Dirichlet-Multinomial Regression Models with Bayesian Variable Selection for Microbiome Data

Matthew D. Koslovsky and Marina Vannucci

1 Introduction

Human microbiome research aims to understand how microbiome communities interact with their host, respond to their environment, and influence disease [32]. High-throughput sequencing technologies have enabled researchers to characterize the composition of the microbiome by quantifying richness, diversity, and abundances. See [14] for a detailed review. However, complex environmental interactions with the microbiome challenge our understanding of community function and its impact on health [23]. Knowledge of the relations between microbial composition and other covariates may help researchers design tailored interventions to help maintain a healthy microbiome community [10, 33].

A popular approach for modeling the relation between microbial data and covariates is the Dirichlet-multinomial (DM) regression model, since it appropriately handles the compositional structure of microbiome data and accommodates overdispersion induced by sample heterogeneity and varying proportions among samples [3, 11–13, 28, 34]. To identify potential covariates, penalized likelihood methods have been developed to simultaneously estimate regression coefficients and perform selection [3, 30]. These models typically have relatively short computation times and demonstrate good predictive accuracy. However, it is challenging to incorporate known relations between covariates into these models due to the requirement of

M. D. Koslovsky
Colorado State University, Fort Collins, CO, USA
e-mail: Matt.Koslovsky@colostate.edu

M. Vannucci (✉)
Rice University, Houston, TX, USA
e-mail: marina@rice.edu

S. Datta, S. Guha (eds.), *Statistical Analysis of Microbiome Data*,
Frontiers in Probability and the Statistical Sciences,
https://doi.org/10.1007/978-3-030-73351-3_10

complex optimization routines [30]. Additionally, they do not accommodate model selection uncertainty while carrying out selection.

Alternatively, Bayesian variable selection methods are able to accommodate the complex high-dimensional data structures found in microbiome studies and fully account for model uncertainty over covariate selection. Commonly, spike-and-slab priors for regression coefficients are embedded into hierarchical Bayesian models to perform variable selection [8]. In this model formulation, regression coefficients' priors depend on latent inclusion indicators which determine a covariate's exclusion or inclusion in the model. Bayesian DM regression models with spike-and-slab priors were originally investigated by Wadsworth et al. [28] to identify KEGG orthology pathways associated with microbiome data. Through simulations, they demonstrate improved performance of their method on selecting covariates when compared to alternative methods, including the penalized likelihood approach of [3]. Recently, the work of Wadsworth et al. [28] was extended to accommodate phylogenetic structure between taxa and known and unknown graphical relations between covariates [11]. Additionally, researchers have leveraged data augmentation techniques to efficiently embed DM regression models into joint modeling frameworks, in order to investigate how the microbiome may mediate the relation between dietary factors and phenotypic responses, such as body mass index [12].

In an effort to make advanced Bayesian methods available to researchers studying the microbiome, we demonstrate how to apply the methods contained in `MicroBVS`, a comprehensive `R` package for identifying covariates associated with compositional data [11]. At the core of `MicroBVS` is a suite of Markov chain Monte Carlo (MCMC) algorithms that generate posterior samples of model parameters for inference. The MCMC algorithms are written in C++ to increase performance time and accessed through `R` wrapper functions using `Rcpp` and `RcppArmadillo` [5, 6]. The package includes various Bayesian variable selection methods for compositional data including Dirichlet-multinomial regression, Dirichlet-tree multinomial regression, and the joint modeling approach proposed in [12]. The package has built-in functionality to simulate data in user-specified research scenarios to assess selection performance and conduct sensitivity analyses. Additionally, various auxiliary `R` functions are incorporated to help researchers assess convergence, draw inference from the MCMC samples, and plot results. The package includes a vignette with worked examples using simulated data and access to open-source data used in our analyses.

In Sect. 2, we describe Dirichlet-multinomial (DM) and Dirichlet-tree multinomial (DTM) regression models with spike-and-slab priors and discuss alternative priors for inclusion indicators that accommodate known and unknown graphical structures between covariates. In Sect. 3, we perform a sensitivity and simulation study for Bayesian DM and DTM regression models and compare them to penalization approaches. Section 4 illustrates the application of the `MicroBVS` package to microbiome data collected in the Multi-omics Microbiome Study—Pregnancy Initiative and a benchmark dataset to investigate the relations between gut microbial taxa and dietary covariates. Section 5 provides concluding remarks.

2 Methods

2.1 *Dirichlet-Multinomial Regression Models for Compositional Data*

In this section, we introduce how to model compositional abundance data via a Dirichlet-multinomial (DM) regression framework and then demonstrate how to embed spike-and-slab priors for variable selection, similar to [28]. We first assume that taxa counts $y_i = (y_{i,1}, \ldots, y_{i,K})$ follow a multinomial distribution

$$y_i \sim \text{Multinomial}(\dot{y}_i | p_i), \tag{1}$$

with $\dot{y}_i = \sum_{k=1}^{K} y_{i,k}$, and p_i defined on the K-dimensional simplex

$$S^{K-1} = \left\{ (p_{i,1}, \ldots, p_{i,K}) : p_{i,k} \geq 0, \forall k, \sum_{k=1}^{K} p_{i,k} = 1 \right\}.$$

To account for overdispersion, we specify a conjugate prior on the taxa probabilities,

$$p_i \sim \text{Dirichlet}(\gamma_i), \tag{2}$$

with the K-dimensional vector $\gamma_i = (\gamma_{i,k} > 0, \forall k \in K)$, similar to [13] and [28]. Typically, the p_i are integrated out of the model for computational convenience, and the y_i are modeled with a Dirichlet-multinomial(γ_i) [28]. To incorporate covariate effects into the model, we use a log-linear regression framework for the concentration parameters γ_i. Specifically, we set $\lambda_{i,k} = \log(\gamma_{i,k})$ and assume

$$\lambda_{i,k} = \alpha_k + \boldsymbol{x}_i' \varphi_k, \tag{3}$$

where $\varphi_k = (\varphi_{k1}, \ldots, \varphi_{kP})'$ represents the covariates' potential relation with the kth compositional taxon, and α_k is a taxon-specific intercept term. Additionally, $\boldsymbol{x}_i$ represents a P-dimensional vector of observed covariates for individual i, e.g., age, sex, medication use, and dietary factors. By exponentiating (3), we ensure positive hyperparameters for the Dirichlet distribution.

2.2 *Variable Selection Priors*

For DM regression models, the number of potential models to choose from when performing variable selection, 2^{PK}, grows quickly even for small covariate spaces. To induce sparsity in the model, we embed multivariate spike-and-slab priors for variable selection that identify covariates that are associated with each compositional taxon [20, 24], as opposed to spike-and-slab constructions that select variables as relevant to either all or none of the responses [2]. We assume that the covariates' inclusion in the model is represented by a latent $K \times P$-dimensional inclusion matrix $\boldsymbol{\zeta}$. As such, $\zeta_{kp} = 1$ indicates that covariate p is associated with compositional taxon k and 0 otherwise. The prior for φ_{kp} given ζ_{kp} follows a mixture of a normal distribution and a Dirac-delta function at zero, δ_0, and is commonly referred to as the spike-and-slab prior. Specifically,

$$\varphi_{kp}|\zeta_{kp}, r_k^2 \sim \zeta_{kp} \cdot N(0, r_k^2) + (1 - \zeta_{kp}) \cdot \delta_0(\varphi_{kp}), \tag{4}$$

where r_k^2 is set large to impose a diffuse prior for the regression coefficients included in the model.

