Biometrics 69, 1053–1063
December 2013
DOI: 10.1111/biom.12079

A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis

Fan Xia,[1] Jun Chen,[2] Wing Kam Fung,[1] Hongzhe Li[2,*]

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong
[2]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
*email: hongzhe@upenn.edu

Summary. Changes in human microbiome are associated with many human diseases. Next generation sequencing technologies make it possible to quantify the microbial composition without the need for laboratory cultivation. One important problem of microbiome data analysis is to identify the environmental/biological covariates that are associated with different bacterial taxa. Taxa count data in microbiome studies are often over-dispersed and include many zeros. To account for such an over-dispersion, we propose to use an additive logistic normal multinomial regression model to associate the covariates to bacterial composition. The model can naturally account for sampling variabilities and zero observations and also allow for a flexible covariance structure among the bacterial taxa. In order to select the relevant covariates and to estimate the corresponding regression coefficients, we propose a group ℓ_1 penalized likelihood estimation method for variable selection and estimation. We develop a Monte Carlo expectation-maximization algorithm to implement the penalized likelihood estimation. Our simulation results show that the proposed method outperforms the group ℓ_1 penalized multinomial logistic regression and the Dirichlet multinomial regression models in variable selection. We demonstrate the methods using a data set that links human gut microbiome to micro-nutrients in order to identify the nutrients that are associated with the human gut microbiome enterotype.

Key words: Hierarchical model; Markov chain Monte Carlo; Over-dispersion; Regularization; Variable selection.

1. Introduction

The human microbiome is the collection of microorganisms that reside in various human body sites, including gut and lung airway. The genomes of these microbiomes comprise an integral part of our genetic and metabolic landscape and consequently contribute to our normal physiology and predisposition to disease (The Human Microbiome Project Consortium, 2012). The targeted amplicon sequencing of 16S ribosomal RNA is commonly used for bacterial ecology studies. The 16S rRNA gene is omnipresent in bacterial organisms and has slowly evolving regions that allow for amplification and fast evolving regions that allow for identification (Kuczynski et al., 2012). To determine bacterial composition, researchers first PCR-amplify the DNA strands in some variable regions of the 16S rRNA gene and then assign the targeted amplicon raw sequences to samples using barcodes. The processed sequences are often further clustered into operational taxonomic units (OTUs) at a certain similarity level in a taxonomic independent way (Chaffron et al., 2010; Caporaso et al., 2010). The OTUs are each characterized by a representative DNA sequence and can be assigned to taxonomic lineages by comparing to the known 16S rRNA database, such as the Ribosomal Database Project (Cole et al., 2009). Finally, at a given taxonomic level (e.g., the genus level), by aggregating the OTUs at the same taxon, taxonomic counts are obtained to characterize the microbiome composition.

The human gut is inhabited by the largest number of microbes with more than 3×10^6 genes. Recent studies have associated the human gut microbiome to obesity and inflammatory bowel disease (see Virgin and Todd, 2011 for a review). It is therefore important to understand the biological/environmental covariates that are associated with microbiome composition. The gut microbiome is dominated by two bacterial divisions, Bacteroidetes and Firmicutes at the phylum level. At the genus level, Arumugam et al. (2011) concluded that the variation of gut microbiome composition is also generally stratified and they classified the human gut microbiome into three discrete enterotypes based on different compositions of Bacteroides, Prevetella, and Ruminococcus. Wu et al. (2011) linked diet intake to the human gut microbiome by performing a simple correlation-based analysis. Our research is motivated by this ongoing microbiome study at the University of Pennsylvania. In this study, stool samples of 98 healthy volunteers were collected and the gut bacterial count data were obtained at different taxonomic levels based on 16S rRNA sequencing. In addition, the food frequency questionnaire was used to collect diet information about these subjects, from which measurements on 214 micro-nutrients were obtained. One of the goals of the study is to identify the micronutrients that are associated with the bacterial composition.

The focus of this article is to develop statistical methods for identifying the covariates that are associated with the gut bacterial composition. Many regression models have been developed for multivariate count or compositional data. Among these, the multinomial logistic (ML) regression model is most commonly used. However, the ML model does not allow for over-dispersion of count data, which is often observed for 16S-based count data. To allow for over-dispersion, Dirichlet multinomial (DM) regression can be applied (Chen and Li, 2013). One drawback of the DM model is that it has a

limited number of parameters to adequately model the variances and covariances of the composition. In particular, the dependence structure between Dirichlet variates cannot be determined independently of their mean values; Dirichlet variates are always negatively correlated, which may not represent the nature of microbiome data. In this article, we propose to apply the additive logistic normal multinomial (LNM) regression model (Aitchison, 1982; Billheimer, Guttorp, and Fagan, 2001) to link covariates with taxonomic counts. In the LNM model, the observed counts are modeled by a multinomial distribution and the underlying bacterial composition is taken as random variables and modeled by a logistic normal (LN) distribution, in which the proportions are transformed to follow a multivariate normal distribution. As suggested by Aitchison (1982), the LN distribution allows for a more flexible covariance structure than the Dirichlet model. However, estimation of such LNM models is not simple, especially when variable selection is the focus of the study. Billheimer et al. (2001) developed a full Bayesian approach using Markov chain Monte Carlo (MCMC) to fit such an LNM model when the number of covariates is small. In addition, Billheimer et al. (2001) only considered the setting when the number of covariates is small and variable selection is not considered.

