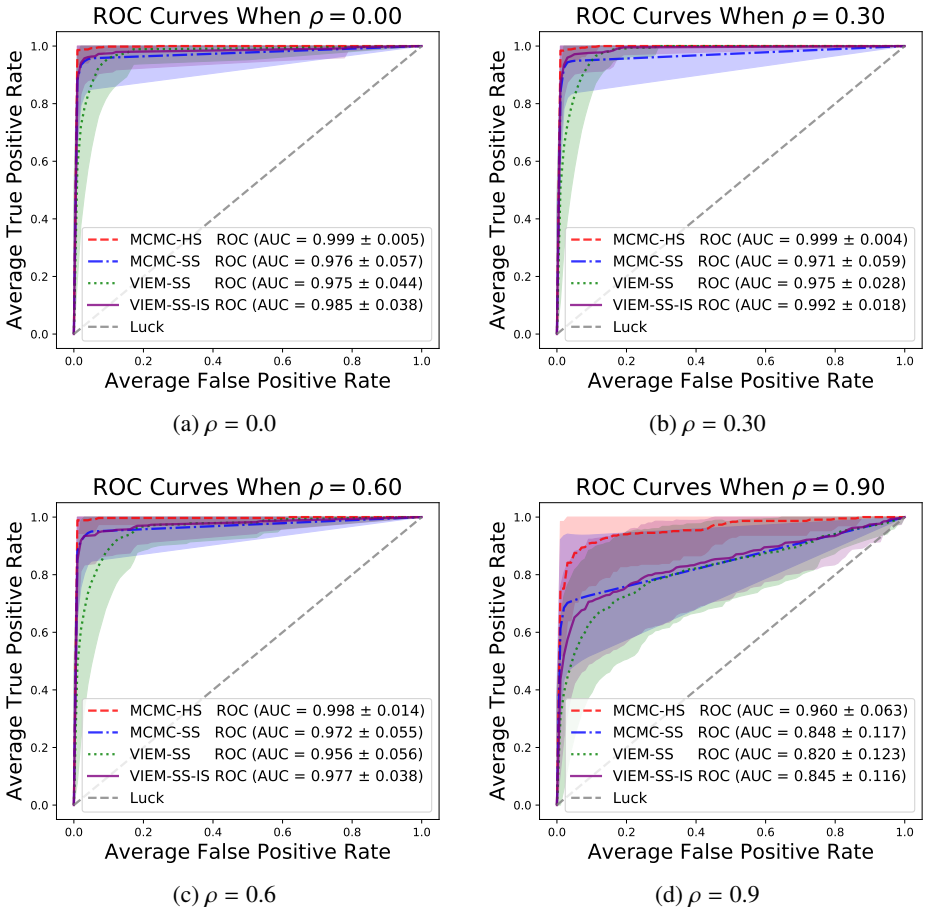
We assessed performances of the Bayesian negative binomial regression model described in this chapter, using MCMC and the VI algorithms for posterior inference. We also considered the LASSO method [47] using the glmnet R package [15]. Finally, we considered the *spike-and-slab* prior versus an adaptive shrinkage horseshoe prior [36]. When fitting the Bayesian models to the data, we imposed a flat Gamma$(0.01, 0.01)$ prior on the over-dispersion r, a Scaled-Inv-$\chi^2\,(10, 1)$ on σ^2_β and an improper uniform prior on the baseline β_0. For the VIEM and the MCMC methods, we set the prior expectation of inclusion to be the true value, while for the VIEM-IS we used 30 equally

spaced grids on the prior log odds of π from -500.0 to -1.0 as the important samples. Results we report here were obtained by running the MCMC algorithms for 23,000 iterations and discarding the initial 3,000 samples as burnin. We assessed convergence of the MCMC algorithms visually via the traceplots of the number of included variables. For the variational algorithms, we terminated the iterations when the absolute changes of the ELBO was less than 0.0001. We utilized 6 threads out of a hexa-core CPU to conduct parallel computation for the VIEM-IS algorithm.

