Effect of Predictor Dependence on Variable Selection for Linear and Log-Linear Regression

Apu Chandra Das
University of Arkansas, Fayetteville

Effect of Predictor Dependence on Variable Selection for Linear and Log-Linear Regression

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Apu Chandra Das

University of Dhaka
Bachelor of Science in Statistics, 2017

July 2020
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Jyotishka Datta, PhD
Thesis Director

Avishek Chakraborty, PhD
Committee Member

Giovanni Petris, PhD
Committee Member

Abstract

We propose a Bayesian approach to the Dirichlet-Multinomial (DM) regression model, which uses horseshoe, Laplace, and horseshoe plus priors for shrinkage and selection. The Dirichlet-Multinomial model can be used to find the significant association between a set of available covariates and taxa for a microbiome sample. We incorporate the covariates in a log-linear regression framework. We design a simulation study to make a comparison among the performance of the three shrinkage priors in terms of estimation accuracy and the ability to detect true signals. Our results have clearly separated the performance of the three priors and indicated that the horseshoe plus prior outperforms both horseshoe and Laplace priors under low dependence for the compositional data model in the Dirichlet-Multinomial regression framework. We have also seen that heavy dependence among the covariates reduces the rate of variable selection and deteriorates the estimation errors compared to low dependence.

Keywords Variable selection, Dirichlet-Multinomial regression, Horseshoe, Horseshoe plus, Laplace, Microbiome data, Overdispersion

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Jyotishka Datta for the consistent assistance during this thesis preparation and beyond. His patience, encouragement, knowledge, and guidance helped me in all the time of research and writing of this work. I could not have imagined having a better advisor and mentor.

I would also like to thank the faculty and staff at the Department of Mathematical Sciences, and especially those affiliated with the Statistics program for their encouragement, teaching, and advice. I also want to thank the committee members Dr. Avishek Chakraborty and Dr. Giovanni Petris.

Finally, I must express my very cordial gratitude to my beloved parents and my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. I want to thank everyone at the University of Arkansas for being so supportive throughout the two years of my master's studies. Thanks to all the faculty and staff for enabling my education goals and providing opportunities for me to be successful. An exceptional thanks to the Director of the Statistics track Dr. Giovanni Petris.

Table of Contents

Chapter 1

Introduction

1.1 Motivation and Background

High-dimensional data with many predictors and comparatively lower sample size have become routine across many modern scientific disciplines as a result of the rapid advances in high-throughput experiments, such as imaging or genomic sequencing for disease models. As a response to this changing landscape, statistical methodologies have evolved to adapt to the 'wide-data' paradigm, i.e. the large p, small n problem, spearheaded by Lasso and its many relatives in the frequentist regime [see e.g. Tibshirani et al., 2005, for a comprehensive review]. On the Bayesian side, a similar growth of methodologies has happened, starting with spike-and-slab priors [Mitchell & Beauchamp, 1988] and culminating in the more recent and popular one-group continuous shrinkage priors also called global-local shrinkage priors, such as the horseshoe prior [Bhadra et al., 2019b, Carvalho et al., 2010]. Inspired by the success of these methods, there has been an explosive development of methodological research in the area of high-dimensional regression and shrinkage methods over the last 15 years.

In spite of this remarkable progress, the performance of shrinkage priors in high-dimensional inference involving non-Gaussian likelihood has not received sufficient attention from the statistical community, perhaps except for sparse Poisson models for count data [Datta & Dunson, 2016]. Evidently, there is a significant applied interest to extend the inferential capacity of the global-local priors to the analysis of compositional data using a modeling framework such as an integrated Dirichlet-Multinomial model. These methods have the potential to be useful for analysis of both high-dimensional compositional data as well as material sciences where compositional data are routinely observed [Holmes et al., 2012, Wadsworth et al., 2017]. The issue of correlation among predictors in Bayesian high-dimensional models is also a relatively unexplored area. We also consider this issue via our simulation studies, in particular the effect on model selection accuracy.

The global-local priors [Bhadra et al., 2016, Carvalho et al., 2010, Polson & Scott, 2010b, 2012] have emerged as a popular and successful method for performing shrinkage in a wide variety of models when there exists high dependence among the covariates. This method can shrink small signals while keeping relatively large signals unshrunk in different models. Sparsity indicates only a few numbers of large signals among a myriad number of noisy observations very close to zero. The purpose of the high-dimensional analysis is to recover the significant low-dimensional signals observed in noisy observations under strong dependence.

The outline of this thesis is as follows. We discuss the effect of strong correlation among predictors on regularization as well as Bayesian shrinkage methods in chapter 1. Next, in chapter 2, we propose a new hierarchical model for learning the sparse parameter structure in an integrated Dirichlet-Multinomial model and show that the global-local shrinkage priors outperform the Bayesian Lasso. We conclude this thesis in chapter 3, with some pointers for future directions of research.

1.1.1 Brief review of shrinkage and selection methods

Consider a linear regression model with independent and identically distributed Gaussian errors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n), \tag{1.1.1}$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ is the vector of response and $\mathbf{x}_j = (x_{1j}, x_{2j}, ..., x_{nj})^T$, $j = 1, 2, \ldots, p$, are covariates in (1.1.1), with $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p]$ being the design matrix. If $p < n$, then ordinary least square (OLS) method is sufficient to estimate all nonzero coefficients. However, if p is far greater than n ($p \gg n$) we cannot use OLS method as the design matrix is no longer a full column rank matrix. In this case, we have to use a suitable regularization method or sparsity-favoring prior to be able to draw inference.

We start by describing the most popular frequentist regularization methods, namely Lasso and adaptive Lasso and the Bayesian sparsity-inducing priors, namely Bayesian Lasso and horseshoe prior.

Regularized Regression: Ridge and Lasso As we described above, the ordinary least squares solution $\hat{\theta}_{OLS}$ is not well-defined for rank-deficient design matrices $\mathbf{X}$. An early solution to address this issue was Ridge regression [Tikhonov, 1963], where a penalty term involving the ℓ_2 norm of the parameter vector is imposed on the negative log-likelihood or the residual sum of squares. Mathematically, the Ridge regression can be written as the following optimization rule:

$$\hat{\boldsymbol{\theta}}^{\text{ridge}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \right\} \tag{1.1.2}$$

$$= \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\text{argmin}} \left\{ ||\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}||_2^2 + \lambda \, ||\boldsymbol{\theta}||_2^2 = \text{RSS} + \lambda \sum_{j=1}^{p} {\theta_j}^2 \right\}, \tag{1.1.3}$$

where RSS denotes the residual sum of squares. It is easy to see that this optimization routine is equivalent to minimizing the residual sum of squares subject to the constraints $||\boldsymbol{\theta}||_2^2 \leq c$. The tuning parameter λ in (1.1.3) decides the amount of penalty to be imposed on $\boldsymbol{\theta}$, and is usually chosen by a k-fold cross-validation method. One of the key advantages of ridge regression is that the solution can be written in an easy analytically closed form:

$$\hat{\boldsymbol{\theta}}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}. \tag{1.1.4}$$

A couple of things can be directly observed from (1.1.3) and (1.1.4): (1) if $\lambda \to \infty$, the solution converges to a degenerate zero and $\lambda \to 0$, will take the ridge solution towards the OLS, and (2) the estimate in (1.1.4) can be also written as the posterior mode under a component-wise Gaussian prior on $\boldsymbol{\theta}$, i.e. $\pi(\theta_j) \propto \exp\{-\lambda\theta_j^2\}$, for all $j = 1, \dots, n$.