Since the main goal of our study is to select the nutrients that are associated with the bacterial composition, variable selection is an essential step in fitting the LNM model. To the best of our knowledge, variable selection for the LNM regression model has not been studied in literature. In this article, we develop a group-penalized likelihood estimation procedure for the LNM model by introducing a group ℓ_1 penalty function to select the relevant variables. Such a group penalty function has been applied in other multivariate regression settings (Meier, van de Geer, and Buhlmann, 2008; Peng et al., 2010). However, because no closed-form likelihood function is available for the LNM regression, the standard coordinate descent algorithm cannot be applied directly to maximize the penalized likelihood function. To address this difficulty, we develop a Monte Carlo EM (MCEM) algorithm to implement the penalized likelihood estimation, where a Metropolis–Hastings (MH) algorithm is used in the expectation (E)-step to sample the unobserved compositions and a group-penalized least-square estimation is used in the maximization (M)-step to select the relevant variables.

The article is structured as follows. Section 2 introduces the LNM regression model linking the high dimensional covariates to microbiome compositional data. Section 3 presents the MCEM algorithm for fitting the models and for selecting the relevant variables. Section 4 presents simulation results to compare the variable selection performance of the LNM model with other models. Section 5 applies the proposed method to the human gut microbiome data set to identify the nutrients that affect the bacterial composition that determines the enterotype. Finally, a brief discussion of the methods and results is given in Section 6.

2. An Additive Logistic Normal Multinomial Regression Model for Microbiome Compositional Data

Consider a human microbiome study where we have obtained the count data on $K+1$ taxa. Let $\mathbf{W} = (W_1, \ldots, W_{K+1})^{\mathrm{T}}$ denote the random vector of counts of $K+1$ bacterial taxa. Let $M = \sum_{k=1}^{K+1} W_k$ be the total number of counts that is determined by the sequencing depth. We treat M as an ancillary statistic and perform the analysis conditioning on M. Let $\mathbf{Z} = (Z_1, \ldots, Z_{K+1})^{\mathrm{T}}$ be the underlying composition of the microbial taxa, where $\sum_{k=1}^{K+1} Z_k = 1$. This implies that $\mathbf{Z}$ is a random vector in K-dimensional simplex $\mathbb{S}^K$. Conditioning on $\mathbf{Z}$, we model the counts $\mathbf{W}$ using a multinomial distribution with the conditional density function given by

$$Pr(\mathbf{W}|\mathbf{Z}) \propto \prod_{k=1}^{K+1} (Z_k)^{W_k}.$$

The expectation and variance of the kth component are $\mathrm{E}(W_k) = MZ_k$ and $\mathrm{Var}(W_q) = MZ_k(1-Z_k)$, respectively. In addition, let $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a p-dimensional vector of the measured covariates. In our data analysis, this represents p micro-nutrient measurements.

One common approach to linking the covariate $\mathbf{X}$ to the composition $\mathbf{Z}$ is through a multinomial-logistic (ML) regression model. However, the actual variation in the observed counts in the microbiome studies is often higher than that the ML model implies. This over-dispersion comes from random variation in the underlying compositions due to individual heterogeneity or the correlations between the taxa. To model such over-dispersed data, we consider an additive logistic normal model for $\mathbf{Z}$. This prior distribution is based on an additive log-ratio transformation (ϕ) to map $\mathbf{Z}$ from the restricted simplex $\mathbb{S}^K$ to the K-dimensional open real space $\mathbb{R}^K$ (Aitchison, 1982). The log-ratio transformation is defined as

$$\mathbf{Y} = \phi(\mathbf{Z}) = \left\{ \log\left(\frac{Z_1}{Z_{K+1}}\right), \ldots, \log\left(\frac{Z_K}{Z_{K+1}}\right) \right\}^{\mathrm{T}}, \qquad (1)$$

where the $(K+1)$th taxon is treated as the base group. The inverse operator, the additive logistic transformation (ϕ^{-1}), which maps $\mathbf{Y}$ to $\mathbf{Z}$, with its kth component given by

$$Z_k = (\phi^{-1}(\mathbf{Y}))_k = \frac{\exp(Y_k)}{\sum_{k=1}^{K} \exp(Y_k) + 1}, \; k = 1, \ldots, K$$