Table 1.1 reports results for precision, recall, MCC, AUC, F1 score, ACC and computing time in seconds, averaged across 50 replicated datasets, with standard deviations in parentheses, and Figure 1.3 shows the corresponding ROC curves, for the different values of ρ. When features are independent ($\rho = 0.0$) or weakly correlated ($\rho = 0.3$), we find that the selection performance of the variational methods closely match that of the sampling-based methods. Figure 1.3a and 1.3b also illustrate that the ROC curves of VIEM (green dotted line) and VIEM-IS (purple solid line) are close in performance to those of the MCMC-HS (red dashed line) and MCMC-SS (blue dot-dash line). When the correlation coefficient ρ increases, we notice a decrease in performance of the VIEM method. This is because the density landscape of the posterior likelihood becomes multi-modal as ρ increases and the EM algorithm is notoriously vulnerable to be trapped in local optima [41]. The VIEM-IS, instead, which utilizes different hyperparameters $\boldsymbol{\vartheta}$ and also several variational parameters $\boldsymbol{\theta}$ to solve each EM optimization independently in parallel, shows more robust performance. When the variables are strongly correlated ($\rho = 0.9$), the fully factorized assumption of the variational approximation in (1.23) becomes invalid and hence the performance of both variational methods become inferior to those of the sampling-based methods. Furthermore, as we can see in Table 1.1, the MCC values of MCMC-HS and MCMC-SS also decrease sharply when ρ increase from 0.6 to 0.9. Therefore, we conclude that sampling methods also suffer to some extent from severe multicollinearity, which is also illustrated in Figure 1.3d. From a computational point of view, as expected, variational methods (VIEM-SS and VIEM-SS-IS) show a dramatic improvement in speed over the sampling-based methods (MCMC-SS and MCMC-HS). In particular, the variational methods are 100 to 1000 times faster than the sampling methods.

Negative Binomial - large p small n example

Next, we considered a simulation with $p = 1000$ and $n = 100$, which we obtained from the previous one simply by adding 950 zero coefficients and adding another 950 columns of independent variables $\tilde{\boldsymbol{X}}_{100\times 950} \sim \text{Normal}(\mathbf{0}, \mathbb{I})$ to the design matrix. We used the same hyperparameter configuration as in the previous example. We dropped the MCMC-HS algorithm, since the moment matrix is not full rank when $p \gg n$. Results are reported in Table 1.2 and Figure 1.4. Both variational methods achieve sim-



(a) $\rho = 0.0$ (b) $\rho = 0.30$

(c) $\rho = 0.6$ (d) $\rho = 0.9$

Figure 1.3: Negative Binomial - small p large n example: Comparison of selection performance (ROC curves). Variational Inference EM *spike-and-slab* (VIEM-SS), Variational Inference EM *spike-and-slab* with importance sampling on π (VIEM-SS-IS), MCMC with *spike-and-slab* prior (MCMC-SS) and MCMC with horseshoe prior (MCMC-HS). The ROC curves and the corresponding standard deviations are averaged over 50 replicated datasets.

ilar performance as the MCMC-SS method but still are around 15 to 75 faster than the MCMC-SS. For the tuning parameter in the Lasso method, we used the default **cv.glmnet** and report the results for the parameter with the smallest cross-validation error.

Table 1.1: Negative Binomial - small p large n example: Performance comparison of Variational Inference EM *spike-and-slab* (VIEM-SS), Variational Inference EM *spike-and-slab* with importance sampling on π (VIEM-SS-IS), MCMC *spike-and-slab* (MCMC-SS), MCMC horseshoe (MCMC-HS) and glmnet (LASSO). Accuracy, recall, precision, F1 score and Matthews correlation coefficient (MCC), averaged over 50 replicated simulated datasets (standard deviation in parentheses).