A drawback of ridge regression is that it does not perform any automatic variable selection, i.e. the solution path at every point includes all of the p covariates in the model regardless of their magnitude. The penalty $\lambda \sum \theta_j^2$ can push the insignificant covariates towards zero, but it can not set any of them exactly to zero (unless $\lambda = \infty$). The lasso, on the other hand, is an alternative to ridge regression which does not have the above caveat. The lasso coefficients, $\hat{\theta}_\lambda^L$ minimize the following quantity:

$$\hat{\boldsymbol{\theta}}^{\text{lasso}} = \sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{p}\theta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\theta_j| = RSS + \lambda\sum_{j=1}^{p}|\theta_j|, \tag{1.1.5}$$

where λ is a tuning parameter to be chosen using a cross-validation technique. The amount of bias increases if λ increases and variance decreases. Unlike ridge regression, the lasso uses an ℓ_1 penalty instead of ℓ_2. For a coefficient vector θ the ℓ_1 norm is denoted as $||\theta||_1 = \sum|\theta_j|$ and it corresponds to a Laplace or double-exponential prior. One of the benefits of ℓ_1 norm in the case of lasso is that it shrinks some of the coefficient estimates to be exactly zero when the value of tuning parameter λ is suitably chosen. Under suitable regularity conditions, such as the theta-min condition and neighborhood stability condition, lasso attains the oracular risk in both estimation and model selection. Informally, the theta-min condition warrants that the true non-zero coefficients are not too close to zero, and the neighborhood stability condition can be represented as a strong irrepresentability condition that restricts the degree of dependence between important and unimportant predictors or columns of the design matrix [Zhao & Yu, 2006]. For various optimality results concerning lasso and other variants, we direct the readers to Bühlmann & van de Geer [2011].

The main disadvantage of the lasso stems from the fact that its performance degrades rapidly when the predictors exhibit a high degree of dependence. Zhao & Yu [2006] showed that the probability of selecting the true model for lasso can drop to near zero when the strong irrepresentability condition is violated in the data.

Adaptive Lasso Undoubtedly, lasso is a popular shrinkage method of variable selection, but it cannot be an oracle procedure [Fan & Li, 2001] and also uses the same weights for all of the coefficients to be equally penalized in the ℓ_1 penalty. The oracle property represents that the method is able to identify the nonzero coefficients correctly with probability converging to one, and estimators of the identified covariates are asymptotically normally distributed with the same means and covariance matrix that they would have if the zero coefficients were known before. An approach that helps to obtain a convex objective function that produces oracle properties in weighted ℓ_1 penalty with a data-dependent weights vector determined by an initial estimator [Zou, 2006]. We can employ different weights to different coefficients, and the weighted lasso is,

$$L_n(\theta) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left|\left|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \theta_j\right|\right|^2 + \lambda \sum_{j=1}^{p} w_j |\theta_j| \tag{1.1.6}$$

where w_j is a data-dependent weights vector. Zou [2006] showed that if the weights vector is chosen adroitly, then the weighted lasso can follow oracle properties. The value $\hat{\theta}_n$ that minimizes L_n is known as adaptive lasso. Compared with the regular lasso, adaptive lasso aims to reduce the estimation bias and increase variable selection accuracy at the same time by granting a relatively higher penalty for zero coefficients and a smaller penalty for non-zero coefficients.

Bayesian Lasso: Tibshirani [1996] pointed out that the Lasso estimate can be interpreted in a Bayesian framework as the posterior mode under component-wise i.i.d. double exponential priors on each θ_i, $i = 1, \ldots, p$.

$$\hat{\theta}_{\text{BLasso}} = \underset{\theta}{\operatorname{argmax}}\ p(\theta \mid\ y, \sigma^2, \tau).$$

When prior distribution is $p(\theta \mid \tau) = (\tau/2)^p \exp(-\tau \, ||\theta||_1)$ and likelihood is $p(y \mid \ \theta, \sigma^2) = N(\mathbf{y} \mid \mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 > 0, \ \tau > 0$, the posterior mode of θ is identical to the lasso estimate with penalty $\lambda = 2\tau\sigma^2$. Park & Casella [2008] showed that the posterior sampling for Bayesian Lasso can be done easily by writing the Laplace density as a Normal scale mixture, facilitating a Gibbs sampler. Park & Casella [2008] also argued that the usual i.i.d. Laplace prior can lead to a bimodal posterior density, that can be prevented by conditioning on σ^2. With this, the prior on θ can be written as:

$$f(\theta \mid \sigma^2) = \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\theta_j|/\sqrt{\sigma^2}}. \tag{1.1.7}$$

While the posterior mode for Bayesian Lasso will have the same optimality properties as the frequentist Lasso estimator, the whole posterior distribution does not share the same optimality properties. Firstly, the Laplace prior has heavy tails and a bounded density near the origin (see Fig. 1.1) that are shown to lead to sub-optimal shrinkage and a non-vanishing bias in the tails [Polson & Scott, 2012]. On the other hand, Castillo et al. [2015] showed that the whole posterior distribution for Laplace prior concentrates around the truth at a sub-optimal rate, unlike the mode: as a result, the posterior distribution becomes useless for uncertainty quantification.

Horseshoe Prior The inadequacies of a Laplace prior are corrected by the horseshoe prior, introduced in Carvalho et al. [2009, 2010], and further investigated in a series of papers Bhadra et al. [2019b], Datta & Ghosh [2013], Polson & Scott [2010a, 2012], van der Pas et al. [2016]. Here the main idea is to use a Gaussian scale mixture prior on the parameter $\boldsymbol{\theta}$ such that the marginal prior has an infinite pole at zero and heavy tails. The hierarchical model for the horseshoe prior is given by:

$$Y_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1), \tag{1.1.8}$$

$$(\theta_i \mid \lambda_i, \tau) \sim \mathcal{N}(0, \lambda_i^2 \tau^2), \tag{1.1.9}$$

$$\lambda_i \sim \mathcal{C}a^+(0,1), \tau \sim \mathcal{C}a^+(0,1), \tag{1.1.10}$$

where $\mathcal{C}a^+(0,1)$ is a half-Cauchy distribution on a positive real line. Carvalho et al. [2009] argued that the horseshoe prior could mimic the superlative performance of a spike-and-slab prior [Mitchell & Beauchamp, 1988], that puts a discrete mixture of a point mass at zero and a continuous (usually Gaussian) prior for each θ_i. The equivalence can be easily demonstrated for the Gaussian sequence model in (1.1.8).

Now, consider the Gaussian sequence model: $Y_i = \theta_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_0^2)$, where θ_i is given a horseshoe prior in (1.1.10). It follows from (1.1.10) that θ_i given y_i and the hyper-parameters, has a posterior distribution which is normal with following mean and variance:

$$E(\theta_i \mid y_i, \lambda_i, \tau, \sigma^2) = (1 - \frac{1}{1 + \lambda_i^2 \tau^2}) y_i$$
$$V(\theta_i \mid y_i, \lambda_i, \tau, \sigma^2) = (1 - \frac{1}{1 + \lambda_i^2 \tau^2}) \sigma^2.$$

Assuming $\kappa_i = \frac{1}{1+\lambda_i^2\tau^2}$, the posterior mean of θ_i is $E(\theta_i \mid y_i, \lambda_i, \tau, \sigma^2) = (1 - \kappa_i) y_i$ and variance is $V(\theta_i \mid y_i, \lambda_i, \tau, \sigma^2) = (1 - \kappa_i)\sigma^2$ and hence by Fubini's theorem

$$E(\theta_i \mid y_i, \lambda_i, \tau, \sigma^2) = (1 - E(k_i \mid y_i, \lambda_i, \tau, \sigma^2)) y_i. \tag{1.1.11}$$

If we compare the posterior mean expression of (1.1.11) with that under a spike-and-slab prior it immediately becomes clear that the quantity ($\hat{\omega}_i = 1 - E(k_i \mid y_i, \lambda_i, \tau, \sigma^2)$) behaves like the posterior inclusion probability $\omega_i = P(\theta_i \neq 0 \mid y_i)$. This means that for the sequence model, we can use the quantities $\hat{\omega}_i$ as a thresholding rule for selecting important θ_i's. The weights $\hat{\omega}_i$ are called pseudo posterior inclusion probabilities. The origin of the nomenclature 'horseshoe' stems from the fact that a half-Cauchy prior on λ_i would lead to a U-shaped beta$(1/2, 1/2)$ prior on $\kappa_i = (\lambda_i^2\tau^2 + 1)^{-1}$. Datta & Ghosh [2013] showed that these pseudo inclusion probabilities could be used as a thresholding rule for multiple testing or parameter selection with optimal risk properties. For a sparse regression problem, the hierarchical model is similar to the one in (1.1.10), with the data generating model given by: $Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2)$. The theory for regression under horseshoe is not as well investigated as for the sparse normal means problem, but it has been shown that they work well under dependence [Bhadra et al., 2019b, Datta & Ghosh, 2015] and outperform common methods like ridge regression for prediction [Bhadra et al., 2019a].