and

$$Z_{K+1} = \frac{1}{\sum_{k=1}^{K} \exp(Y_k) + 1}.$$

We then assume that the random variable $\mathbf{Y}$ follows a multivariate normal distribution $N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the density function

$$f(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\, \boldsymbol{\Sigma} \,|^{-1/2} \exp\left\{ -\tfrac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \right\},$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix of the log-ratio transformed compositional data. In order to model the effects of covariates on the latent bacterial composition, following Billheimer et al. (2001), we

assume that

$$\mu_k = \beta_{k0} + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_k,$$

for $k = 1, \ldots, K$, where β_{k0} is the intercept and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^{\mathrm{T}}$ is the p-dimensional regression coefficient that measures the effects of the p covariates on the kth mean. Let $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{K0})^{\mathrm{T}}$ be the intercept vector and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ be the $p \times K$ dimensional coefficient matrix. Further, we denote the lth row of the coefficient matrix $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_{l\cdot}$ for $l = 1, \ldots, p$.

To assist the interpretation of the regression coefficient matrix $\boldsymbol{\beta}$, following Aitchison (1986) and Billheimer et al. (2001), we define a perturbation operator $\oplus$ for the composition $\mathbf{Z}$ and another variable $\mathbf{u}$ in $\mathbb{S}^K$ by

$$\mathbf{Z}' = \mathbf{Z} \oplus \mathbf{u} = \left(\frac{Z_1 u_1}{\sum_{k=1}^{K+1} Z_k u_k}, \frac{Z_2 u_2}{\sum_{k=1}^{K+1} Z_k u_k}, \ldots, \frac{Z_{K+1} u_{K+1}}{\sum_{k=1}^{K+1} Z_k u_k} \right)^{\mathrm{T}}.$$

This operator represents the original composition $\mathbf{Z}$ being "perturbed" by $\mathbf{u}$ to form a new composition $\mathbf{Z}'$. When $u_k > (K+1)^{-1}$, the kth taxon increases its relative abundance. Based on the transformed data, we have $\phi(\mathbf{Z}') = \phi(\mathbf{Z}) + \phi(u)$. It is also easy to verify that $\phi(\mathbf{Z}^a) = a\phi(\mathbf{Z})$ and $\phi(\mathbf{J}_K) = 0$, where a is any number in real space and $\mathbf{J}_K = \{1/(K+1), \ldots, 1/(K+1)\}^{\mathrm{T}}$ is the center of simplex $\mathbb{S}^K$, representing identical composition.

Based on this perturbation operator, we have

$$\phi^{-1}\left(\boldsymbol{\beta}_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right) = \phi^{-1}(\boldsymbol{\beta}_0) \oplus \phi^{-1}(\boldsymbol{\beta}_{1\cdot})^{X_1} \oplus \cdots \oplus \phi^{-1}\left(\boldsymbol{\beta}_{p\cdot}\right)^{X_p},$$

where $\boldsymbol{\beta}_{l\cdot}$ represents the lth row of the $p \times K$ coefficient matrix $\boldsymbol{\beta}$. The vector $\phi^{-1}(\boldsymbol{\beta}_0)$ represents the location of the $K+1$ components in the simplex $\mathbb{S}^K$ without disturbance from the covariates, therefore, it is interpreted as the "baseline composition." The vector $\phi^{-1}(\boldsymbol{\beta}_{l\cdot})$ measures the shift in composition from the baseline by a unit change in the covariate l when the other covariates are unchanged. Furthermore, when $\left\{\phi^{-1}(\boldsymbol{\beta}_{l\cdot})\right\}_k > 1/(K+1)$, the covariate l is positively associated with the kth taxon and otherwise, the association is negative. The magnitude of the disturbance in bacterial compositions from one unit change in covariate l is measured by

$$\| \phi^{-1}(\boldsymbol{\beta}_{l\cdot}) \|_2 = \sqrt{\boldsymbol{\beta}_{l\cdot}(\mathbf{I}_K + \mathbf{1}_K\mathbf{1}_K^{\mathrm{T}})^{-1}\boldsymbol{\beta}_{l\cdot}^{\mathrm{T}}}.$$

where $\mathbf{I}_K$ is the $K \times K$ identity matrix and $\mathbf{1}_K$ is the vector of 1s.

3. Variable Selection and Parameter Estimation Via Penalized Likelihood Estimation

Suppose we have n *i.i.d* copies of the random vector $(\mathbf{W}, \mathbf{X})$, denoted by $(\mathbf{W}_i, \mathbf{X}_i)$ for $i = 1, \ldots, n$. Let $\mathbf{Z}_i \in S^K$ be the unobserved composition for the ith samples and $\mathbf{Y}_i = \phi(\mathbf{Z}_i)$ be its log-ratio transformation as defined in (1). We consider the scenario when the number of the covariates p is large and our main goal of analysis is to select the variables that have an effect on the bacterial composition. Even when p is small, the likelihood-based inference of the LNM model is not simple because no closed-form log-likelihood function is available.