	MCMS-HS	MCMC-SS	VIEM-SS	VIEM-SS-IS	LASSO
			$\rho = 0.0$		
Precision	0.987 (0.056)	0.966 (0.091)	0.807 (0.215)	0.980 (0.061)	0.184 (0.110)
Recall	0.967 (0.086)	0.959 (0.115)	0.958 (0.095)	0.939 (0.102)	0.997 (0.022)
MCC	0.973 (0.058)	0.956 (0.090)	0.858 (0.163)	0.953 (0.062)	0.300 (0.151)
AUC	0.999 (0.005)	0.976 (0.057)	0.975 (0.044)	0.985 (0.038)	0.209 (0.148)
F1	0.974 (0.057)	0.957 (0.089)	0.862 (0.159)	0.954 (0.061)	0.297 (0.139)
ACC	0.995 (0.011)	0.993 (0.014)	0.969 (0.040)	0.992 (0.010)	0.079 (0.026)
Time	32.270 (1.187)	29.012 (3.559)	0.047 (0.039)	0.325 (0.256)	0.339 (0.091)
			$\rho = 0.3$		
Precision	0.983 (0.067)	0.957 (0.088)	0.751 (0.221)	0.960 (0.084)	0.186 (0.095)
Recall	0.963 (0.087)	0.947 (0.116)	0.957 (0.086)	0.937 (0.102)	0.992 (0.037)
MCC	0.969 (0.071)	0.944 (0.096)	0.818 (0.157)	0.940 (0.069)	0.296 (0.139)
AUC	0.999 (0.004)	0.971 (0.059)	0.975 (0.028)	0.992 (0.018)	0.220 (0.139)
F1	0.971 (0.066)	0.947 (0.092)	0.823 (0.154)	0.942 (0.066)	0.302 (0.122)
ACC	0.994 (0.015)	0.990 (0.017)	0.955 (0.047)	0.990 (0.012)	0.085 (0.031)
Time	32.476 (1.352)	29.754 (4.112)	0.043 (0.033)	0.311 (0.230)	0.350 (0.069)
			$\rho = 0.6$		
Precision	0.964 (0.103)	0.971 (0.083)	0.727 (0.254)	0.971 (0.067)	0.196 (0.116)
Recall	0.964 (0.091)	0.948 (0.108)	0.921 (0.123)	0.905 (0.120)	0.983 (0.076)
MCC	0.957 (0.077)	0.953 (0.087)	0.782 (0.202)	0.928 (0.073)	0.306 (0.154)
AUC	0.998 (0.014)	0.972 (0.055)	0.956 (0.056)	0.977 (0.038)	0.190 (0.143)
F1	0.958 (0.077)	0.955 (0.081)	0.790 (0.196)	0.931 (0.071)	0.312 (0.142)
ACC	0.992 (0.016)	0.991 (0.018)	0.945 (0.060)	0.988 (0.013)	0.085 (0.034)
Time	33.501 (4.543)	29.291 (4.125)	0.048 (0.034)	0.341 (0.233)	0.452 (0.121)
			$\rho = 0.9$		
Precision	0.889 (0.189)	0.887 (0.190)	0.587 (0.280)	0.772 (0.223)	0.234 (0.099)
Recall	0.743 (0.197)	0.705 (0.226)	0.624 (0.218)	0.580 (0.222)	0.901 (0.159)
MCC	0.784 (0.176)	0.766 (0.211)	0.545 (0.247)	0.618 (0.186)	0.342 (0.131)
AUC	0.960 (0.063)	0.848 (0.117)	0.820 (0.123)	0.845 (0.116)	0.192 (0.117)
F1	0.791 (0.167)	0.774 (0.198)	0.591 (0.208)	0.624 (0.187)	0.356 (0.113)
ACC	0.962 (0.037)	0.961 (0.039)	0.910 (0.067)	0.937 (0.043)	0.089 (0.038)
Time	32.456 (1.343)	28.670 (3.468)	0.036 (0.031)	0.269 (0.225)	0.803 (0.160)

Table 1.2: Negative Binomial - large p small n example: Performance comparison of Variational Inference EM *spike-and-slab* (VIEM-SS), Variational Inference EM *spike-and-slab* with importance sampling on π (VIEM-SS-IS), MCMC *spike-and-slab* (MCMC-SS) and glmnet [15] (LASSO). Values averaged over 50 replicated simulated datasets (standard deviation in the parentheses).