The DM model can incorporate different sparsity levels and can accommodate various structural relations between covariates through the specification of the prior probability of inclusion for each covariate, w_{kp}. Commonly, a beta-binomial distribution is assumed. With this prior, we let each ζ_{kp} follow a Bernoulli distribution

$$p(\zeta_{kp}|w_{kp}) = w_{kp}^{\zeta_{kp}}(1 - w_{kp})^{1-\zeta_{kp}}$$

and further assume $w_{kp} \sim \text{Beta}(a, b)$. By integrating out w_{kp}, we obtain

$$p(\zeta_{kp}) = \frac{\text{Beta}(\zeta_{kp} + a, 1 - \zeta_{kp} + b)}{\text{Beta}(a, b)},$$

where the hyperparameters a and b can be set to impose different levels of sparsity in the model. In practice, the authors in [28] suggest using a weakly informative prior probability of inclusion by setting $a + b = 2$, where the prior expected mean value $m = a/(a + b)$. Thus, setting $a = 0.1$ and $b = 1.9$ reflects a prior belief that 5% of the covariates will be selected. A non-informative prior is assumed by setting $a = b = 1$ (i.e., $m = 0.50$). See [28] for a detailed sensitivity analysis regarding hyperparameter specification for DM regression models. To complete the model's specification, we assume that the intercept terms α_k follow a $N(0, \sigma_k^2)$, where σ_k^2 are set large to impose diffuse priors.

2.3 Network Priors

Under the beta-binomial prior, inclusion indicators are assumed independent. In other settings, researchers may be interested in incorporating prior information for the probability of inclusion of a covariate based on known relations with other covariates. For example, when covariates are chosen as gene expression levels, a network of covariate interactions may be known based on biological information [15, 25]. This graphical structure can be incorporated into the model via Markov random field (MRF) priors, which are parameterized to increase a covariate's inclusion probability if neighboring covariates in the graph are included. MRFs are undirected graphical models for random variables whose distribution follows Markovian properties.

To use this information to help guide variable selection, the prior probability of inclusion for each covariate is set according to the given relations between covariates $\boldsymbol{x}$. Specifically, we assume an MRF prior on ζ_k that increases the probability of inclusion for a covariate if covariates in its neighborhood in the graph are also included. Given the graph G, an adjacency matrix that represents the relations between covariates, the prior probability of inclusion for indicators ζ_k follows

$$p(\zeta_k|G) \propto \exp(a_G \mathbf{1}'\zeta_k + b_G \zeta_k' G \zeta_k),$$

where $\mathbf{1}$ is a P-dimensional vector of 1s and a_G and b_G control the global probability of inclusion and the influence of neighbors' inclusion on a covariate's inclusion, respectively. Previous studies have demonstrated how small increments in b_G can drastically increase the number of covariates included in the model [15, 25]. Li and Zhang [15] provide a detailed description of how to select a value for b_G. Note that if there is no structure between covariates, the prior probabilities of inclusion simplify to independent Bernoulli(exp$(a_G)/(1 + \exp(a_G))$).

2.3.1 Unknown G

When less is known about the relations between covariates, the network structure, G, can be inferred. Efficient sampling algorithms for learning the structure of high-dimensional data with Gaussian graphical models [29] have allowed researchers to embed them into Bayesian variable selection models that simultaneously perform variable selection while learning the relations between covariates [19].

Let $X \sim MVN(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \Sigma^{-1}$ is a $P \times P$ precision matrix. Following [29], we assume a hierarchical prior that models conditional dependence between covariates through edge detection in an undirected graph. Let graph G contain P nodes, corresponding to the set of potential covariates in the model. Let $g_{st} \in \{0, 1\}$ represent a latent inclusion indicator for an edge between nodes s and t, for $s < t$. The inclusion of edge g_{st} corresponds to $\omega_{st} \neq 0$, where ω_{st}, $1 \leq s < t \leq P$, are the off-diagonal elements of $\boldsymbol{\Omega}$. The prior distribution for $\boldsymbol{\Omega}$ is the product of P exponential distributions for diagonal components and $P(P - 1)/2$ mixtures of normals for off-diagonal components of the precision matrix. Specifically,

$$p(\boldsymbol{\Omega}|G, v_0, v_1, \theta) = \{C(G, v_0, v_1, \theta)\}^{-1} \prod_{s<t} N(\omega_{st}|0, v_{st}^2) \prod_{s} \text{Exp}\,(\omega_{ss}|\theta/2)\, I_{\{\boldsymbol{\Omega} \in M^+\}},$$

where $\text{Exp}(\cdot|\theta/2)$ represents an exponential distribution with mean $2/\theta$, $C(G, v_0, v_1, \theta)$ is a normalizing constant, and $I_{\{\boldsymbol{\Omega} \in M^+\}}$ is an indicator function that constrains $\boldsymbol{\Omega}$ to be a symmetric positive definite matrix. Here, $v_{st}^2 = v_1$ if the edge inclusion indicator $g_{st} = 1$, and $v_{st}^2 = v_0$ if $g_{st} = 0$. In practice, $v_0 > 0$ is set small to concentrate ω_{st} around zero for excluded edges, and $v_1 > 0$ is set large so that ω_{st} is freely estimated via a diffuse prior for included edges. The prior for the edge inclusion indicator g_{st} follows

$$p(G, v_0, v_1, \theta, \pi) = \{C(v_0, v_1, \theta, \pi)\}^{-1} C(G, v_0, v_1, \theta) \prod_{s<t} \left\{ \pi^{g_{st}} (1-\pi)^{1-g_{st}} \right\},$$

where $C(v_0, v_1, \theta, \pi)$ is a normalizing constant and π represents the prior probability of inclusion for an edge. Following the recommendations of [29], the specification of π should reflect prior belief in the sparsity of the graph, and θ is typically set to one. The latter implies a relatively vague prior for ω_{ss}, since the data are usually standardized prior to analysis. See [29] for more details regarding prior specification.

2.4 Dirichlet-Tree Multinomial Models

In this section, we describe Bayesian variable selection for Dirichlet-tree multinomial regression models, similar to [11]. The DM model described in Sect. 2.1 assumes that counts are negatively correlated. Alternatively, the Dirichlet-tree multinomial model (DTM) inherits the DM's ability to handle overdispersed data, can model general correlation structures between counts, and can naturally incorporate structural information [4, 17]. In microbiome research, this allows us to model evolutionary relations among taxa represented by a phylogenetic tree [11, 26, 27, 30].

To accommodate a tree-like structure among counts, the multinomial distribution is deconstructed into the product of multinomial distributions for each of the sub-trees in the tree, and the conjugate Dirichlet-tree prior is assumed [4]. Specifically, let tree T have K leaf nodes and V internal nodes. Let C_v represent the set of child nodes for each individual node $v \in V$. For each subject, the branch probability between parent node v and child node c is represented as $p_{i,vc}$, where $\sum_{c=1}^{|C_v|} p_{i,vc} = 1$ and $|C_v|$ is the number of child nodes of v. Under this parameterization, we assume that $y_{i,v} = (y_{i,v1}, \ldots, y_{i,vC})'$ follows a Multinomial($\dot{y}_{i,v}, p_{i,v}$), where $p_{i,v} = \{p_{i,vc}, c \in C_v\}$. We assume a Dirichlet($\gamma_{i,v}$) prior for each $p_{i,v}$, where $\gamma_{i,v} = (\gamma_{i,vc} > 0, \forall c \in C_v)$. Integrating the $p_{i,v}$ out, we model $\dot{y}_{i,v}$ with a Dirichlet-multinomial($\gamma_{i,v}$) and take the product of the v Dirichlet-multinomial models for each sub-tree, to obtain the Dirichlet-tree multinomial (DTM) distribution as

$$p(y_i|\gamma_i, v \in V) = \prod_{v \in V} \frac{\Gamma(\sum_{c \in C_v} y_{i,vc} + 1)\Gamma(\sum_{c \in C_v} \gamma_{i,vc})}{\Gamma(\sum_{c \in C_v} y_{i,vc} + \sum_{c \in C_v} \gamma_{i,vc})} \times \prod_{c \in C_v} \frac{\Gamma(y_{i,vc} + \gamma_{i,vc})}{\Gamma(y_{i,vc} + 1)\Gamma(\gamma_{i,vc})},$$

where Γ represents the gamma function. The generalized DM model and the DM model are special cases of the DTM class of models [30]. Specifically, the generalized DM model can be represented as a DTM with a binary cascading tree (i.e., at each level of the tree, the rightmost branch splits into two), and the DM can be represented with a tree containing only one root node and K leaf nodes.