The general class of global-local shrinkage priors was proposed in Polson & Scott [2012], and has seen a lot of developments over the last decade. The general family of global-local priors are defined as:

$$\theta_i \sim \mathcal{N}(0, \lambda_i^2\tau^2) \tag{1.1.12}$$

$$\lambda_i \sim \pi(\lambda_i), \tau \sim \pi(\tau), \tag{1.1.13}$$

where the parameters λ_i and τ are called local and global shrinkage parameters, respectively, as they help in tagging the large signals and adapting to the level of sparsity, respectively. Figure. 1.1 shows the general shape of the induced priors on θ_i, $p(\theta_i)$ under a general class of global-local shrinkage priors (1.1.13).

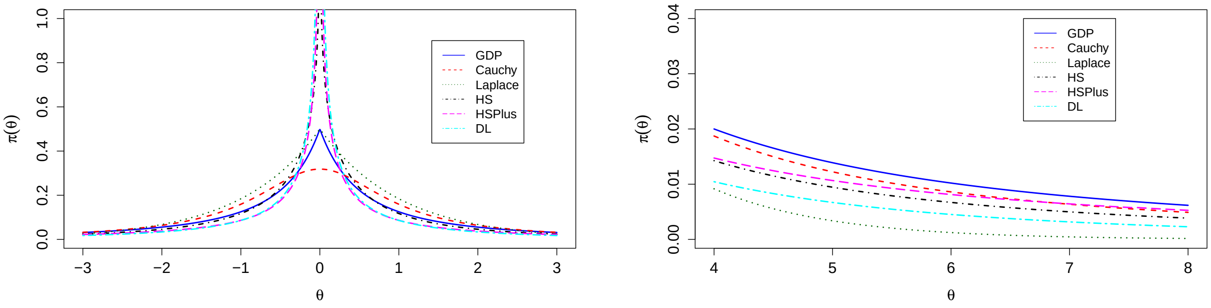


Figure 1.1: Probability density functions for Generalized Double Pareto, Cauchy, Laplace, Horseshoe, and Horseshoe+ prior.

The common characteristics are a peak near the origin and heavy tails. The peak helps in achieving sparsity, and the heavy tails are critical for robustness to large signals. Several theoretical optimality properties are known for the class of global-local shrinkage priors, such as risk optimality for multiple testing [Ghosh et al., 2016] and near-minimaxity [van der Pas et al., 2014, 2016] and adaptive uncertainty quantification [van der Pas et al., 2017]. For an extensive survey of these priors vis-a-vis Lasso and other regularized methods, we refer the readers to Bhadra et al. [2019b], and the references therein.

Horseshoe+ Prior The horseshoe+ prior, an extension of the horseshoe prior, has been successful in detecting and estimating sparse signals, and at the same time, achieved some theoretical properties while enjoying computational feasibility. Bhadra et al. [2017a] proved that the horseshoe+ posterior concentrates at a faster rate than the horseshoe, and also, the estimator has a lower posterior mean squared error in estimating true signals.

Considering $(y_i|\theta_i) \sim \mathcal{N}(\theta_i, 1)$,the horseshoe hierarchical model can be defined by the following set of conditional distributions

$$\begin{aligned}(\theta_i|\lambda_i, \tau) &\sim \mathcal{N}(0, \lambda_i^2\tau^2), \\ (\lambda_i|\tau) &\sim \mathcal{C}a^+(0, \tau),\end{aligned} \tag{1.1.14}$$

where C^+ stands for half-Cauchy distribution with scale parameter λ_i and density function

$$p(\lambda_i|\tau) = \frac{2}{\pi\tau\{1 + (\lambda_i/\tau)^2\}}, \tag{1.1.15}$$

The horseshoe+ hierarchical model extend this by adding an extra layer of shrinkage parameters:

$$\begin{aligned}(\theta_i|\lambda_i, \eta_i, \tau) &\sim \mathcal{N}(0, \lambda_i^2\tau^2), \\ (\lambda_i|\eta_i, \tau) &\sim \mathcal{C}a^+(0, \tau\eta_i),\end{aligned} \tag{1.1.16}$$

$$\eta_i \sim \mathcal{C}a^+(0,1),$$

where we have introduced the half-Cauchy distribution again mixing with variable η_i. The local shrinkage parameter λ_i's are conditionally independent given a new level of local shrinkage parameter η_i's, in addition to τ. The density of λ_i is as follows

$$p(\lambda_i|\tau) = \frac{4}{\pi^2 \tau} \frac{\log(\lambda_i/\tau)}{(\lambda_i/\tau)^2 - 1}. \tag{1.1.17}$$

The additional $\log(\lambda_i/\tau)$ term in the numerator leads to some different and surprising properties of the proposed estimator than its predecessor horseshoe.

We can write the horseshoe+ prior as a member of one-group global-local shrinkage priors in order to develop the distributional properties

$$p(\theta_i|\tau) = \int_0^\infty p(\theta_i|\lambda_i, \tau) p(\lambda_i|\tau) d\lambda_i.$$

Using the transformation $k_i = 1/(1 + \lambda_i^2 \tau^2)$ yields

$$p(\theta_i|\tau) = \int_0^1 p(\theta_i|k_i, \tau) p(k_i|\tau) dk_i, \text{with } p(\theta_i|k_i, \tau) \sim \mathcal{N}\left(0, \frac{1 - k_i}{k_i}\right),$$

where $k_i \in [0, 1]$ is a shrinkage weight. The induced prior density of κ_i, given below, will push more mass towards the extremities $\kappa_i = 0$ and $\kappa_i = 1$, ensuring stronger shrinkage compared to the horseshoe prior:

$$p_{HS+}(\kappa_i) = \frac{\tau}{\sqrt{\kappa_i(1 - \kappa_i)}} \frac{1}{(1 + \kappa_i(\tau^2 - 1))}.$$

Now, we turn to describing the hierarchical modeling framework for the Dirichlet–Multinomial regression.

Chapter 2

Dirichlet-multinomial regression

The human body is inhabited by lots of microorganisms such as bacteria, viruses, and some eukaryotes, and the number of microbial cells is approximately ten times than of the total number of human cells. Recent studies show that there is an association between the microbiome and human diseases such as obesity and diabetes [Virgin & Todd, 2011].

In this thesis, we use the sparse integrated Bayesian approach based on Dirichlet-Multinomial (henceforth referred to as DM) regression to find an association between available taxa counts and high-dimensional covariates from microbiome data. Recently, La Rosa et al. [2012] introduced the use of DM distributions for hypothesis testing and power calculations. Holmes et al. [2012] proposed a finite mixture of DM distributions to model the taxa counts directly. Chen & Li [2013] suggested that the likelihood ratio test can be used to test the effects of covariates on taxa proportions. Wadsworth et al. [2017] used the integrative Bayesian methodology regarding the use of DM distributions and spike-and-slab priors [Bogdan et al., 2011, Efron, 2008, 2010, Johnstone & Silverman, 2004] as global-local shrinkage priors for studying the association between the available covariates and taxa abundance. In this thesis, we propose a subtle modification of previous paper considering DM distributions and the horseshoe prior [Carvalho et al., 2010] as global-local shrinkage priors for the selection of significant associations between a set of available covariates and taxa. Even though both priors have their own advantages and downsides, the first one places a discrete mixture of a point mass at zero (the spike) and an absolutely continuous density (the slab) on each parameter, while the latter imposes absolutely continuous shrinkage priors on the whole parameter space that shrinks the small signals (close to zero).