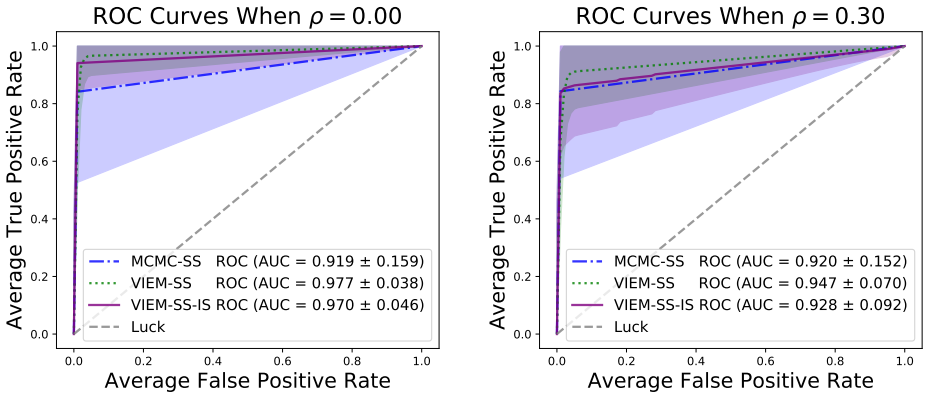
	MCMC-SS	VIEM-SS	VIEM-SS-IS	LASSO
		$\rho = 0.0$		
Precision	0.755 (0.227)	0.460 (0.274)	0.945 (0.103)	0.100 (0.054)
Recall	0.840 (0.320)	0.965 (0.071)	0.925 (0.108)	0.912 (0.199)
MCC	0.799 (0.242)	0.639 (0.205)	0.931 (0.075)	0.275 (0.069)
AUC	0.919 (0.160)	0.977 (0.039)	0.970 (0.047)	0.014 (0.014)
F1	0.817 (0.203)	0.584 (0.240)	0.928 (0.079)	0.168 (0.068)
ACC	0.997 (0.003)	0.988 (0.011)	0.999 (0.001)	0.005 (0.002)
Time	1039.920 (94.772)	14.256 (29.387)	82.463 (177.261)	0.815 (0.109)
		$\rho = 0.3$		
Precision	0.773 (0.221)	0.424 (0.301)	0.814 (0.284)	0.117 (0.046)
Recall	0.842 (0.308)	0.909 (0.132)	0.854 (0.184)	0.921 (0.124)
MCC	0.812 (0.221)	0.588 (0.240)	0.818 (0.227)	0.310 (0.064)
AUC	0.920 (0.154)	0.947 (0.070)	0.928 (0.093)	0.012 (0.012)
F1	0.801(0.234)	0.534 (0.271)	0.807 (0.242)	0.203 (0.071)
ACC	0.997 (0.003)	0.985 (0.014)	0.996 (0.009)	0.006 (0.002)
Time	1086.613 (163.832)	14.228 (20.609)	74.665 (107.522)	0.782 (0.110)

Dirichlet-multinomial example

Finally, we conducted a simulation study to assess performances of the Bayesian Dirichlet-multinomial regression model for multivariate responses. We used the approximate variational method described in this chapter and the MCMC posterior sampling of [51], which employs *spike-and-slab* priors. We also considered the penalized likelihood approach of [9]. We simulated data with $n = 100$, $J = 50$ and $p = 50$. More specifically, for each sample $i = 1, \ldots, n$, we generated a matrix of covariates $\boldsymbol{x}_i \sim \text{Normal}\,(0, \Sigma)$, where the (l, m)-th entry of the covariance matrix was set to be $\Sigma_{lm} = \rho^{|l-m|}$ for $l \neq m$. Here, we set $\rho = 0.4$. The responses were sampled from a Multinomial-Dirichlet regression model of the type

$$\boldsymbol{y}_i \sim \text{Multinomial}(y_{i+}, \boldsymbol{\phi}_i) \tag{1.51}$$

$$\boldsymbol{\phi}_i \sim \text{Dirichlet}(\xi_{i1}, ..., \xi_{i50}), \tag{1.52}$$



(a) $\rho = 0.0$ (b) $\rho = 0.30$

Figure 1.4: Negative Binomial - large p small n example: Comparison of selection performance (ROC curves). Variational Inference EM *spike-and-slab* (VIEM-SS), Variational Inference EM *spike-and-slab* with importance sampling on π (VIEM-SS-IS), MCMC with *spike-and-slab* prior (MCMC-SS) and MCMC with horseshoe prior (MCMC-HS). The ROC curves and the corresponding standard deviations are averaged over 50 replicated datasets.

with $y_{i+} \sim \text{Uniform}\,(1000, 2000)$ as the observed total count of each sample, and where $\boldsymbol{\phi}_i$ denotes the 50×1 vector of multinomial parameters. In order to evaluate the effect of different assumptions about overdispersion in the data, we set the parameters of the Dirichlet prior by letting $\xi_{ij} = \frac{\xi_{ij}}{\xi_i^+} \times \frac{1-r}{r}$, $j = 1, \ldots, 50$, where small values of r lead to more overdispersed data. ξ_{ij} was associated to the covariates through a log link of the type