Similar to Eq. (3), covariate effects can be incorporated into the model using a log-linear regression framework. Specifically, we set $\lambda_{i,vc} = \log(\gamma_{i,vc})$ and assume

$$\lambda_{i,vc} = \alpha_{vc} + \boldsymbol{x}_i' \varphi_{vcp},$$

where $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,P})'$ represents a set of measurements on P covariates and $\boldsymbol{\varphi}_{vc} = (\varphi_{vc1}, \ldots, \varphi_{vcP})'$. We assume that the intercept terms α_{vc} follow a $N(0, \sigma^2_{vc})$, where σ^2_{vc} are set large to impose vague priors on α_{vc}. Similar prior specifications for variable selection presented in Sect. 2.2 can be applied to each of the DM components of this model.

2.5 Posterior Inference

In Bayesian inference, the posterior distribution is proportional to the product of the likelihood of the data and the prior distributions for the parameters. For both DTM and DM models, researchers have implemented Metropolis–Hastings algorithms within a Gibbs sampler for inference [11, 28]. Since the DTM model is a generalization of the DM model, we present a general MCMC algorithm in the context of DTM models. Assuming a beta-binomial prior probability of inclusion, the parameter space is described as $\Phi = \{\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\zeta}\}$, and the posterior distribution is

$$p(\Phi|\boldsymbol{Y}, \boldsymbol{x}) \propto f(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x}) p(\boldsymbol{\alpha}) p(\boldsymbol{\varphi}|\boldsymbol{\zeta}) p(\boldsymbol{\zeta}).$$

We use a two-step update approach to sample regression coefficients and inclusion indicators for covariates, following [21].

A generic iteration of the MCMC algorithm is described as follows:

- Update each α_{vc}—Metropolis step with random walk proposal from $\alpha'_{vc} \sim N(\alpha_{vc}, 0.50)$. Accept proposal with probability

$$\min\left\{ \frac{f(\boldsymbol{Y}|\boldsymbol{\alpha}', \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x}) p(\alpha'_{vc})}{f(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x}) p(\alpha_{vc})}, 1 \right\}.$$

- Jointly update a ζ_{vcp} and φ_{vcp}
  - *Between-Model Step*: Randomly select a ζ_{vcp} term.

    Add: If the covariate is currently excluded ($\zeta_{vcp} = 0$), change it to $\zeta'_{vcp} = 1$. Then, sample a $\varphi'_{vcp} \sim N(\varphi_{vcp}, 0.50)$. Accept proposal with probability

$$\min\left\{\frac{f(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\varphi}',\boldsymbol{\zeta}',\boldsymbol{x})p(\varphi'_{vcp}|\zeta'_{vcp})p(\zeta'_{vc})}{f(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\varphi},\boldsymbol{\zeta},\boldsymbol{x})p(\zeta_{vc})},1\right\}.$$

Delete: If the covariate is currently included ($\zeta_{vcp}=1$), change it to $\zeta'_{vcp}=0$ and set $\varphi'_{vcp}=0$. Accept proposal with probability

$$\min\left\{\frac{f(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\varphi}',\boldsymbol{\zeta}',\boldsymbol{x})p(\zeta'_{vc})}{f(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\varphi},\boldsymbol{\zeta},\boldsymbol{x})p(\varphi_{vcp}|\zeta_{vcp})p(\zeta_{vc})},1\right\}.$$

- *Within-Model Step*: Propose a $\varphi'_{jp}\sim N(\varphi_{jp},0.50)$ for each covariate currently selected in the model ($\zeta_{vcp}=1$). Accept each proposal with probability

$$\min\left\{\frac{p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\varphi}',\boldsymbol{\zeta},\boldsymbol{x})p(\varphi'_{vcp}|\zeta_{vcp})}{p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\varphi},\boldsymbol{\zeta},\boldsymbol{x})p(\varphi_{vcp}|\zeta_{vcp})},1\right\}.$$

To include a known graphical structure and impose an MRF prior for selection, the algorithm simply replaces $p(\zeta)$ with $p(\zeta|G)$. If the relational structure between the covariates is unknown, the posterior distribution of the model is redefined as

$$p(\Phi|\boldsymbol{Y},\boldsymbol{X})\propto f(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\varphi},\boldsymbol{\zeta},\boldsymbol{X})f(\boldsymbol{X}|\boldsymbol{\Omega})p(\boldsymbol{\alpha})p(\boldsymbol{\Omega}|G)p(\boldsymbol{\varphi}|\boldsymbol{\zeta})p(\boldsymbol{\zeta}|G)p(G),$$

where $\Phi=\{\boldsymbol{\alpha},\boldsymbol{\varphi},\boldsymbol{\zeta},\boldsymbol{\Omega},G\}$. Note that this parameterization treats the covariates $\boldsymbol{X}$ as random and not fixed. For implementation, the MCMC algorithm requires two additional steps to simultaneously learn the graphical relations. We update $\boldsymbol{\Omega}$ and G following the approach outlined in [29].

For implementation, the algorithms are initiated at a set of arbitrary parameter values and then used to generate samples of the posterior distribution. After burn-in, the remaining samples are used for inference. To determine inclusion in the model, the marginal posterior probability of inclusion (MPPI) for each of the covariates is determined by taking the average of their respective inclusion indicator's MCMC samples. Note that a covariate has a unique inclusion indicator for each of the taxon. Commonly, variables are included in the model if their MPPI ≥ 0.50 [1]. Alternatively, the authors in [18] propose using a threshold based on a Bayesian false discovery rate (BFDR) to control for multiplicity.

3 Simulated Data

In this section, we demonstrate the selection performance for the DM and DTM models using simulated data. For the DM models, we compared the performances using different variable selection priors, i.e., a beta-binomial prior, an MRF prior with fixed graphical structure (i.e., G set to the truth and G learned a priori), and an MRF prior with unknown graphical structure.

For variable selection, all models were assessed on the basis of sensitivity (1—false negative rate), specificity (1—false positive rate), and Matthew's correlation coefficient (MCC) (a measure of overall selection accuracy). These are defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

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

where TN, TP, FN, and FP represent the true negatives, true positives, false negatives, and false positives, respectively. Covariates were determined to be associated with the compositional and response data, respectively, if their MPPI ≥ 0.50 [1]. Results we report below were obtained by averaging over 30 replicated datasets.