We propose to develop a penalized likelihood approach to select the covariates that are associated with the bacterial composition. Note that if the lth covariate is not associated with the bacterial composition, we have $\boldsymbol{\beta}_{l\cdot} = 0$ and therefore $\| \boldsymbol{\beta}_{l\cdot} \|_2 = (\sum_{k=1}^{K} \boldsymbol{\beta}_{lk}^2)^{1/2} = 0$. This motivates estimating the parameters $\boldsymbol{\eta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ based on maximizing the following penalized log-likelihood function,

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} l(\boldsymbol{\eta}; \mathbf{W}_i, \mathbf{X}_i) - \lambda \sum_{l=1}^{p} \| \boldsymbol{\beta}_{l\cdot} \|_2 \right\}, \quad (2)$$

where $l(\boldsymbol{\eta}; \mathbf{W}_i, \mathbf{X}_i)$ is the log-likelihood function of the ith observation based on the LNM model, and λ is the tuning parameter. The group ℓ_1 penalty function (Yuan and Lin, 2006) combines information across all the taxa and induces row sparsity of the coefficient matrix $\boldsymbol{\beta}$. Since there is no closed-form expression of the log-likelihood function $l(\boldsymbol{\eta}; \mathbf{W}_i, \mathbf{X}_i)$, we cannot perform the optimization (2) directly.

3.1. *An MCEM Algorithm for LNM Model Estimation*

For a given tuning parameter λ, the complete data group ℓ_1 penalized log-likelihood can be written as

$$\begin{aligned} l(\boldsymbol{\eta}) &= \log\left[\prod_{i=1}^{n} \left\{ Pr(\mathbf{W}_i|\mathbf{Y}_i) \times f(\mathbf{Y}_i|\mathbf{X}_i, \eta) \right\}\right] - \lambda \sum_{l=1}^{p} \| \boldsymbol{\beta}_{l\cdot} \|_2 \\ &= \sum_{i=1}^{n} \log\left\{ Pr(\mathbf{W}_i|\mathbf{Y}_i) \right\} - \frac{1}{2} n \log(|\boldsymbol{\Sigma}|) \\ &\quad - \frac{1}{2} \sum_{i=1}^{n} \left\{ \left(\mathbf{Y}_i - \boldsymbol{\beta}_0 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{Y}_i - \boldsymbol{\beta}_0 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\right) \right\} \\ &\quad - \lambda \sum_{l=1}^{p} \| \boldsymbol{\beta}_{l\cdot} \|_2 . \end{aligned} \quad (3)$$

To implement the EM algorithm, at the tth E-step we need to compute the expected complete data penalized log-likelihood (3),

$$\begin{aligned} &Q\left(\eta|\eta^{(t-1)}\right) \\ &= -\frac{1}{2} n \log(|\boldsymbol{\Sigma}|) \\ &\quad - \frac{1}{2} \sum_{i=1}^{n} E\{(\mathbf{Y}_i - \boldsymbol{\beta}_0 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\beta}_0 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\} \\ &\quad - \lambda \sum_{l=1}^{p} \| \boldsymbol{\beta}_{l\cdot} \|_2, \end{aligned} \quad (4)$$

where the expectation is with respect to the conditional distribution of $\mathbf{Y}_i|(\mathbf{W}_i, \mathbf{X}_i; \boldsymbol{\eta}^{(t-1)})$, and $\boldsymbol{\eta}^{(t-1)} = (\boldsymbol{\beta}_0^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)})$ are the parameter values at the $(t-1)$th M-step.

For the ith sample, the conditional distribution of $\mathbf{Y}_i$ given the observed data and the current parameter values is

$$\pi(\mathbf{Y}_i|\mathbf{W}_i,\mathbf{X}_i) \propto Pr(\mathbf{W}_i|\mathbf{Y}_i)\times f(\mathbf{Y}_i|\mathbf{X}_i,\eta^{(t-1)}) \propto \frac{\prod_{k=1}^{K}\exp(W_{ik}Y_{ik})}{\left[\left\{\sum_{k=1}^{K}\exp(Y_{ik})+1\right\}^{M_i}\right]} \times \exp\left[-\frac{1}{2}\left\{\mathbf{Y}_i^{(t-1)*T}\left(\boldsymbol{\Sigma}^{(t-1)}\right)^{-1}\mathbf{Y}_i^{(t-1)*}\right\}\right],$$