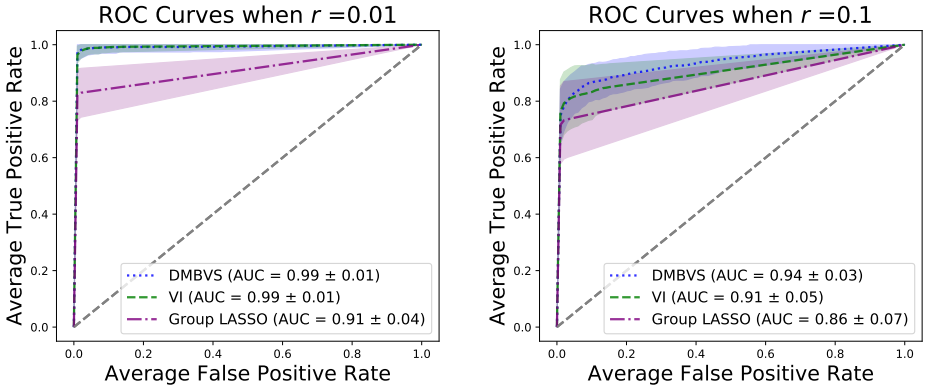
$$\log(\xi_{ij}) = \alpha_j + \sum_{k=1}^{p} \beta_{kj}\, x_{ik}, \tag{1.53}$$

with intercept $\alpha_j \sim \text{Uniform}\,(-2.3, 2.3)$, similarly as in [51] and [9].

Table 1.3 reports the results for precision, recall, MCC, AUC, F1 score, and accuracy, averaged across 50 replicated datasets, with standard deviations in parentheses, and Figure 1.5 reports the ROC curves. For the Bayesian methods, in each dataset relevant associations were selected to ensure a Bayesian false discovery rate (FDR) control of 0.1. Results show that the proposed variational Bayes approach performs comparably with the MCMC approach, although it is characterized by lower recall values. The performance of the penalized group LASSO appears to degrade with in-

Table 1.3: Dirichlet-multinomial example: Performance comparisons of variational inference with non-local prior (VI), MCMC spike-and-slab (MCMC) and the Group penalized lasso approach (Group LASSO). The selection performance is evaluated using accuracy, recall, precision, F1 score and Matthews correlation coefficient (MCC), all averaged over 50 replicated simulated datasets (standard deviation in parentheses).

	$r = 0.01$			$r = 0.1$		
	DMBVS	VI	Group LASSO	DMBVS	VI	Group LASSO
Precision	0.99 (0.02)	0.95 (0.06)	0.60 (0.07)	0.98 (0.04)	0.76 (0.11)	0.33 (0.07)
Recall	0.48 (0.10)	0.41 (0.13)	0.81 (0.09)	0.28 (0.14)	0.44 (0.10)	0.63 (0.14)
MCC	0.68 (0.08)	0.61 (0.10)	0.69 (0.07)	0.51 (0.14)	0.57 (0.08)	0.48 (0.10)
AUC	0.99 (0.01)	0.99 (0.01)	0.90 (0.04)	0.94 (0.03)	0.91 (0.05)	0.86 (0.10)
F1	0.64 (0.10)	0.56 (0.13)	0.68 (0.07)	0.42 (0.17)	0.55 (0.09)	0.45 (0.09)
ACC	0.99 (0.001)	0.99 (0.001)	0.99 (0.002)	0.99(0.001)	0.99 (0.001)	0.98 (0.004)



(a) $r = 0.01$ (b) $r = 0.10$

Figure 1.5: Dirichlet-multinomial example: Comparison of selection performance (ROC curves). DMBVS, VI (variational inference) and Group LASSO. The ROC curves and the corresponding standard deviations are averaged over 50 replicated datasets.

creasing overdispersion.

6. Benchmark Applications

Next, we show performances of the methods on some benchmark applications and case study data. In particular, we use the well know Boston housing dataset for an application of the negative Binomial model and apply the Dirichlet-multinomial model

to a case study dataset on microbiome data.

Boston Housing Data

The Boston housing dataset, collected by the U.S Census Service, can be obtained from the **StatLib** archive at http://lib.stat.cmu.edu/datasets/boston, and has been used extensively to benchmark different algorithms. The dataset consists of 506 observations on 14 variables. Here we use the non-negative attribute **medv** (median value of owner-occupied home in $1,000) as the outcome and the remaining 13 features as predictors. We preprocessed the data by standardizing the predictors to account for the different units of measurement. We also created a larger dataset by adding 300 noise random features sampled from a standard Gaussian distribution.