3.1 Simulation Study for DM Regression Models

Similar to simulation schemes adopted by [3, 12, 28], we simulated $N = 100$ subjects with $P = 30$ covariates and $K = 75$ compositional taxa. Covariates $\boldsymbol{x}$ were simulated from a $N_P(\mathbf{0}, \Sigma)$, where Σ was set to a block diagonal matrix with one along the diagonal and three 5×5 exchangeable covariance structures (for the first 15 covariates) with σ_{ij} = 0.7, 0.5, and 0.3, respectively. In each of the replicate datasets, we randomly selected 25 of the 2250 covariate–taxon combinations to be associated with the compositional data. Corresponding regression coefficients φ were randomly sampled from $\pm[0.75, 1.25]$. Intercept terms $\boldsymbol{\alpha}$ were simulated from a Uniform$[-2.3, 2.3]$. The compositional data $\mathbf{Y}$ were sampled from a Multinomial$(\dot{y}_i, p_i^*)$, where $\dot{y}_i \sim$ Uniform[5,000, 10,000] and $p_i^* \sim$ Dirichlet$(\boldsymbol{\gamma}_i^*)$, where $\boldsymbol{\gamma}_i^* = (\gamma_{i,1}^*, \gamma_{i,2}^*, \ldots, \gamma_{i,K}^*)$. We let $\gamma_{i,k}^* = \frac{\gamma_{i,k}}{\sum_{k=1}^K \gamma_{i,k}} \frac{1-d}{d}$, $k = 1, \ldots, K$, where $\gamma_{i,k}$ was determined using Eq. (3), and d serves as an overdispersion parameter which was set at 0.01. As a result, the data-generating model differs from our model assumptions.

When running the MCMC algorithm, we set hyperparameters $a = 1$ and $b = 9$ for the beta-binomial prior and $a_G = \log(0.1/0.9)$ for the MRF prior, representing a prior expectation of 10% of the total number of covariates included in both models. For the MRF prior with known graphical structure, we set $b_G = 0.2$ and the graph G equal to a $P \times P$-dimensional block diagonal matrix, with 3, 5×5 blocks of 1s for the first 15 elements. Additionally, we set G equal to the graphical structure learned

Table 1 Simulation results for the DM regression model with various inclusion indicator prior assumptions. **# Selected**—the number of selected covariates and **MCC**—Matthew's correlation coefficient. Results are presented as mean (SD) over 30 replicate datasets

Prior	# Selected	Sensitivity	Specificity	MCC
beta-binomial	24.3 (3.2)	0.904 (0.092)	0.999 (0.001)	0.917 (0.062)
MRF fixed G-*true*	56.6 (13.8)	0.987 (0.028)	0.986 (0.006)	0.665 (0.083)
MRF fixed G-*learned*	43.8 (12.4)	0.975 (0.032)	0.991 (0.006)	0.748 (0.086)
MRF unknown G	42.6 (8.5)	0.979 (0.029)	0.992 (0.003)	0.766 (0.073)

using [29]. For the MRF prior with unknown graphical structure, we set $b_G = 0.2$, $v_0 = 0.01$, $v_1 = 10$, $\lambda = 1$, and $\pi = 2/(P-1)$, similar to [29]. Simulations were run for 10,000 iterations and thinned to every 10th iteration. This resulted in 1,000 iterations, of which the first 500 iterations were treated as burn-in and the remaining 500 used for inference. Each run was initiated with $\zeta_{pk} = 0$ and α_k sampled from a standard normal distribution.

Results are found in Table 1. Overall, the DM model with MRF prior and fixed graphical structure among covariates had the highest number of selected covariates on average. These results were expected since the baseline prior probability of inclusion using the MRF (a_G) was set to impose a 10% prior probability of inclusion, similar to the beta-binomial model, and any graphical structure (known or unknown) would only increase the probability of inclusion in the model. As a result, the MRF with G fixed to the truth had the highest sensitivity overall. However, since it typically overselected, it achieved the lowest specificity and MCC as well. Overall, the DM with a beta-binomial prior had the highest MCC (∼92%). Lastly, we observed a marginal improvement in selection performance when learning the graphical structure simultaneously in the model versus a priori. It is important to note that the MRF model with unknown graphical structure had similar performance to the MRF with known graphical structure while additionally providing inference on the relations among covariates.

3.2 DM Sensitivity Analysis

To assess the model's sensitivity to hyperparameter settings, we set each of the hyperparameters to default values and then evaluated the effect of manipulating each term on selection performance. We investigated the model's sensitivity to specification of the beta-binomial prior hyperparameters a and b, MRF prior hyperparameters a_G and b_G, and hyperparameters associated with the Gaussian graphical models, v_0, v_1, and π. For the default parameterization, we set the hyperparameters for the beta-binomial prior inclusion indicators to $a = 1$ and $b = 9$. For the MRF priors, we set the hyperparameters $a_G = \log(0.1/0.9)$ and $b_G = 0.2$. The default values for the Gaussian graphical model hyperparameters were $v_0 = 0.01$, $v_1 = 10$, and $\pi = 2/(P-1)$. We ran our MCMC algorithm on

Table 2 Sensitivity results for the beta-binomial and MRF prior probability of inclusion parameters b and b_G, respectively, the exclusion variance for graphical edge selection v_0, and the prior probability of edge inclusion π. **# Selected**—the number of selected covariates **MCC**—Matthew's correlation coefficient. Results are presented as mean (SD) over 30 replicate datasets

Prior		$b = 1$	$b = 99$
beta-binomial	# Selected	37.0 (9.4)	21.5 (2.5)
	Sensitivity	0.97 (0.04)	0.83 (0.12)
	Specificity	0.99 (0.00)	1.00 (0.00)
	MCC	0.81 (0.09)	0.89 (0.08)
		$b_G = 0.05$	$b_G = 0.5$
MRF fixed G	# Selected	41.2 (10.9)	893.6 (72.4)
	Sensitivity	0.97 (0.04)	1.00 (0.00)
	Specificity	0.99 (0.00)	0.61 (0.03)
	MCC	0.77 (0.09)	0.13 (0.01)
		$v_0 = 0.001$	$v_0 = 0.1$
MRF unknown G	# Selected	43.9 (10.7)	41.9 (9.8)
	Sensitivity	0.98 (0.03)	0.97 (0.03)
	Specificity	0.99 (0.00)	0.99 (0.00)
	MCC	0.75 (0.09)	0.76 (0.08)
		$\pi = 0.02$	$\pi = 0.5$
MRF unknown G	# Selected	42.5 (11.6)	47.1 (12.3)
	Sensitivity	0.98 (0.04)	0.98 (0.03)
	Specificity	0.99 (0.01)	0.99 (0.01)
	MCC	0.76 (0.09)	0.72 (0.08)

the 30 replicated datasets generated in the simulation study, using 10,000 iterations, treating the first 5,000 iterations as burn-in, and thinning to every 10th iteration.

The results of the sensitivity analysis are presented in Table 2. As expected, we found that increasing (decreasing) b in the beta-binomial prior reduced (increased) the number of covariates selected in the model. Here, we observed a positive relation between sensitivity and the prior probability of inclusion. However, since the model overselected covariates with smaller b values, the specificity diminished as a result. Using an MRF prior with a fixed underlying graphical structure, we found that as b_G increased, so did the number of selected covariates on average. In our analysis, the models seemed to experience a phase transition, in which the number of covariates selected in the model dramatically increased, for $b_G = 0.5$. See [15] for recommendations on selecting the appropriate b_G in practice. With unknown graphical structure, we found marginal differences in results relative to changes in v_0 and π.

3.3 Simulation Study for DTM Regression Models

For the DTM model, we compared selection performances to the penalized DTM approach of [30]. We simulated $N = 100$ subjects with $P = 75$ covariates and $K = 30$ compositional taxa. Covariates $\boldsymbol{x}$ were simulated from a $N_P(\mathbf{0}, \Sigma)$, where

$\sigma_{ij} = \omega^{|i-j|}$ and $\omega = 0.3$. In each of the replicate datasets, we randomly selected 15 of the 4,350 covariate–branch combinations to be associated with the compositional data. Corresponding regression coefficients φ were randomly sampled from $\pm[0.75, 1.50]$. Intercept terms $\boldsymbol{\alpha}$ were simulated from a Uniform$[-1.3, 1.3]$. The multivariate count data $\mathbf{Y}$ were sampled from a DTM regression model with total counts for each individual uniformly distributed between 7,500 and 10,000. For each dataset, we simulated a random tree using sequential binary separation [7], in which the parent node and subsequent internal nodes are split into two branches until the total number of leaf nodes K is obtained.