2.1 Brief Review of Dirichlet-multinomial Regression

Assume there are J bacterial taxa available and their counts $Y = (Y_1, Y_2, Y_3, ..., Y_J)$ are random variables. $\mathbf{y} = (y_1, y_2, ..., y_J)$ denotes the vector of observed counts.
$\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_p})$ is a data matrix with order $n \times p$ which indicates obtained observations on P covariates. The count data y can be modeled as multinomial distribution and its probability function is given as,

$$\mathbf{y} \mid \boldsymbol{\phi} \sim \text{Multinomial}(y_+, \boldsymbol{\phi})$$

$$f_M(y_1, y_2, ..., y_J; \phi) = \binom{y_+}{y} \prod_{j=1}^{J} \phi_j^{y_{ij}} \tag{2.1.1}$$

where $y_+ = \sum_{j=1}^{J} y_j$ is sum of the all counts and $\boldsymbol{\phi}$'s are defined on the J dimensional simplex

$$S^{J-1} = \{(\phi_1, \phi_2, ..., \phi_J) : \phi_j \geq 0, \forall_j, \sum_{j=1}^{J} \phi_j = 1\},$$

The mean and variance of the above multinomial distribution are:

$$E(Y_j) = y_+\phi_j, Var(Y_j) = y_+\phi_j(1 - \phi_j) \tag{2.1.2}$$

The actual variation is larger than what we would predict by the multinomial distribution for microbiome composition data because of the heterogeneous characteristics of the microbiome samples and the proportions vary among samples. We impose a conjugate Dirichlet prior on the underlying proportions $(\phi_1, \phi_2, ..., \phi_j)$, where the proportions themselves are positive random variables$(\Phi_1, \Phi_2, ..., \Phi_j)$ with a constraint $\sum_{j=1}^{J} \Phi_j = 1$.

$$f_D(\Phi_1, \Phi_2, ..., \Phi_j; \gamma) = \frac{\Gamma(\gamma_+)}{\prod_{j=1}^{J} \Gamma(\gamma_j)} \prod_{j=1}^{J} \phi_j^{\gamma_j - 1}, \tag{2.1.3}$$

where $\gamma = (\gamma_1, \gamma_2, ..., \gamma_J)$ is a set of J-dimensional vector of strictly positive parameters, $\gamma_+ = \sum_{j=1}^{J} \gamma_j$ and $\Gamma(.)$ is the Gamma function. Dirichlet component $\Phi(j = 1, ..., J)$ has the following mean and variance:

$$E(\Phi_j) = \frac{\gamma_j}{\gamma_+}, Var(\Phi_j) = \frac{\gamma_j(\gamma_+ - \gamma_j)}{(1 + \gamma_+)\gamma_+^2} \tag{2.1.4}$$

There is an obvious benefit of the above hierarchical formulation and conjugacy helps to integrate $\boldsymbol{\phi}$ out, achieving the Dirichlet-Multinomial distribution [Mosimann, 1962], $\mathbf{y} \sim$ DM(γ), with the following probability mass function,

$$\begin{aligned} f_{DM}(y; \gamma) &= \int f_M(y_{i1}, y_{i2}, ..., y_{iJ}; \phi_i) f_D(\phi; \gamma) d\phi \\ &= \binom{y_+}{y} \frac{\Gamma(y_+ + 1)\Gamma(\gamma_+)}{\Gamma(y_+ + \gamma_+)} \times \prod_{j=1}^{J} \frac{\Gamma(y_j + \gamma_+)}{\Gamma(\gamma_j)\Gamma(y_j + 1)} \end{aligned} \tag{2.1.5}$$

where $\gamma_+ = \sum_{j=1}^{J} \gamma_j$. For overdispersed multivariate count data, the DM(γ) has more flexibilty than Multinomial distribution.

The above formulated Dirichlet-Multinomial (DM) posterior distribution has the following mean and variance:

$$E(Y_j) = y_+ E(\Phi_j), Var(Y_j) = y_+ E(\Phi_j)\{1 - E(\Phi_j)\} \left(\frac{y_+ + \gamma_+}{1 + \gamma_+}\right) \tag{2.1.6}$$

Comparing 2.1.6 with 2.1.2, we see that the variance of the DM distribution is increased by a factor of $(y_+ + \gamma_+)/(1 + \gamma_+)$, where the over-dispersion is taken into account of by the term γ_+ with a bigger value indicating less over-dispersion [Chen & Li, 2013].

For a given microbial sample, the DM model without the covariate term can be implemented to create more accurate estimates of taxa proportions than a multinomial model because of its ability to deal with over-dispersion. Along with the proportion estimation, ecologists are also interested in finding an association between microbiome composition and environmental covariates. Assume n microbiome samples are available with J species. Let $\mathbf{X} = (x_{ij})_{n \times p}$ be the microbiome data for n samples with p covariates and $\mathbf{Y} = (y_{ij})_{n \times J}$ be the observed count matrix for the n samples. We include the predictors into the regression model in a log-linear regression framework and assume that parameters $\gamma_j (j = 1, 2, ..., J)$ depend on the available covariates through the following model

$$\gamma_j(\mathbf{x}^i) = \exp\left(\alpha_j + \sum_{p=1}^{P} \theta_{jp} x_{ip}\right), \tag{2.1.7}$$

where $\mathbf{x}^i$ indicates the ith observation vector in the design matrix $\mathbf{X}$ and θ_{jp} represents the coefficient for the jth taxon with respect to the pth covariate which explains the effect of the pth covariate on jth taxon.

We can model the γ_j terms in a log-linear regression framework with a suitable shrinkage prior for achieving sparsity. Wadsworth et al. [2017] used a component-wise spike and slab prior for each θ_j for selecting strong associations. Here we propose to replace the spike-and-slab priors with the more efficient global-local priors for gaining better computational efficiency.

Equation 2.1.9 below shows the hierarchical model for applying horseshoe prior to the integrated Dirichlet-multinomial framework. We compare this model with two other priors, namely Bayesian Lasso [Park & Casella, 2008] and horseshoe+ [Bhadra et al., 2017b]. The hierarchical models for applying the two other candidate priors are very similar and not shown here.

$$\eta_{ij} = \log(\gamma_{ij}),\ \eta_{ij} = \theta_{0j} + \sum_{l=1}^{p} X_{il}\theta_{lj}, \quad i = 1, \ldots, N, \tag{2.1.8}$$

$$\theta_{lj} \sim \mathcal{N}(0, \lambda_{lj}^2 \tau_j^2),\ \lambda_{lj}^2 \sim \mathcal{C}a^+(0, \tau_j),\ \tau_j^2 \sim \mathcal{C}a^+(0, 1), l = 1, \ldots, p,\ j = 1, \ldots, J. \tag{2.1.9}$$