For this dataset, we focused in particular on the predictive accuracy of the method and consider prediction results averaged over 100 random splits of the whole dataset into training (405 observations, 80%) and validation (101 observations, 20%) sets. To test the goodness-of-fit, we evaluate the widely used metric in GLMs called Pearson residuals on the training set,

$$E = \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i \left(1 + \hat{\kappa}\hat{\mu}_i\right)}} \right)^2,$$

where $\hat{\mu}$ and $\hat{\kappa}$ are the estimated mean and quasi-dispersion ($\hat{\kappa} = 0$ for Poisson and $\hat{\kappa} = \hat{r}^{-1}$ for the negative binomial regression models). We also compute the root mean squared predictive error (RMSPE) on the testing set. We compared performances of the two variational based algorithms (VIEM-SS and VIEM-SS-IS), the two sampling based algorithms (MCMC-HS and MCMC-SS) and the LASSO method. For the Bayesian methods, we used the same hyper-parameter setting as in the simulation study and ran 13,000 Gibbs sampling iterations with the initial 3,000 samples discarded as burnin. For the variational algorithms, we terminated them when changes of the ELBO was less than 0.001. For the LASSO method, we again used the default **cv.glmnet** function [15] with cross-validation.

Results are summarized in Table 1.4, where we again observe that the two variational methods achieve similar performance than the MCMC methods, but at a much faster computational speed. In terms of goodness of fit measured by Poisson residuals, the LASSO based Poisson model performs the worst due to its unrealistic equal-dispersion assumption, while the negative binomial model significantly improves the performance when assuming a gamma distributed multiplicative random effect term r [56]. When looking into the variable selection performances, we noticed that all Bayesian methods would choose **lstat** as the only important feature, while LASSO tended to include more covariates in the model (results not shown).

Table 1.4: Boston Housing Data: completion times in second, Pearson residuals and RMSPE. Values averaged over 100 random splits of the whole dataset into training and validation sets (standard deviations in parentheses).

Methods	Time(s)	Pearson Residuals	RMSPE
Small Dataset (p = 13)			
MCMC-HS	98.280 (13.190)	34.705 (1.518)	6.063 (0.702)
MCMC-SS	68.830 (3.860)	37.670 (3.508)	5.849 (0.807)
VIEM-SS	0.005 (0.002)	38.779 (2.030)	5.730 (0.594)
VIEM-SS-IS	0.036 (0.005)	38.841 (1.885)	5.729 (0.592)
LASSO (1SE)	0.182 (0.026)	356.213 (25.330)	4.413 (0.650)
LASSO (MIN)	0.183 (0.018)	307.235 (17.790)	4.171 (0.576)
Large Dataset (p = 313)			
MCMC-HS	211.260 (21.350)	37.016 (1.916)	5.862 (0.629)
MCMC-SS	486.810 (77.710)	37.815 (2.538)	5.765 (0.595)
VIEM-SS	0.050 (0.010)	39.035 (1.959)	5.752 (0.635)
VIEM-SS-IS	0.400 (0.030)	38.913 (1.803)	5.728 (0.591)
LASSO (1SE)	3.541 (0.344)	418.373 (31.456)	4.831 (0.622)
LASSO (MIN)	3.561 (0.319)	344.157 (26.746)	4.510 (0.645)

Microbiome Data

We apply our variational method with non-local prior to a human gut microbiome dataset, which has been previously used in [52] to investigate the association of dietary and environmental variables with the gut microbiota. Here, the multivariate outcome $\boldsymbol{y}_i$ represents the vector of counts obtained as the taxonomic abundances of q taxa. More specifically, the dataset contains microbiome 16S rDNA sequencing data from a cross-sectional analysis of $n = 98$ healthy volunteers. The original microbiome abundance table contained 3068 OTUs (excluding the singletons), which were further combined into 127 genera. More specifically, here we follow [9] and consider a subset of 30 relatively common genera that appeared in at least 25 subjects. Diet information was also collected on all subjects, using a food frequency questionnaire and then converting to nutrient intake values, which were summarized in a $n = 98 \times p = 117$ matrix of representative nutrients. We considered the squared root transformed values of taxa abundance, similarly as in [9].

We applied the Dirichlet-multinomial model, with non-local priors and VI inference. Our method selected 4 genera and 8 nutrient types, after controlling for a Bayesian FDR of 0.1, corresponding to a posterior probability of inclusion of 0.745.