where $\mathbf{Y}_i^{(t-1)*}=\mathbf{Y}_i-\boldsymbol{\beta}_0^{(t-1)}-(\boldsymbol{\beta}^{(t-1)})^{\mathrm{T}}\mathbf{X}_i$. In order to compute the conditional expectations, we use a MH algorithm to sample from this conditional distribution. Specifically, for the rth MH step, we propose a new vector $\mathbf{Y}_i$ from a multivariate normal distribution, $q(\mathbf{Y}_i|\mathbf{Y}_i^{(r-1)})=N(\mathbf{Y}_i^{(r-1)},I)$, and then calculate the Metropolis acceptance ratio

$$r\left(\mathbf{Y}_i|\mathbf{Y}_i^{(r-1)}\right)=\min\left(1,\frac{\pi(\mathbf{Y}_i)}{\pi\left(\mathbf{Y}_i^{(r-1)}\right)}\right). \tag{5}$$

We then draw a random number u from the uniform distribution $U(0,1)$ and accept the proposed new value $\mathbf{Y}_i$ if $u\le r(\mathbf{Y}_i|\mathbf{Y}_i^{(r-1)})$ and keep the previous value otherwise. After 500 initial burn-ins, we choose the next R MH samples to calculate the conditional expectations in the E-step.

For the M-step, it is easy to check that the updating formula for $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}_0$ are given as

$$\boldsymbol{\Sigma}^{(t)}=\frac{1}{R}\sum_{r=1}^{R}\left\{\frac{\sum_{i=1}^{n}\left(\mathbf{Y}_i^{(r*)}\right)\left(\mathbf{Y}_i^{(r*)}\right)^{\mathrm{T}}}{n}\right\},\quad \boldsymbol{\beta}_{0k}^{(t)}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{R}\sum_{r=1}^{R}Y_{ik}^{(r)}\right)$$

for $k=1,...,K$, where $\mathbf{Y}^{(r)*}=\mathbf{Y}_i^{(r)}-\boldsymbol{\beta}_0^{(t-1)}-(\boldsymbol{\beta}^{(t-1)})^{\mathrm{T}}\mathbf{X}_i$.

Given $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}_0$, we next update $\boldsymbol{\beta}$ by minimizing the following negative penalized expected value of the complete data log-likelihood,

$$Q^*(\boldsymbol{\beta}|\boldsymbol{\beta}_0,\boldsymbol{\Sigma})=\frac{1}{2}\sum_{i=1}^{n}\left(E(\mathbf{Y}_i)-\boldsymbol{\beta}_0^{(t)}-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\right)^{\mathrm{T}}\left(\boldsymbol{\Sigma}^{(t)}\right)^{-1} \times\left(E(\mathbf{Y}_i)-\boldsymbol{\beta}_0^{(t)}-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\right)+\lambda\sum_{l=1}^{p}\parallel\boldsymbol{\beta}_{l\cdot}\parallel_2,$$

where $E(\mathbf{Y}_i)$ is estimated by $\bar{\mathbf{Y}}_i^{(t)}=1/R\sum_{r=1}^{R}\mathbf{Y}_i^{(r)}$. Since $\boldsymbol{\Sigma}^{(t)}$ is a positive-definite matrix, this is a convex optimization problem and has a unique solution. We can employ the block coordinate descent algorithm to find $\boldsymbol{\beta}$. Specifically, we write $(\boldsymbol{\Sigma}^{(t)})^{-1}=\mathbf{U}^{\mathrm{T}}\mathbf{U}$ and define $\tilde{\mathbf{Y}}_i=\mathbf{U}\bar{\mathbf{Y}}_i^{(t)}$, $\tilde{\boldsymbol{\beta}}_0=\mathbf{U}\boldsymbol{\beta}_0^{(t)}$ and $\tilde{\boldsymbol{\beta}}^{\mathrm{T}}=(\tilde{\boldsymbol{\beta}}_{1\cdot}^{\mathrm{T}},...,\tilde{\boldsymbol{\beta}}_{p\cdot}^{\mathrm{T}})$ with $\tilde{\boldsymbol{\beta}}_{l\cdot}^{\mathrm{T}}=\mathbf{U}\boldsymbol{\beta}_{l\cdot}$ for $l=1,...,p$. Since $\mathbf{U}$ is a lower diagonal matrix, $\boldsymbol{\beta}_{l\cdot}$ is a zero vector if and only if $\tilde{\boldsymbol{\beta}}_{l\cdot}$ is a zero vector. Then the objective function $Q^*(\boldsymbol{\beta}|\boldsymbol{\beta}_0,\boldsymbol{\Sigma})$ can be written as

$$Q^{**}\left(\tilde{\boldsymbol{\beta}}|\tilde{\boldsymbol{\beta}}_0\right)=\frac{1}{2}\sum_{i=1}^{n}\left(\tilde{\mathbf{Y}}_i-\tilde{\boldsymbol{\beta}}_0-\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i\right)^{\mathrm{T}}\left(\tilde{\mathbf{Y}}_i-\tilde{\boldsymbol{\beta}}_0-\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i\right)+\lambda\sum_{l=1}^{p}\parallel\tilde{\boldsymbol{\beta}}_{l\cdot}\parallel_2. \tag{6}$$

Minimizing (6) is then reduced to the standard group lasso problem for multivariate regression and the coordinate descent algorithm developed in Peng et al. (2010) can be applied directly.