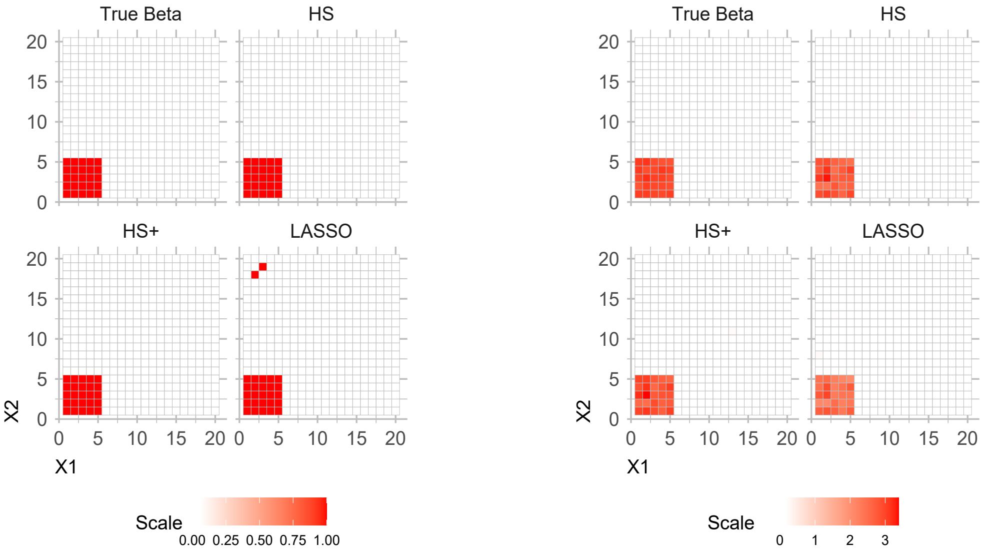
We use a Hamiltonian Monte Carlo based approach for posterior sampling using the popular Stan interface Carpenter et al. [2017]. The stan codes for applying horseshoe prior for the integrated Dirichlet-Multinomial model is provided in the Appendix §3.

2.2 Simulation Study

In this section, we carry out a simulation study to measure the performance of our proposed models and compare the results. In the simulation study, we used $n = 50, p = 20$, and $q = 20$. Here q and p represent the number of column in the response variable y and covariate matrix X respectively. Both q and p have only 5 non-zero columns which warrants their sparsity. Coefficient vector θ is generated from a normal distribution with mean 3 and standard deviation 0.1. Then an intercept vector is introduced in the model and simulated from a uniform distribution with some certain parameters $a = -2.3$ and $b = 2.3$. We generated the covariate matrix X from a Multivariate-Normal $(0, \Sigma)$ distribution where $\Sigma = \rho^{|i-j|}$ and ρ is set to 0.4 which is an indication of weak correlation among the predictors. Each response vectors, y_i has been drawn from a Dirichlet-Multinomial distribution, $\mathbf{y}_i \sim$ Multinomial $(N_i, \boldsymbol{\pi}_i^*)$, where $\boldsymbol{\pi}_i^* = (\pi_{i1}^*, \pi_{i2}^*, ..., \pi_{iJ}^*) \sim$ Dirichlet $(\boldsymbol{\gamma}^*)$. This design is repeated 50 times, and at each iteration we apply the horseshoe, Bayesian Lasso, and horseshoe plus 1000 times to each of the 50 generated models.

For the Bayesian methods, we run the Markov chain for 1000 samples, discarding the first 500 as a burn-in step and finally thinning every two samples. The goal of this simulation is to compare horseshoe, Bayesian Lasso, and horseshoe plus priors under Dirichlet-Multinomial (DM) setting.

Figure 2.1 shows the performance of horseshoe, horseshoe plus, and Bayesian Lasso in terms of variable selection in our proposed simulation design and clearly shows that the horseshoe and horseshoe plus select more true signals than the Bayesian Lasso.



(a) Variable selection

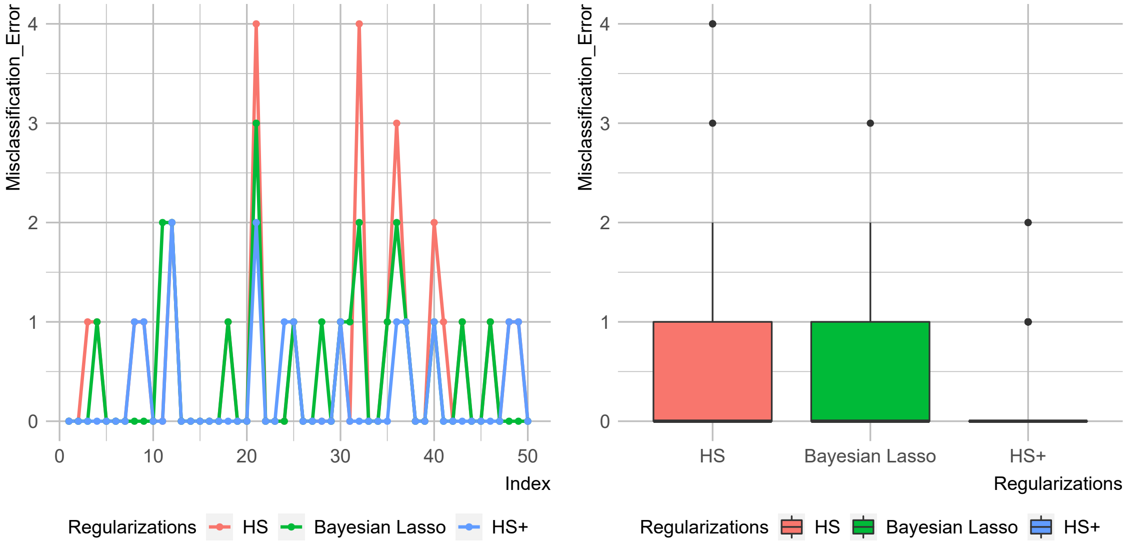
(b) Estimation accuracy

Figure 2.1: Comparison of estimated $\hat{\boldsymbol{\theta}}$ matrices and selected entries of $\hat{\boldsymbol{\theta}}$ matrices across the candidate priors for a single simulated data set

2.2.1 Comparison Under Low Dependency

Misclassification error has been calculated from the above simulation design to compare the performance of the proposed shrinkage priors in terms of variable selection. In this design, we used 50 iterations for horseshoe, Bayesian Lasso, and horseshoe plus methods. Then misclassification errors were calculated and sorted out to make the figure visually clear and understandable. Horizontal axis and vertical axis represent the number of iteration and misclassification errors, respectively. The following figure 2.2a shows that both horseshoe and horseshoe plus methods have fewer misclassification errors compared to the Bayesian Lasso, which indicates that they are more efficient than their competitor Bayesian Lasso in terms of true signals recovering.

The following figure 2.2b shows a summary statistics of misclassification errors for the three proposed methods. We can see that the value of the 1st quartile, median, and 3rd quartile are exactly zero for horseshoe plus except some outliers. The summary statistics are equivalent except some outliers for both horseshoe and Bayesian Lasso, which means they perform almost equally to recover the true signals but not better than horseshoe plus.

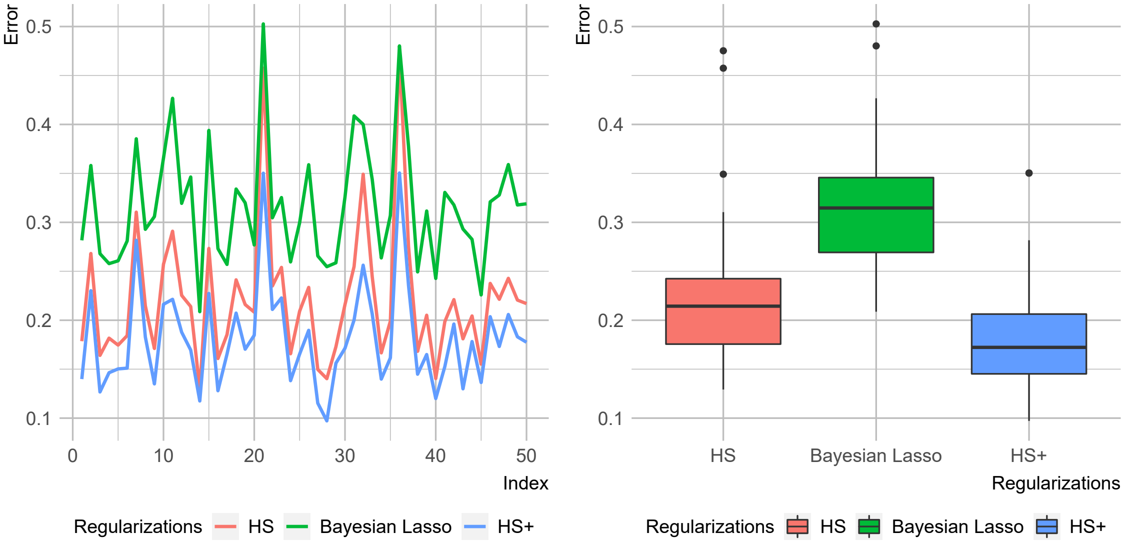


(a) Misclassification errors for each replication

(b) Boxplot of the misclassification errors

Figure 2.2: Comparison of the distributions and per-iteration misclassification errors among horseshoe, Bayesian Lasso, and horseshoe plus.