We chose a beta-binomial inclusion prior and set $a = 1$ while varying b as $b = 1, 9$, and 99, to investigate the model's sensitivity to hyperparameter specification. The MCMC algorithms were run for 40,000 iterations, treating the first 20,000 as burn-in and thinning to every 10th iteration. For the penalized approach of [30], it is necessary to choose tuning parameters γ and λ, which control the sparsity of the model. When $\gamma = 0$ and $\gamma = 1$, the model generates the lasso and group lasso estimate, respectively. Following the recommendations of [30], we set $\gamma = \{0.0, 0.25, 0.5, 1.0\}$ and fit the model over a grid of λ values. The best model for each γ was then chosen by minimizing the Bayesian information criterion [22].

Similar to the DM model, we found that the DTM was sensitive to the prior probability of inclusion (Table 3). Specifically, as b increased (decreased), the number of covariate–branch association decreased (increased), as expected. We found that the model with $b = 9$ had the best selection performance overall (MCC $= 0.544$), and the non-informative model (i.e., $a = b = 1$) showed the worst performance overall (MCC $= 0.219$). All prior specifications achieved a relatively high specificity (>0.97). Similar specificity results were found with the penalized approach (Table 4). However, the penalized approach, regardless of tuning parameter γ, had extremely low sensitivity, resulting in low MCC values as well. When $\gamma = 1$, the penalized model did not select any covariate–branch terms (results not shown).

Table 3 Simulation results for the Bayesian variable selection method for DTM regression models at various prior probabilities of inclusion

Prior	# Selected	Sensitivity	Specificity	MCC
$a = 1$ and $b = 1$	135.0 (42.1)	0.642 (0.184)	0.971 (0.010)	0.219 (0.085)
$a = 1$ and $b = 9$	15.5 (5.2)	0.564 (0.239)	0.998 (0.001)	0.544 (0.202)
$a = 1$ and $b = 99$	5.2 (2.6)	0.293 (0.145)	1.00 (0.00)	0.491 (0.156)

Table 4 Simulation results for the penalized DTM regression approach of [30]. For each γ, the optimal model is chosen over a grid of λ values using the Bayesian information criterion

γ	# Selected	Sensitivity	Specificity	MCC
0.0	47.5 (36.3)	0.122 (0.172)	0.989 (0.008)	0.071 (0.098)
0.25	28.3 (25.1)	0.107 (0.173)	0.994 (0.006)	0.090 (0.142)
0.50	17.3 (21.7)	0.071 (0.139)	0.996 (0.005)	0.076 (0.118)

Table 5 Sensitivity results for high and low count associations with the Bayesian beta-binomial ($a = 1$ and $b = 9$) and Penalized DTM regression models. **# Selected**— the number of selected covariates and **MCC**—Matthew's correlation coefficient. Results are presented as mean (SD) over 30 replicate datasets

Branch count	Model	# Selected	Sensitivity	Specificity	MCC
High	Bayesian	19.2 (3.5)	0.507 (0.106)	0.998 (0.001)	0.575 (0.097)
	Penalized	49.6 (35.8)	0.800 (0.089)	0.993 (0.008)	0.629 (0.135)
Low	Bayesian	17.7 (5.4)	0.466 (0.165)	0.999 (0.001)	0.546 (0.131)
	Penalized	255.1 (320.1)	0.542 (0.220)	0.944 (0.074)	0.270 (0.189)

3.4 DTM Sensitivity Analysis

In this sensitivity analysis, we investigate how selection performance is affected by branch count. Specifically, we simulated data similar to Sect. 3.3, with the exception that we targeted high (upper quartile) and low (lower 50th percentile) branch count regions in the tree when setting the associated terms. In the first (second) setting, we activated 25 terms across 5 high (low) count branches. We applied the Bayesian and penalized approaches used in the simulation study in this analysis and present results for the best performing parameterizations. For the Bayesian approach, we assumed a beta-binomial prior for inclusion indicators, ($a = 1$ and $b = 9$), and for the penalized approach, we set $\gamma = 0.50$.

The results of our sensitivity analysis are presented in Table 5. Here, we found that the Bayesian model was quite robust to branch counts. In both the high and the low settings, it generated selection performance results similar to the simulation study (MCC $\sim$ 0.55). The penalized method showed the best performance overall when the covariates were associated with high branch counts (MCC $=$ 0.63). However, in the low branch count setting, it over-selected, which greatly reduced its overall performance. Thus, in practice, the Bayesian method may be preferred in more sparse settings, whereas the penalized approach may be better suited for studies with higher numbers of taxa reads.

4 Applications

In this section, we apply the DM and DTM Bayesian variable selection methods to data collected in two microbiome studies, in order to demonstrate how to implement the MCMC algorithms provided in `MicroBVS` and how to draw inference on the results. First, we apply the DM regression model with beta-binomial and MRF priors for inclusion indicators to open-source data collected in the Multi-omics Microbiome Study—Pregnancy Initiative (MOMS-PI) [9]. This study was funded by the NIH Roadmap Human Microbiome Project with the aim of understanding the relations between the microbiome and pregnancy-related health outcomes. Then, we demonstrate the functionality of the DTM regression model by applying it to a benchmark dataset collected to study the relation between the dietary intake and the human gut microbiome [31]. The data used in this analysis consist of 28 genera-level OTU counts obtained from 16S rRNA sequencing and a corresponding set of 97 dietary intake covariates derived from diet information collected using a food frequency questionnaire on 98 subjects.

4.1 Multi-omics Microbiome Study—Pregnancy Initiative (MOMS-PI)

To demonstrate the application of the DM regression models with various inclusion indicator priors, we use the open-source data collected in the Multi-omics Microbiome Study—Pregnancy Initiative (MOMS-PI). Data were obtained from the `HMP2Data` package in `R`, which contains observations on 596 subjects. Women enrolled in the study provided microbiome samples from the mouth, skin, vagina, and rectum longitudinally. We investigated relations between the vaginal microbiome and cytokine abundances, which help regulate the composition of the vaginal microbiome. For this analysis, we used baseline measures on 225 subjects with accompanying cytokine abundances. The dataset is available as part of the `MicroBVS` `R` package [11]. To install the package, follow the instructions in the README found at http://github.com/mkoslovsky/MicroBVS. Once installed, load the package, as well as the abundance, cytokine, and taxonomic data, into the `R` environment by running:

```
library(MicroBVS)
data("momspi16S")
data("momspiCyto")
data("momspi16S_tax")
```

We further limited analyses to only those taxa identified in at least 10% of participants (i.e., 123 taxa), to reduce the number of spurious relationships detected. We also standardized the cytokine values before analysis. When running the model with an MRF prior with an unknown graphical structure, cytokine abundances were log transformed and centered. Prior to transformation, cytokine values ≤ 0 were replaced with relatively small pseudovalues.

To fit the DM regression model with a non-informative beta-binomial prior for inclusion indicators, simply run

```
model1 <- DMbvs_R(iterations = 50000, thin = 10,
                  z = momspi16S, x = momspiCyto,
                  prior = "BB", a = 1, b = 1, seed = 1)
```

For the results given below, we ran the model for 50,000 iterations, thinning to every 10th and setting the initial seed at 1 for reproducibility. To extract the results from the `DMbvs_R` object, use the `selected_DM()` function as follows:

```
out <- selected_DM( model1, threshold = 0.5, burnin =
    2500)
```

The out object contains a # Selected covariates × 2-dimensional matrix of associations, where the first (second) column represents the row (column) of the corresponding *momspiCyto* term selected using a burn-in of 2500 iterations and a marginal posterior probability of inclusion threshold of $\geq$0.5, following the median model approach [1]. Additionally, the out object contains the MPPIs for all of the corresponding cytokine–taxon associations. Figure 1 presents a plot of the MPPIs for each covariate–taxon pair, and Fig. 2 is a heatmap of identified associations' regression coefficients. For this analysis, the model selected 43 covariate–taxon associations.