Since MCMC is used in the E-step of the EM algorithm, we do not expect exact convergence of the parameter values. Instead we check to make sure that the estimates are stabilized by examining the plots of the parameter estimates. For our simulated and real data analysis, the algorithm stabilizes very fast, usually within 10 EM steps. Due to the uncertainty of the sampling in the E-step, after the estimates stabilize, we run additional S EM steps and take the median of the estimates as the final estimates of the parameters. We use $S=5$ for simulations and $S=10$ for real data analysis.

3.2. *Tuning Parameter Selection*

We use a fivefold cross-validation to choose the tuning parameter λ. For each λ, we partition the samples into five non-overlapping subsets. Denote the subset v by $\{\mathbf{W}^{(v)},\mathbf{X}^{(v)}\}$ and the collection of the remaining subsets of individuals by $\{\mathbf{W}^{(-v)},\mathbf{X}^{(-v)}\}$. Based on $\{\mathbf{W}^{(-v)},\mathbf{X}^{(-v)}\}$, we perform the MCEM algorithm to obtain $\hat{\boldsymbol{\beta}}_0^{(-v)}(\lambda)$, $\hat{\boldsymbol{\beta}}^{(-v)}(\lambda)$, and then calculate the relative prediction error of data set $\{\mathbf{W}^{(v)},\mathbf{X}^{(v)}\}$ by

$$\mathrm{MSPE}^{(v)}(\lambda)=\frac{1}{n^{(v)}}\sum_{i=1}^{n^{(v)}}\frac{\parallel\mathbf{W}_i^{(v)}-\hat{\mathbf{W}}_i^{(v)}\parallel_2^2}{\parallel\mathbf{W}_i^{(v)}\parallel_2^2},$$

where $n^{(v)}$ is the number of samples in the fold v, $\mathbf{W}_i^{(v)}$, and $\hat{\mathbf{W}}_i^{(v)}$ are the observed and predicted count vectors for the ith sample in the fold, with $\hat{\mathbf{W}}_i^{(v)}=M_i^{(v)}\phi^{-1}(\hat{\boldsymbol{\beta}}_0^{(-v)}(\lambda)+\mathbf{X}_i^{(v)}\hat{\boldsymbol{\beta}}^{(-v)}(\lambda))$ and $M_i^{(v)}$ being the corresponding total count. The cross validation score is given by $CV(\lambda)=\sum_{v=1}^{5}\mathrm{MSPE}^{(v)}(\lambda)$. The tuning parameter λ with the smallest score is selected.

4. Simulation Evaluations

4.1. *Models for Simulations and Methods Compared*

We simulate data sets with taxa count data $\mathbf{W}$ and covariate data $\mathbf{X}$. Given $(n,p,K+1)$, we generate the the compositional vector of the i-th individual by

$$\mathbf{Z}_i=\phi^{-1}\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}+\epsilon_i\right),$$

with the component $K+1$ being set as the base component in the operator ϕ^{-1}. The covariate vector $\mathbf{X}_i$ is simulated from a p-dimensional multivariate normal distribution with mean

zero and a polynomial decay covariance matrix Σ_X given by $(\mathbf{\Sigma_X})_{pp'} = \rho_{\mathbf{X}}^{|p-p'|}$. To simulate a sparse covariate effect, we select $q = \alpha p$ number of relevant covariates among the p covariates with α controlling the model sparsity. The non-zero elements of the coefficient matrix $\boldsymbol{\beta}$ is generated from a uniform distribution over the interval $[-3, -1) \bigcup (1, 3]$. The error vector ϵ_i is simulated from $N(0, \Sigma_\epsilon)$ with $(\mathbf{\Sigma}_\epsilon)_{ll'} = \sigma_\epsilon^2 \rho_\epsilon^{|l-l'|}$, where σ_ϵ^2 is chosen to control the averaged signal-to-noise ratio (SNR). The level of SNR determines the degree of over-dispersion in the model, where a smaller SNR indicates an overall lower influence from the covariates on the compositions, therefore generating a higher level of over-dispersion in the model. In contrast, a larger SNR specifies a lower level of over-dispersion. In the extreme case when there is no extra variability except for covariates' effects, that is when $\epsilon_i's$ are fixed to be zeros, the data set is simulated without over-dispersion and SNR goes to infinite. Finally, we generate the taxa read depth M_i from a uniform distribution over the interval $[m_{\min}, m_{\max}]$, and simulate the observed count vector for individual i by $\mathbf{W}_i \sim \text{Multinomial}(M_i, \mathbf{Z}_i)$.