Figure 2.3 shows a distinct pattern among horseshoe, Bayesian Lasso, and horseshoe plus models and tells that errors for Bayesian Lasso are far larger than both horseshoe and horseshoe plus for all of the iterations. Compared to the horseshoe model, horseshoe plus errors are a bit lower, so it outperforms the other two models.



(a) Estimation errors for each replication

(b) Boxplot of the estimation errors

Figure 2.3: Comparison of the distributions and per-iteration estimation errors among horseshoe, Bayesian Lasso, and horseshoe plus.

The following table 2.1 is a list of estimation errors for several different numbers of iteration and shows that errors for horseshoe are close to the horseshoe plus but a bit larger. On the other hand, errors for Bayesian Lasso are fairly large compared to the other two.

Table 2.1: Estimation errors under low dependency

	Number of Iteration					
Regularizations	1	10	20	30	40	50
Horseshoe	0.179	0.257	0.208	0.216	0.140	0.217
Bayesian Lasso	0.282	0.366	0.277	0.326	0.243	0.319
Horseshoe Plus	**0.140**	**0.216**	**0.185**	**0.171**	**0.120**	**0.177**

Table 2.2 is a short summary statistics for the figure 2.3 and again shows that error mean and median for horseshoe plus are close to the horseshoe but a bit lower, on the other hand, somewhat smaller than Bayesian Lasso.

Table 2.2: Summary Statistics of the estimation errors

Regularizations	Mean	Median	Standard Deviation
Horseshoe	0.221	0.214	0.068
Bayesian Lasso	0.317	0.315	0.061
Horseshoe Plus	0.181	0.172	0.052

2.2.2 Correlation among predictors

One of the most common problems of high-dimensional data is that covariates are strongly dependent on each other. It is cumbersome to recover the true signals when the above situation arises. We do not know how it affects a log-linear regression model, and we are going to look at one simulation study to get an idea. For this simulation, we set $\rho = 0.9$, which warrants high dependence among covariates. The following figure 2.4 shows that Bayesian lasso correctly identifies more coefficients than horseshoe and horseshoe plus. Both methods recovered only 10 true signals out of 25, while the earlier one recovered 12.

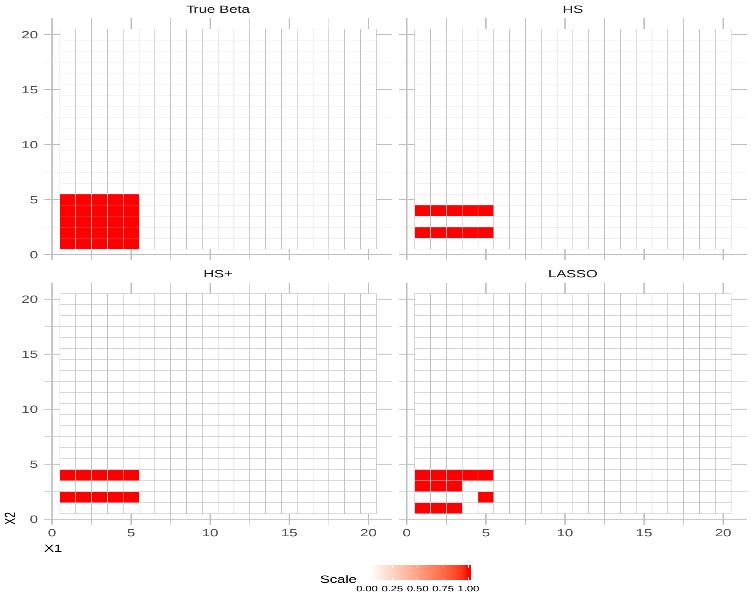
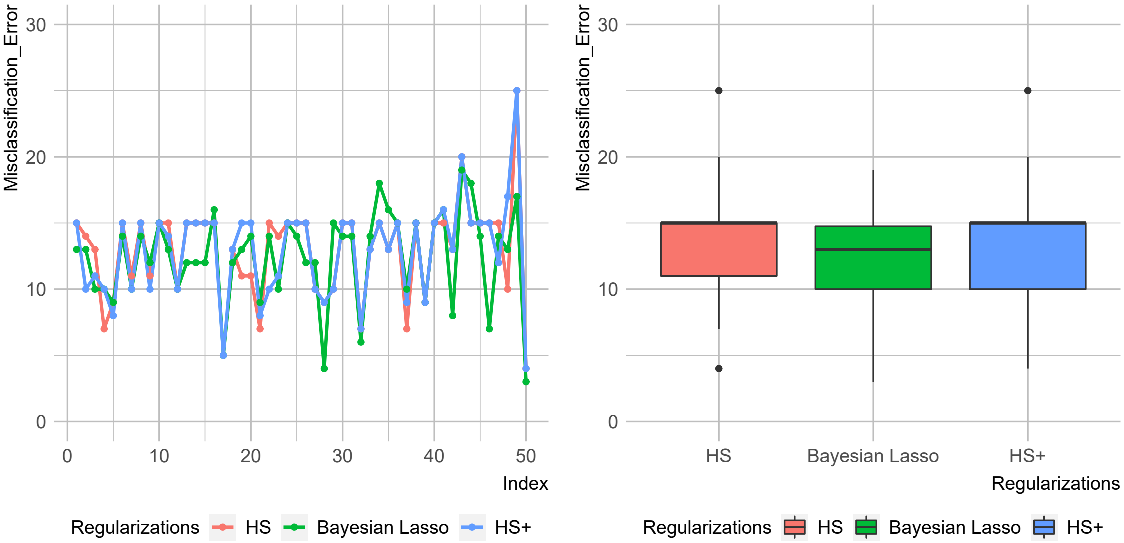


Figure 2.4: Comparison of variable selection performance among the horseshoe, Bayesian Lasso, and horseshoe plus priors when covariates are highly dependent.

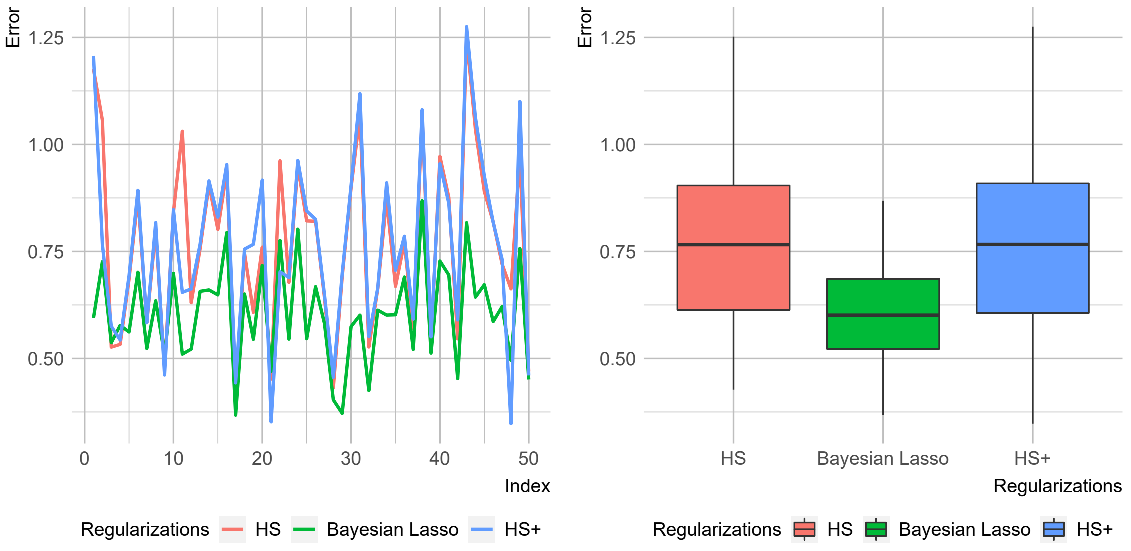
Figure 2.5 shows there is no substantial difference in performance among the candidate priors. The line plots 2.5a do not show a clear dominance of one method over another, as we see in the low dependence case. So the performance of the candidate priors varies in different situations. Though the horseshoe plus candidate prior outperforms the other two priors in the low dependence case, all three priors perform almost equally in the high dependence case.