Next, we ran the DM regression model with an MRF prior with an unknown graphical structure as

```
model2 <- DMbvs_R(iterations = 50000, thin = 10,
                  z = momspi16S, x = momspiCyto,
                  prior = "MRF_unknown",
                  a_G = 0, b_G = 0.2, v0 = 0.01, v1 = 10,
                  pie = 2/(ncol(momspiCyto)-1), lambda = 1)
```

We assumed the baseline prior probability of inclusion, a_G, equal to zero (analogous to the non-informative beta-binomial prior), and the rest of the hyperparameters were set similarly to our simulation study. Results from model2 can be extracted using the selected_DM function as above. To extract the learned graphical

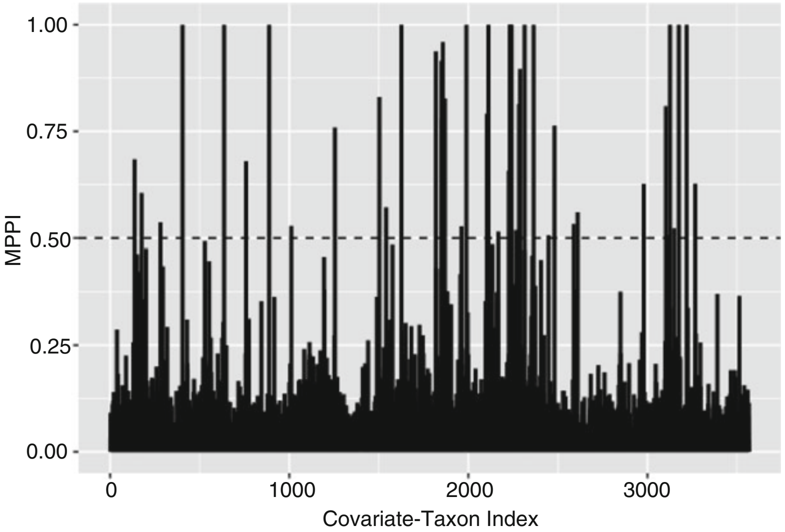

Fig. 1 MOMS-PI study: resulting marginal posterior probability of inclusion from DM regression model with beta-binomial priors for inclusion indicators. MPPI threshold of 0.50 indicated with dotted line

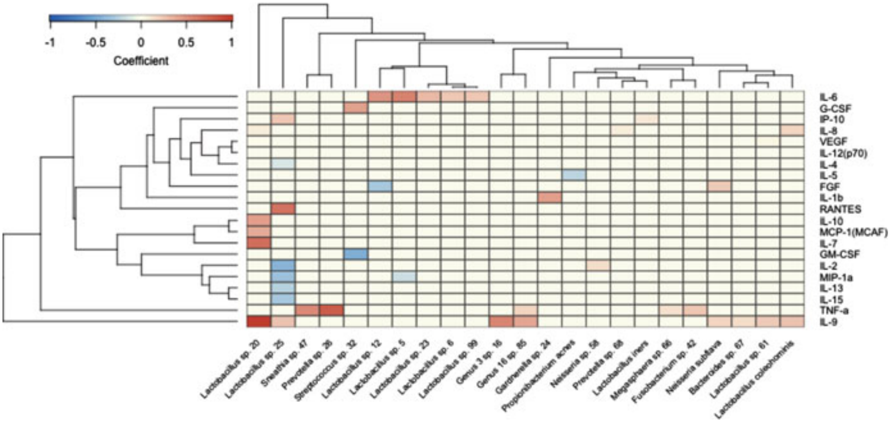

Fig. 2 MOMS-PI study: heatmap of cytokine-taxon associations identified with DM regression model with beta-binomial priors. Taxa are indexed by genus and species

structure in the cytokine data, additionally set the argument `G = T`. This generates an additional `estimated_G` element for the `selected_DM` object, which is a # cytokines × # cytokines-dimensional adjacency matrix. A network plot of the learned structure is presented in Fig. 3. With the MRF prior, the number of included covariate–taxon associations increased to 64, as expected from the simulation study. A plot of the MMPIs for `model2` is presented in Fig. 4, and the corresponding heatmap of identified association is presented in Fig. 5. To fit the DM regression model with fixed graphical structure between covariates, set the `DMbvs_R` function argument `prior = "MRF_fixed"` and `G` equal to an adjacency matrix representing the assumed graphical structure. Additional examples on simulated data can be found in the vignette provided with the `MicroBVS` package.

4.2 *Gut Microbiome Study*

In this section, we demonstrate how to apply the DTM Bayesian variable selection method to a benchmark dataset collected to study the relation between the dietary intake and the human gut microbiome [31]. Previously, Wang and Zhao [30] proposed a penalized DTM regression model to identify dietary intake covariates associated with genus-level operational taxonomic units (OTUs) on a subset of these data. We illustrate the Bayesian DTM model on the same subset. To load the necessary R packages and data into the R environment, run

```
library(MicroBVS)
library(phyloseq)
data("Gut_micro")
```

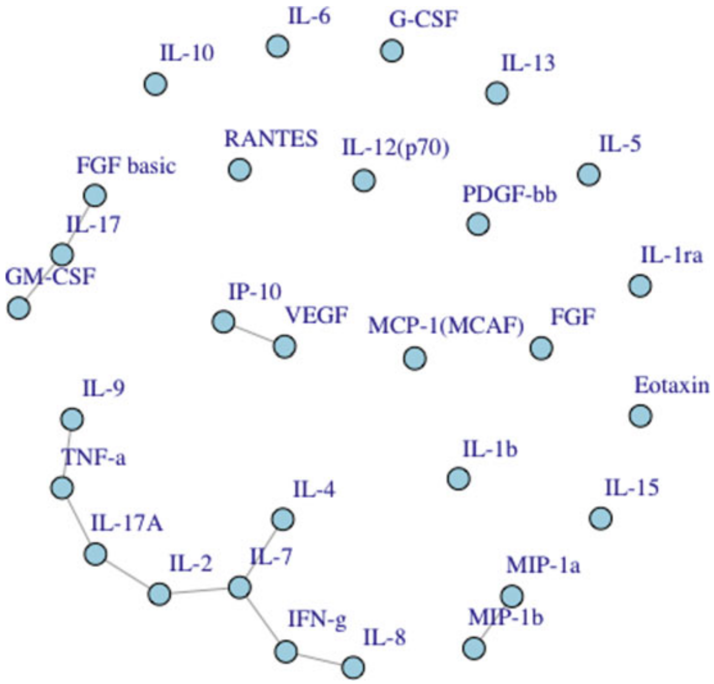

Fig. 3 MOMS-PI study: learned graphical structure of cytokine data

```
data("Gut_dietary")
data("tree")
```

The phylogenetic tree used in this example is presented in Fig. 6. We assumed a non-informative beta-binomial prior for inclusion indicators ($a = b = 1$). The MCMC algorithm was run for 150,000 iterations thinning to every 100th sample. After a burn-in of 750 samples, inference was drawn from the remaining 750.

```
model_gut <- DTMbvs_R( iterations = 150000, thin =
    100, tree = tree, Y = Gut_micro, X = Gut_dietary,
     prior = "BB", seed = 1)
```

In this example, we used a Bayesian false discovery rate of 0.01 to determine a covariate's inclusion in the model. To identify the corresponding MPPI threshold for inclusion, run the selected_DTM function to obtain the matrix of MPPIs. Then, run the bfdr function at the prespecified error level, i.e., 0.01 in this application. Next, run the selected_DTM function with the BFDR threshold, MPPI $\geq$ 0.89 in this example. To label the covariates, we supplied the column names for the Gut_dietary matrix. While not shown here, the function also has an argument for edge labels (edge_lab) to help with inference. See the vignette for more details.