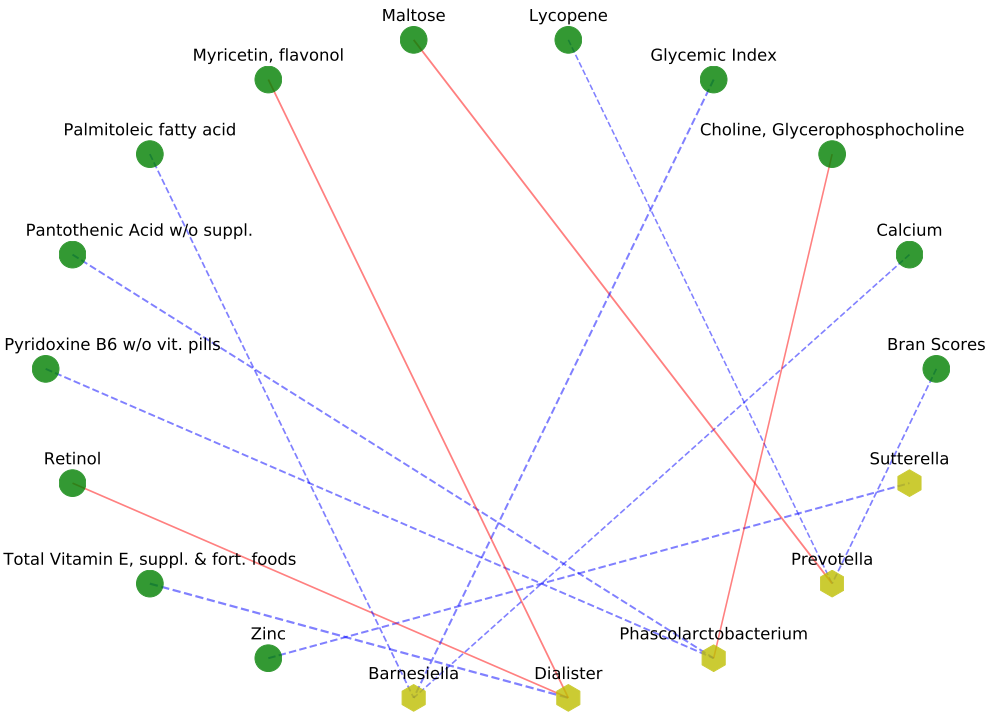


Figure 1.6: Microbiome data: Bipartite graph of selected taxa-covariate associations. The blue dashed lines denote negative associations; the red solid lines denote positive associations.

Selected associations are visualized in a bipartite graph in Figure 1.6. Similarly as in [52], Prevotella is found to be highly associated with maltose, which is a common disaccharide, indicative of a high carbohydrates diet. At the same time, Prevotella is found to be negatively associated with nutrients typical of a high fiber diet, a finding which has also been confirmed in the literature (see, e.g., [22]). Also, increased Barnesiella has been linked to diets rich in gluten, characterized by high glycemic index [29]. The Group penalized Lasso approach selected a larger number of significant associations, involving 12 nutrient types and 10 genera (result not shown).

7. Conclusion

We have developed Bayesian variable selection approaches using variational inference for the Negative Binomial and Multinomial Dirichlet regression models. For the NB model, we have introduced two data augmentation schemes to obtain deterministic

update rules for the parameters of interest via variational EM approaches. For the DM model, we have proposed a low-variance stochastic gradient method to optimize the ELBO objective. The variational algorithms we have developed can be applied to other Bayesian regression settings, with variable selection. We have shown on simulated data that the variational schemes have similar selection perfromance as the sampling-based MCMC methods.

Some of the shortcomings of the variational approach can be explained by the approximating family distributions. While the proposed factorization in Equation (1.25) allows for a tractable closed form computation, the independence assumption can cause the model to underestimate the posterior variance of the latent variables. In situations with correlated explanatory variables, the performance is sensitive to initialization and can be subject to poor optima. To overcome the problems mentioned above, attempts have been made to specify an expressive variational distribution while maintaining efficient computation [39] and to make posterior inference robust to initialization via by constraining the optimization path [2].