(a) Misclassification errors for each replication (b) Boxplot of the misclassification errors

Figure 2.5: Comparison of the distributions and per-iteration misclassification errors among horseshoe, Bayesian Lasso, and horseshoe plus under high dependency.

The following figure 2.6 tells that both horseshoe and horseshoe plus have an almost similar distribution of the errors, but their competitor Bayesian Lasso has a lot of overlaps and lower median error. Unlike the performance in the low dependence case, it says Bayesian Lasso outperforms both horseshoe and horseshoe plus candidate priors. The line plots 2.6a do not show any explicit ordering among the three methods, but there are some obvious overlaps. That is a departure from the usual high-dimensional regression example. As shown in Bhadra et al. [2019b], Fan & Li [2001], the performance of convex methods and their Bayesian counterparts, e.g., Lasso, deteriorates with increasing dependence between covariates. However, non-convex regularizers or shrinkage priors like horseshoe are relatively more immune to the issue of dependence [see Mazumder et al., 2010].



(a) Estimation errors for each replication (b) Boxplot of the estimation errors

Figure 2.6: Comparison of the distributions and per-iteration estimation errors among horseshoe, Bayesian Lasso, and horseshoe plus under high dependency.

Table 2.3 is a list of estimation errors for some different number of iterations and shows that there is no precise sequence of estimation errors between horseshoe and horseshoe plus priors, but they are very close. On the other hand, errors for Bayesian Lasso are relatively smaller than the other two.

Table 2.3: Estimation errors under high dependency

	Number of Iteration					
Regularizations	1	10	20	30	40	50
Horseshoe	1.176	0.842	0.760	0.899	0.972	0.470
Bayesian Lasso	**0.595**	**0.697**	**0.717**	**0.575**	**0.728**	**0.451**
Horseshoe Plus	1.207	0.848	0.917	0.905	0.955	0.460

Table 2.4 is a short summary statistics for the figure 2.6 which shows that error mean and median for horseshoe plus are very similar to the horseshoe but, somewhat larger than their counterpart Bayesian Lasso.

Table 2.4: Summary Statistics of the estimation errors

Priors	Mean	Median	Standard Deviation
Horseshoe	0.769	0.766	0.205
Bayesian Lasso	0.605	0.601	0.117
Horseshoe Plus	0.765	0.767	0.213

That is a surprising departure from what we see for the linear regression case, and a more in-depth study, both theoretical and methodological, is needed to understand this phenomenon.

Chapter 3

Future Scopes and Conclusion

In this thesis, we developed a Bayesian approach to the Dirichlet-Multinomial model using horseshoe, Laplace, and horseshoe plus priors to find the association between covariates and taxa in a log-linear regression model and then performed a simulation study to make a comparison among the performance of the three priors in terms of true signals recovering. Based on our simulation, both the horseshoe and horseshoe plus priors outperform the Bayesian Lasso for the compositional data model under low dependency. While a theoretical investigation is beyond the scope of this thesis, we plan to take this up on a future endeavor. Our conjecture is that the heavy tails of global-local shrinkage priors coupled with the spike at zero are responsible for the superior performance compared to Bayesian Lasso. We also plan to apply our developed methodology on a human microbiome data-set, such as the one analyzed in Wadsworth et al. [2017]. The methodological extensions of this project are considering a generalized Dirichlet distribution that offers a more flexible dependence structure and an explicit zero-inflation model for response. We describe these ideas briefly below:

Generalized Dirichlet A possible extension is to use the Zero-Inflated Generalized Dirichlet prior to handle the sparsity in $\boldsymbol{\pi}_i$ resulting from rare species. The Generalized Dirichlet distribution, proposed by Connor & Mosimann [1969] can be described as follows:

$$\begin{aligned}
&Z_j \sim \operatorname{Beta}\left(a_j, b_j\right), j = 1, \ldots, K \\
&\pi_j = Z_j \prod_{i=1}^{j-1} (1 - Z_i) \Rightarrow Z_j = \pi_j / \left(1 - \sum_{i=1}^{j-1} \pi_i\right) \\
&\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \mathrm{GD}(\mathbf{a}, \mathbf{b})
\end{aligned}$$

The Dirichlet distribution can be derived as a special case of Generalized Dirichlet distribution if $b_{j-1} = a_j + b_j$, in particular the symmetric Dirichlet density $\mathrm{Dir}(\alpha/K, \ldots, \alpha/K)$ results if $a_j = \alpha/K, b_j = \alpha(1 - j/K)$. The Generalized Dirichlet distribution has a more general covariance structure compared to the Dirichlet, and it maintains the nice properties of Dirichlet such as conjugacy to multinomial likelihood and complete neutrality [Connor & Mosimann, 1969].

Zero-inflation The Generalized Dirichlet distribution does not account for excess zeroes found in a typical microbiome data. To correct for this, Tang & Chen [2018] proposes a zero-inflated Generalized Dirichlet distribution where they augment the Beta distribution for each Z_j with a point mass at 0. The resulting hierarchy would be:

$$\begin{aligned}
&Z_j \sim \begin{cases} 0 \text{ with probability } \phi_j \\ \mathrm{Beta}\left(a_j, b_j\right), \text{ with probability } 1 - \phi_j \end{cases}, \quad j = 1, \ldots, K, \\
&\pi_j = Z_j \prod_{i=1}^{j-1} (1 - Z_i) \Rightarrow Z_j = \pi_j / \left(1 - \sum_{i=1}^{j-1} \pi_i\right) \\
&\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \mathrm{ZIGD}(\mathbf{a}, \mathbf{b}, \boldsymbol{\phi}).
\end{aligned}$$

Another possible application area is ecological data, where the relative abundance of species depends on a high dimensional covariate, and model selection provides interpretability for such data set. Another future direction is considering time as a covariate that is building a dynamic shrinkage prior to compositional data where the relative abundance of species could vary with time. It appears that such dynamic modeling will require additional modification to the Bayesian strategy beyond the log-linear model covered in this thesis.