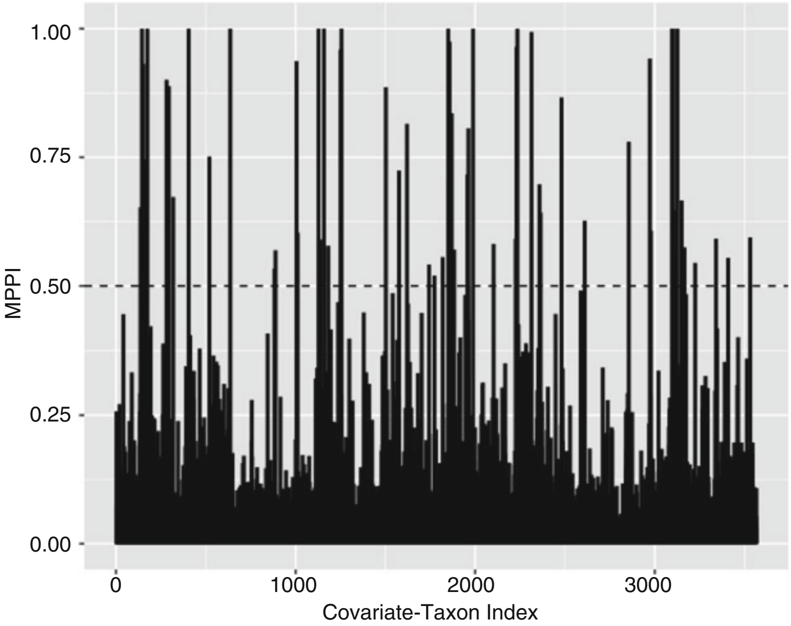

Fig. 4 MOMS-PI study: resulting marginal posterior probability of inclusion for results from DM regression model with MRF prior for inclusion indicators. MPPI threshold of 0.50 indicated with dotted line

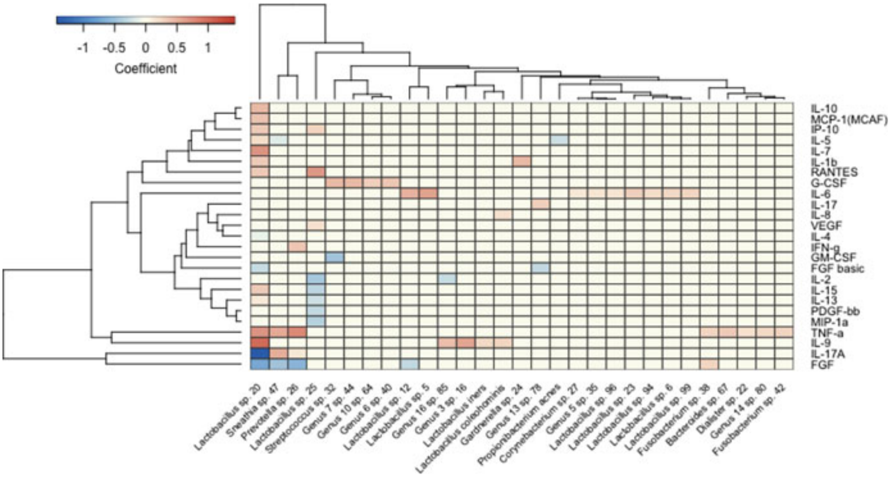

Fig. 5 MOMS-PI study: heatmap of cytokine-taxon associations identified with DM regression model with MRF priors. Taxa are indexed by genus and species

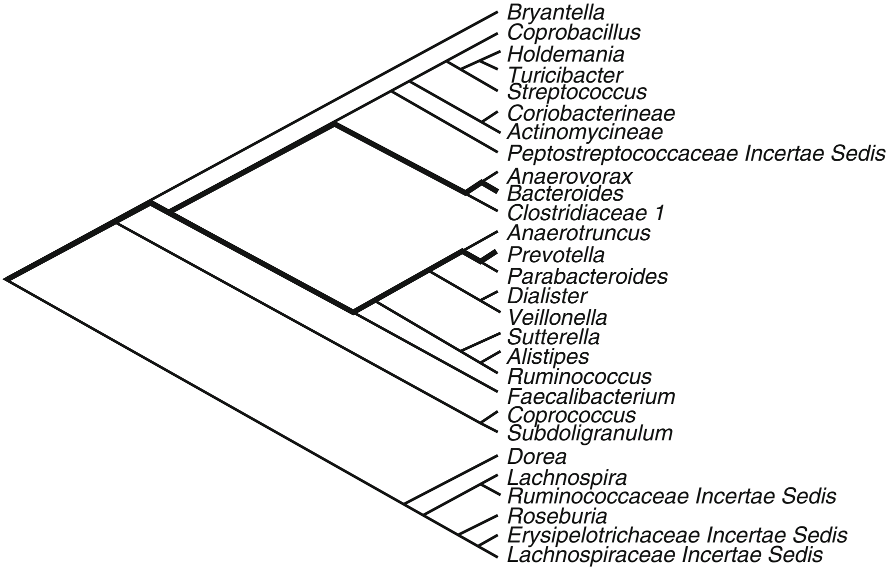

Fig. 6 Gut microbiome study: phylogenetic tree for identifying dietary intake covariates associated with genus-level OTUs in a Dirichlet-tree multinomial model regression

```
MPPI <- selected_DTM( model_gut, burnin = 750)$mppi_
   zeta
bfdr_fit <- bfdr( MPPI, threshold = 0.01 )
out <- selected_DTM( model_gut, burnin = 750,
            threshold = bfdr_fit$threshold,
            cov_lab = colnames(Gut_dietary) )
```

For inference, we are interested in the dietary covariates associated with branches along the path from a particular taxon to the root node. For demonstration, we focus on two genera researchers that have previously targeted in these data [11, 30], Bacteroides and Prevotella. To find the unique covariates associated with the branches corresponding to Bacteroides, run the `branch_covariates` function as below:

```
bact_cov <- branch_covariates( tree = tree, dtm_obj =
model_gut, covariate_name = colnames( Gut_dietary ),
branch_name = "Bacteroides", threshold = bfdr_fit\$
   threshold )
```

This function generates a vector of the unique covariates associated with a given taxon. Note that the `branch_name` provided must match an element in the `covariate_name` vector. In Table 6, we present the dietary intake covariates selected by the Bayesian DTM regression model for Bacteroides and Prevotella.