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
2. Altosaar, J., Ranganath, R., and Blei, D. M. (2017). Proximity variational inference. *arXiv preprint arXiv:1705.08931*.
3. Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174.
4. Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877.
5. Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):627–641.
6. Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.
7. Carbonetto, P., Stephens, M., et al. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.
8. Chandra, R., Dagum, L., Kohr, D., Maydan, D., Menon, R., and McDonald, J. (2001). *Parallel programming in OpenMP*. Morgan kaufmann.
9. Chen, J. and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1).
10. Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., and Stine, R. A. (2001). The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134.
11. Durrett, R. (2010). *Probability: theory and examples*. Cambridge University Press.
12. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.
13. Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101.
14. Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.
15. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.
16. George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian Variable selection. *Statistica Sinica*, 7, 339-373.
17. Griffin, J. E., Brown, P. J., et al. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.
18. Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401.
19. Huang, J. C., Morris, Q. D., and Frey, B. J. (2007). Bayesian inference of MicroRNA targets from sequence and expression data. *Journal of Computational Biology*, 14(5):550–563.
20. Huang, X., Wang, J., and Liang, F. (2016). A variational algorithm for Bayesian variable selection. *arXiv preprint arXiv:1602.07640*.
21. Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with Gumbel-Softmax. *CoRR*, abs/1611.01144.
22. Jang, H. B., Choi, M.-K., Kang, J. H., Park, S. I., and Lee, H.-J. (2017). Association of dietary patterns with the fecal microbiota in korean adolescents. *BMC Nutrition*, 3(1):20.
23. Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170.
24. Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
25. Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.

26. Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, 11(1):58.
27. Louizos, C., Welling, M., and Kingma, D. P. (2018). Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*.
28. Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *CoRR*, abs/1611.00712.
29. Marietta, E. V., Gomez, A. M., Yeoman, C., Tilahun, A. Y., Clark, C. R., Luckey, D. H., Murray, J. A., White, B. A., Kudva, Y. C., and Rajagopalan, G. (2013). Low incidence of spontaneous type 1 diabetes in non-obese diabetic mice raised on gluten-free diets is associated with changes in the intestinal microbiome. *PloS one*, 8(11):e78687.
30. Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.
31. Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82.
32. Nathoo, F., Babul, A., Moiseev, A., Virji-Babul, N., and Beg, M. (2014). A variational Bayes spatiotemporal model for electromagnetic brain mapping. *Biometrics*, 70(1):132–143.
33. O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117.
34. Ormerod, J. T., You, C., Müller, S., et al. (2017). A variational Bayes approach to variable selection. *Electronic Journal of Statistics*, 11(2):3549–3594.
35. Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.
36. Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538.
37. Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.
38. Quenouille, M. H. (1949). A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics*, 5(2):162–164.
39. Ranganath, R., Tran, D., and Blei, D. (2016). Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333.
40. Roberts, S. J. (2010). *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature*. PhD thesis, University of Oxford.
41. Ročková, V. (2018). Particle EM for variable selection. *Journal of the American Statistical Association*, pages 1–14.
42. Ruiz, F. R., AUEB, M. T. R., and Blei, D. (2016). The generalized reparameterization gradient. In *Advances in neural information processing systems*, pages 460–468.
43. Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
44. Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2):1053.
45. Stepleton, T. (2008). Understanding the "Antoniak equation". *Unpublished manuscript*.
46. Stingo, F. C. et al. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *The Annals of Applied Statistics*, 4(4):2024–2048.
47. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
48. Titsias, M. K. (2017). Learning model reparametrizations: Implicit variational inference by fitting MCMC distributions. *arXiv preprint arXiv:1708.01529*.
49. Titsias, M. K. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems*, pages 2339–2347.
50. Van Dyk, D. A. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796.
51. Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). An integrative Bayesian dirichlet-multinomial regression model for the analysis of taxo-

nomic abundances in microbiome data. *BMC Bioinformatics*, 18(1):94.

52. Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108.
53. Zhang, C.-X., Xu, S., and Zhang, J.-S. (2018). A novel variational Bayesian method for variable selection in logistic regression models. *Computational Statistics & Data Analysis*.
54. Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., Vannucci, M., et al. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. *The Annals of Applied Statistics*, 10(2):638–666.
55. Zhou, M. and Carin, L. (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320.
56. Zhou, M., Li, L., Dunson, D., and Carin, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Proceedings of the International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access.
57. Zhou, Q., Guan, Y., et al. (2018). Fast model-fitting of Bayesian variable selection regression using the iterative complex factorization algorithm. *Bayesian Analysis*.
58. Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.