We compare the variable selection performance of the LNM model with three alternative models with similar group ℓ_1 penalization:

1. The logistic normal (LN) model that treats the log-ratio transformation as the multivariate response and then applies the group variable selection procedure of Peng et al. (2010) to select the variables.
2. The Dirichlet multinomial (DM) model with group ℓ_1 penalty function proposed by Chen and Li (2013).
3. The multinomial logistic (ML) model that models the count data directly with the group ℓ_1 penalized multinomial logistic regression model.

In the LN model, the zero counts in the simulated data sets are replaced by a pseudo count value 0.05 to facilitate the estimation. The reference component is randomly assigned among the $K + 1$ components in constructing the log-ratios. Fivefold cross-validation is used for choosing the tuning parameter in all four models.

We consider several different parameter settings to assess the effects of the model sparsity, over-dispersion and the level of read depth on variable selection performance. For each model, 50 replications are conducted and the variable selection performance is evaluated by the averages of the following measures

$$\text{recall} = \frac{TP}{TP + FN}, \quad \text{precision} = \frac{TP}{TP + FP},$$
$$\text{F1} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}, \tag{7}$$

where TP and FP are the number of true positives and false positives, respectively, and F_1 gives an overall measurement of the variable selection performance.

4.2. *Simulation Results*

For the first set of simulations, we examine the effect of the proportion of relevant covariates on variable selection performance. We choose $n = 75$, $p = 100$, $K + 1 = 3$, and consider three levels of model sparsity, $\alpha = 0.05, 0.1, 0.15$. The other parameters are chosen as SNR=4.5, $\rho_{\mathbf{X}} = 0.5$, $\rho_\epsilon = 0.5$, $m_{\min} = 100$, $m_{\max} = 500$. The top panel of Figure 1 presents the averaged values of recall, precision and F_1 values for models with different sparsity levels. When the model is sparse ($\alpha = 0.05$), we observe that the four methods have similar performance with the proposed LNM being slightly better than the other three. As α increases and the model becomes less sparse, the selection performance is reduced for all four methods. However, the LNM model shows higher recall rates than other methods. For example, when $\alpha = 0.15$, the LNM identifies on average about 75% of the relevant covariates while LN, DM, and ML models only identify about 60%, 28%, and 20% of the relevant variables, respectively. In terms of precision, the LN and LNM methods show lower precision with increasing α. However, in terms of the overall measure of F_1 score, the LNM still outperforms the other three methods.

We next evaluate the effect of over-dispersion on model selection performance. We consider models with large, moderate and no overdispersion by fixing SNR at 1.5, 4.5, and ∞, respectively. We choose $n = 75$, $p = 100$, $K + 1 = 3$, $\alpha = 0.1$, $m_{\min} = 100$, $m_{\max} = 500$, $\rho_{\mathbf{X}} = 0.5$, $\rho_\epsilon = 0.5$. The results are shown in the middle panel of Figure 1. For all methods, smaller over-dispersion results in better variable selection performance. The LNM method has higher recall rates than other methods with lower precision rates. However, the overall variable selection performance of LNM as measured by the F_1 score is better than the other three models. This is largely due to the selection of the tuning parameter using cross-validation since when the over-dispersion is very large, the CV tends to select larger models to achieve better prediction. This leads to selection of more covariates in the model and lower precision and higher recall rates. If a smaller tuning parameter is used, the proposed model can result in similar rate of precision as the other models we compared.

We lastly examine the performance of variable selection for data of different sequencing depths. Note that sequencing depth determines the number of counts we observe for each of the bacterial taxa. We choose $n = 75$, $p = 100$, $K + 1 = 5$, $\alpha = 0.1$, SNR=4.5, $\rho_{\mathbf{X}} = 0.5$, and $\rho_\epsilon = 0.5$. We consider $\{m_{\min}, m_{\max}\} = (\{10, 50\}, \{100, 500\}, \{1000, 2000\})$. The results are shown in the bottom panel of Figure 1. We can see that the LNM method outperforms the other methods with higher recall rates and F1 scores for all the settings with only slightly lower precision rates. Furthermore, the performances of the ML and DM methods heavily depend on the total number of the taxa counts. When the sequencing depth is large, the LM and the LNM methods perform similarly.