Table 6 Gut microbiome study: dietary factors identified as associated with Bacteroides and Prevotella using the DTM model with Bayesian variable selection

Bacteroides	Prevotella
Protein	Saturated fat
Saturated fat	Palmitic fatty acid
Palmitic fatty acid	Stearic fatty acid
Stearic fatty acid	Natural food folate
Natural food folate	Retinol equivalents of vitamin A
Vitamin E, food fortification	Vitamin E, food fortification
Maltose	Palmitelaidic trans fatty acid
Total trans	c9,t11 conjug diene isomer 18:2 Linoleic
Isoleucine	Total trans
Lysine	Isoleucine
Phenylalanine	Arginine
Histidine	Serine
Serine	Delphinidin, anthocyanidin
Naringenin, flavanone	Petunidin, anthocyanidin
Delphinidin, anthocyanidin	Proanthocyanidin, trimers
Petunidin, anthocyanidin	
Proanthocyanidin, trimers	
Proanthocyanidin, polymers	

5 Conclusion

In this chapter, we have detailed the use of Dirichlet-multinomial-based approaches with Bayesian variable selection for microbiome studies. We have explored various priors for inclusion indicators using the DM regression model and additionally demonstrated how to incorporate phylogenetic structure into the analysis using DTM models. While we have only shown beta-binomial inclusion indicator priors for the DTM model, the `MicroBVS` package can support MRF priors for DTM models as well. Additionally, the `MicroBVS` package includes functionality to implement the joint model proposed in [12] and additional code to simulate data for each of these models. Step-by-step worked examples using simulated data are provided in the vignette. Frequentist variable selection methods for microbiome data are covered in Chap. 8.

The computational burden of the models described in this chapter is largely dependent on the dimension of the data, tree complexity, prior specification, and the sparsity of the model. For reference, the DTM model run in the gut microbiome analysis took around 9 h to run 150,000 iterations (0.23 seconds/iteration) on a 2.5 GHz dual-core Intel Core i5 processor with 8 GB RAM. To maintain reasonable computation times and selection performance, the authors in [11] recommend applying DTM models to small-to-medium sized microbiome datasets, that is, with less than 100 compositional components and moderate-to-large tree structures when

$B \times P >> n$. Larger datasets might be analyzed by employing the DM models, which do not incorporate the phylogenetic tree. For comparison, the application of the DM model with beta-binomial priors for inclusion indicators took 24 min (0.14 s/iteration) to run with roughly four times as many taxa (123 versus 28). Using the MRF prior with unknown graphical structure also increases the computation time with larger covariate spaces. For our analysis of the MOMS-PI data, the addition of the Gaussian graphical model increased the computation time to 36 min (0.22 s/iteration). As an avenue for future work, variational inference approaches to DM models have shown promising variable selection results [16].

Acknowledgments We would like to thank Hongzhe Lee for sharing the data from the gut microbiome study with us [31]. Additionally, we would like to thank the owners of the MOMS-PI study data for open-sourcing their data and code for extraction.

References

1. Barbieri, M.M., Berger, J.O., et al.: Optimal predictive model selection. Ann. Stat. **32**(3), 870–897 (2004)
2. Brown, P.J., Vannucci, M., Fearn, T.: Multivariate Bayesian variable selection and prediction. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **60**(3), 627–641 (1998)
3. Chen, J., Li, H.: Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. Ann. Appl. Stat. **7**(1), 418–442 (2013)
4. Dennis III, S.Y.: On the hyper-Dirichlet type 1 and hyper-Liouville distributions. Commun. Stat.-Theory Methods **20**(12), 4069–4081 (1991)
5. Eddelbuettel, D., Sanderson, C.: RcppArmadillo: accelerating R with high-performance C++ linear algebra. Comput. Stat. Data Anal. **71**, 1054–1063 (2014)
6. Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., Bates, D.: Rcpp: seamless R and C++ integration. J. Stat. Softw. **40**(8), 1–18 (2011)
7. Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. Math. Geol. **37**(7), 795–828 (2005)
8. George, E.I., McCulloch, R.E.: Approaches for Bayesian variable selection. Stat. Sin. **7**, 339–373 (1997)
9. Integrative, H.: The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe **16**(3), 276 (2014)
10. Knights, D., Parfrey, L.W., Zaneveld, J., Lozupone, C., Knight, R.: Human-associated microbial signatures: examining their predictive value. Cell Host Microbe **10**(4), 292–296 (2011)
11. Koslovsky, M.D., Vannucci, M.: MicroBVS: Dirichlet-tree multinomial regression models with Bayesian variable selection - an R package. BMC Bioinf. **21**, 301 (2020). https://doi.org/10.1186/s12859-020-03640-0
12. Koslovsky, M.D., Hoffman, K.L., Daniel, C.R., Vannucci, M.: A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. Ann. Appl. Stat. **14**(3), 1471–1492 (2020)
13. La Rosa, P.S., Brooks, J.P., Deych, E., Boone, E.L., Edwards, D.J., Wang, Q., Sodergren, E., Weinstock, G., Shannon, W.D.: Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS One **7**(12), e52078 (2012)
14. Li, H.: Microbiome, metagenomics, and high-dimensional compositional data analysis. Ann. Rev. Stat. Appl. **2**, 73–94 (2015)

15. Li, F., Zhang, N.R.: Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. J. Am. Stat. Assoc. **105**(491), 1202–1214 (2010)
16. Miao, Y., Kook, J.H., Lu, Y., Guindani, M., Vannucci, M.: Scalable Bayesian variable selection regression models for count data. In: Flexible Bayesian Regression Modelling, pp. 187–219. Elsevier, Amsterdam (2020)
17. Minka, T.: The Dirichlet-tree distribution (1999)
18. Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P.: Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics **5**(2), 155–176 (2004)
19. Peterson, C.B., Stingo, F.C., Vannucci, M.: Joint Bayesian variable and graph selection for regression models with network-structured predictors. Stat. Med. **35**(7), 1017–1031 (2016)
20. Richardson, S., Bottolo, L., Rosenthal: Bayesian models for sparse regression analysis of high dimensional data. In: Bayesian Statistics, vol. 9, pp. 539–569. Oxford University Press, Oxford (2010)
21. Savitsky, T., Vannucci, M., Sha, N.: Variable selection for nonparametric Gaussian process priors: models and computational strategies. Stat. Sci.: Rev. J. Inst. Math. Stat. **26**(1), 130–149 (2011)
22. Schwarz, G., et al.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
23. Shetty, S.A., Hugenholtz, F., Lahti, L., Smidt, H., de Vos, W.M.: Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. FEMS Microbiol. Rev. **41**(2), 182–199 (2017)
24. Stingo, F.C., Chen, Y.A., Vannucci, M., Barrier, M., Mirkes, P.E.: A Bayesian graphical modeling approach to microRNA regulatory network inference. Ann. Appl. Stat. **4**(4), 2024–2048 (2010)
25. Stingo, F.C., Chen, Y.A., Tadesse, M.G., Vannucci, M.: Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. Ann. Appl. Stat. **5**(3), 1978–2002 (2011)
26. Tang, Z.Z., Chen, G., Alekseyenko, A.V., Li, H.: A general framework for association analysis of microbial communities on a taxonomic tree. Bioinformatics **33**(9), 1278–1285 (2017)
27. Tang, Y., Ma, L., Nicolae, D.L., et al.: A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. Ann. Appl. Stat. **12**(1), 1–26 (2018)
28. Wadsworth, W.D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S.A., Vannucci, M.: An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. BMC Bioinf. **18**(1), 94 (2017)
29. Wang, H., et al.: Scaling it up: Stochastic search structure learning in graphical models. Bayesian Anal. **10**(2), 351–377 (2015)
30. Wang, T., Zhao, H.: A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. Biometrics **73**(3), 792–801 (2017)
31. 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.: Linking long-term dietary patterns with gut microbial enterotypes. Science **334**(6052), 105–108 (2011)
32. Xia, Y., Sun, J.: Hypothesis testing and statistical analysis of microbiome. Genes Dis. **4**(3), 138–148 (2017)
33. Xu, Z., Knight, R.: Dietary effects on human gut microbiome diversity. Br. J. Nutr. **113**(S1), S1–S5 (2015)
34. Zhang, Y., Zhou, H., Zhou, J., Sun, W.: Regression models for multivariate count data. J. Comput. Graph. Stat. **26**(1), 1–13 (2017)