Under the Zero-Inflated Generalized Dirichlet prior, our hierarchical model would become:

$$\mathbf{Y}_i \mid \boldsymbol{\pi}_i \sim \text{Multinomial}(y_{i+} \mid \boldsymbol{\pi}_i),\ y_{i+} = \sum_{j=1}^{J} y_{ij}, i = 1, \ldots, N, \tag{3.0.1}$$

$$\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ}) \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\phi} \sim \text{ZIGD}(\boldsymbol{\pi}_i \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\phi}), i = 1, \ldots, N. \tag{3.0.2}$$

Then we incorporate the covariates into the model using a log-linear regression approach by using the log-shape parameter of the Zero-Inflated Generalized Dirichlet distribution as a response variable for the covariates. Our hierarchical model is:

$$\mathbf{Y}_i \mid \boldsymbol{\pi}_i \sim \text{Multinomial}(y_{i+} \mid \boldsymbol{\pi}_i),\ y_{i+} = \sum_{j=1}^{J} y_{ij}, i = 1, \ldots, N, \tag{3.0.3}$$

$$\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ}) \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\phi} \sim \text{ZIGD}(\boldsymbol{\pi}_i \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\phi}), i = 1, \ldots, N, \tag{3.0.4}$$

where $\mathbf{a}$, $\mathbf{b}$, $\boldsymbol{\phi}$ are the two shape parameters and the zero-augmentation probabilities for the Zero-Inflated Generalized Dirichletdistribution. We can model the $\eta_j = \log(a_j)$ terms in a log-linear framework as follows:

$$\eta_{ij} = \log(a_{ij}),\ \eta_{ij} = \beta_{0j} + \sum_{l=1}^{p} X_{il}\beta_{lj}, \quad i = 1, \ldots, N \tag{3.0.5}$$

$$\mu_{ij} = \log(b_{ij}),\ \mu_{ij} = \gamma_{0j} + \sum_{l=1}^{p} X_{il}\gamma_{lj}, \quad i = 1, \ldots, N \tag{3.0.6}$$

$$\beta_{lj} \sim \mathcal{N}(0, \lambda_{lj}^2\tau_j^2),\ \lambda_{lj}^2 \sim \mathcal{C}a^+(0, \tau_j), \tau_j^2 \sim\sim \mathcal{C}a^+(0, 1),\ l = 1, \ldots, p,\ j = 1, \ldots, J \tag{3.0.7}$$

$$\gamma_{lj} \sim \mathcal{N}(0, \tilde{\lambda}_{lj}^2\tilde{\tau}_j^2),\ \tilde{\lambda}_{lj}^2 \sim \mathcal{C}a^+(0, \tilde{\tau}_j), \tilde{\tau}_j^2 \sim\sim \mathcal{C}a^+(0, 1)\ l = 1, \ldots, p,\ j = 1, \ldots, J. \tag{3.0.8}$$

A potential issue with Zero-Inflated Generalized Dirichlet modeling framework is over-parametrization. To see this, first note that the Zero-Inflated Generalized Dirichlet distribution has almost three times as many parameters compared to a typical Dirichlet distribution - how do we handle all of them? Tang & Chen [2018] model the mean and variance of the i, j^{th} Beta distribution ($a_{ij}/(a_{ij} + b_{ij})$ and $1/(1 + a_{ij} + b_{ij})$) as well as the spike mass ϕ_{ij} as three separate log-linear regression framework. A possible future direction is to investigate whether there exists a natural way to reduce the complexity.

Bibliography

Bhadra, A., Datta, J., Li, Y., Polson, N. G. & Willard, B. T. (2019a). Prediction risk for the horseshoe regression. *Journal of Machine Learning Research* **20**, 1–39.

Bhadra, A., Datta, J., Polson, N. G. & Willard, B. (2016). Default bayesian analysis with global-local shrinkage priors. *Biometrika* **103**, 955–969.

Bhadra, A., Datta, J., Polson, N. G., Willard, B. et al. (2017a). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **12**, 1105–1131.

Bhadra, A., Datta, J., Polson, N. G., Willard, B. et al. (2017b). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **12**, 1105–1131.

Bhadra, A., Datta, J., Polson, N. G., Willard, B. et al. (2019b). Lasso meets horseshoe: A survey. *Statistical Science* **34**, 405–427.

Bogdan, M., Chakrabarti, A., Frommlet, F. & Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics* **39**, 1551–1579.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer-Verlag Berlin Heidelberg.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software* **76**.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2009). Handling sparsity via the horseshoe. *Journal of Machine Learning Research W&CP* **5**, 73–80.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

Castillo, I., Schmidt-Hieber, J. & van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43**, 1986–2018.

Chen, J. & Li, H. (2013). Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The annals of applied statistics* **7**.

Connor, R. J. & Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association* **64**, 194–206.

Datta, J. & Dunson, D. B. (2016). Bayesian inference on quasi-sparse count data. *Biometrika* **103**, 971–983.

Datta, J. & Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis* **8**, 111–132.

Datta, J. & Ghosh, J. K. (2015). In search of optimal objective priors for model selection and estimation. *Current Trends in Bayesian Methodology with Applications* , 225.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 1–22.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. 1. Cambridge University Press.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**, 1348–1360.

Ghosh, P., Tang, X., Ghosh, M. & Chakrabarti, A. (2016). Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Anal.* **11**, 753–796.

Holmes, I., Harris, K. & Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one* **7**, e30126.

Johnstone, I. M. & Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* **32**, 1594–1649.

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G. & Shannon, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one* **7**, e52078.

Mazumder, R., Hastie, T. & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research* **11**, 2287–2322.

Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* **83**, 1023–1032.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49**, 65–82.

Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

Polson, N. G. & Scott, J. G. (2010a). Large-scale simultaneous testing with hypergeometric inverted-beta priors. *arXiv preprint arXiv:1010.5223* .

Polson, N. G. & Scott, J. G. (2010b). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics* **9**, 501–538.

Polson, N. G. & Scott, J. G. (2012). Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 287–311.

Tang, Z.-Z. & Chen, G. (2018). Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* .

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B* **58**, 267–288.

Tibshirani, R., Saunders, M. A., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **67**, 91–108.

Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Meth. Dokl.* **4**, 1035–1038.

van der Pas, S., Kleijn, B. & van der Vaart, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8**, 2585–2618.

van der Pas, S., Szabó, B. & van der Vaart, A. (2016). How many needles in the haystack? Adaptive inference and uncertainty quantification for the horseshoe. *arXiv:1607.01892* .

van der Pas, S., Szabó, B. & van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe. *arXiv:1702.03698* .

Virgin, H. W. & Todd, J. A. (2011). Metagenomics and personalized medicine. *Cell* **147**, 44–56.

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A. & Vannucci, M. (2017). An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC bioinformatics* **18**, 94.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research* **7**, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**, 1418–1429.

Appendix
Stan code: Horseshoe prior for Dirichlet-Multinomial model

```
functions {
// for likelihood estimation
  real dirichlet_multinomial_lpmf(int[] y, vector alpha) {
    real alpha_plus = sum(alpha);
    return lgamma(alpha_plus) + lgamma(sum(y)+1) + sum(lgamma(alpha + to_vector(y)))
                - lgamma(alpha_plus+sum(y)) - sum(lgamma(alpha))-sum(lgamma(to_vector(y)+1));
  }
}

data {
  int<lower=1> N; // total number of observations
  int<lower=2> ncolY; // number of categories
  int<lower=2> ncolX; // number of predictor levels
  matrix[N,ncolX] X; // predictor design matrix
  int <lower=0> Y[N,ncolY]; // data // response variable
  //real<lower=0> sd_prior;
  real<lower=0> sd_prior;
  real<lower=0> psi;
}
parameters {
  matrix[ncolX, ncolY] beta_raw; // coefficients (raw)
  vector[N] beta0; // intercept
  matrix<lower=0>[ncolX,ncolY] lambda_tilde; // truncated local shrinkage
  vector<lower=0>[ncolY] tau; // global shrinkage
}
transformed parameters{
  matrix[ncolX,ncolY] beta; // coefficients
  matrix<lower=0>[ncolX,ncolY] lambda; // local shrinkage
  lambda = diag_post_multiply(lambda_tilde, tau);
  beta = beta_raw .* lambda;
}

model {
// prior:
for(k in 1:N){
  beta0[k] ~ normal(0, 10);
}
for (k in 1:ncolX) {
      for (l in 1:ncolY) {
        tau[l] ~ cauchy(0.1, 1); // flexible
        lambda_tilde[k,l] ~ cauchy(0, 1);
        beta_raw[k,l] ~ normal(0,sd_prior);
    }
  }
// likelihood
for (i in 1:N) {
    vector[ncolY] logits;
    for (j in 1:ncolY){
      logits[j] = beta0[i]+X[i,] * beta[,j];
      }
     Y[i,] ~ dirichlet_multinomial(softmax(logits)*(1-psi)/psi);
    }
}
}
```