4.3. *Computational Complexity, Sensitivity to the MH Samples, and Convergence of the MCEM*

To assess the sensitivity of the parameter estimates to the number of the MH samples in our MCMC implementation, we simulate a data set with $n = 75$, $p = 100$, $K + 1 = 3$, $\rho_X = 0.5$, $\rho_\epsilon = 0.5$, and $\alpha = 0.05$. The non-zero elements of the coefficient matrix β are fixed to be 2. Further we set SNR=1.5 and $m_{\min} = 10$, $m_{\max} = 50$. This model corresponds to a high level of over-dispersion and large multinomial sampling variation. We apply fivefolds cross-validation to select the tuning

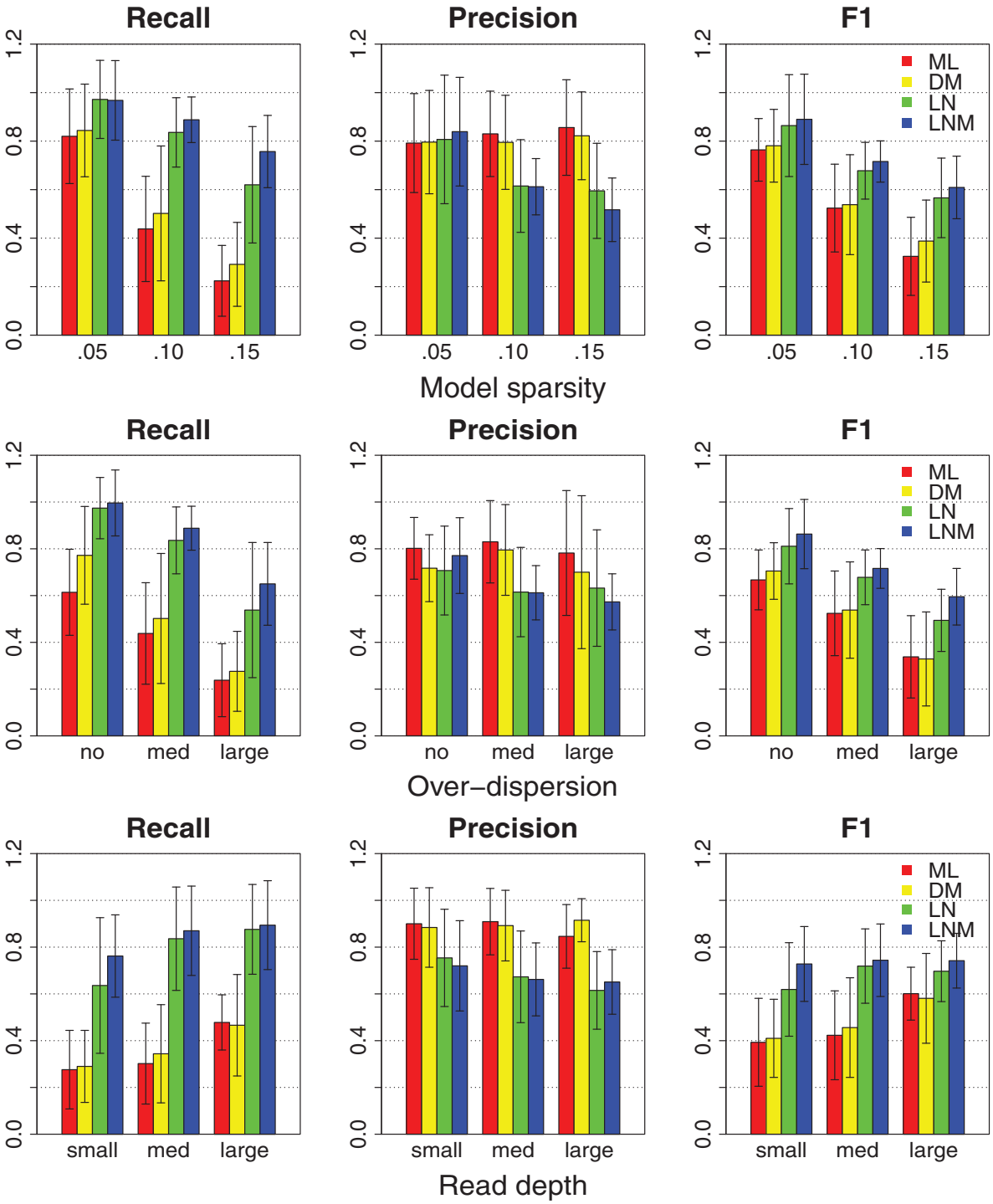

Figure 1. Simulation study to evaluate the effects of the proportion of the relevant covariates (top panel), effects of over-dispersion (middle panel) and the effects of the sequencing depths (bottom panel) on the performance of three different methods for identifying the relevant covariate. For each model and method, the recall, precision, and F1 measurements are reported as averaged measures of 50 replicates with the standard error bars. ML, multinormial logistic; DM, dirichlet multinomial; LN, logistic normal; LNM, logistic normal multinomial. This figure appears in color in the electronic version of this article.

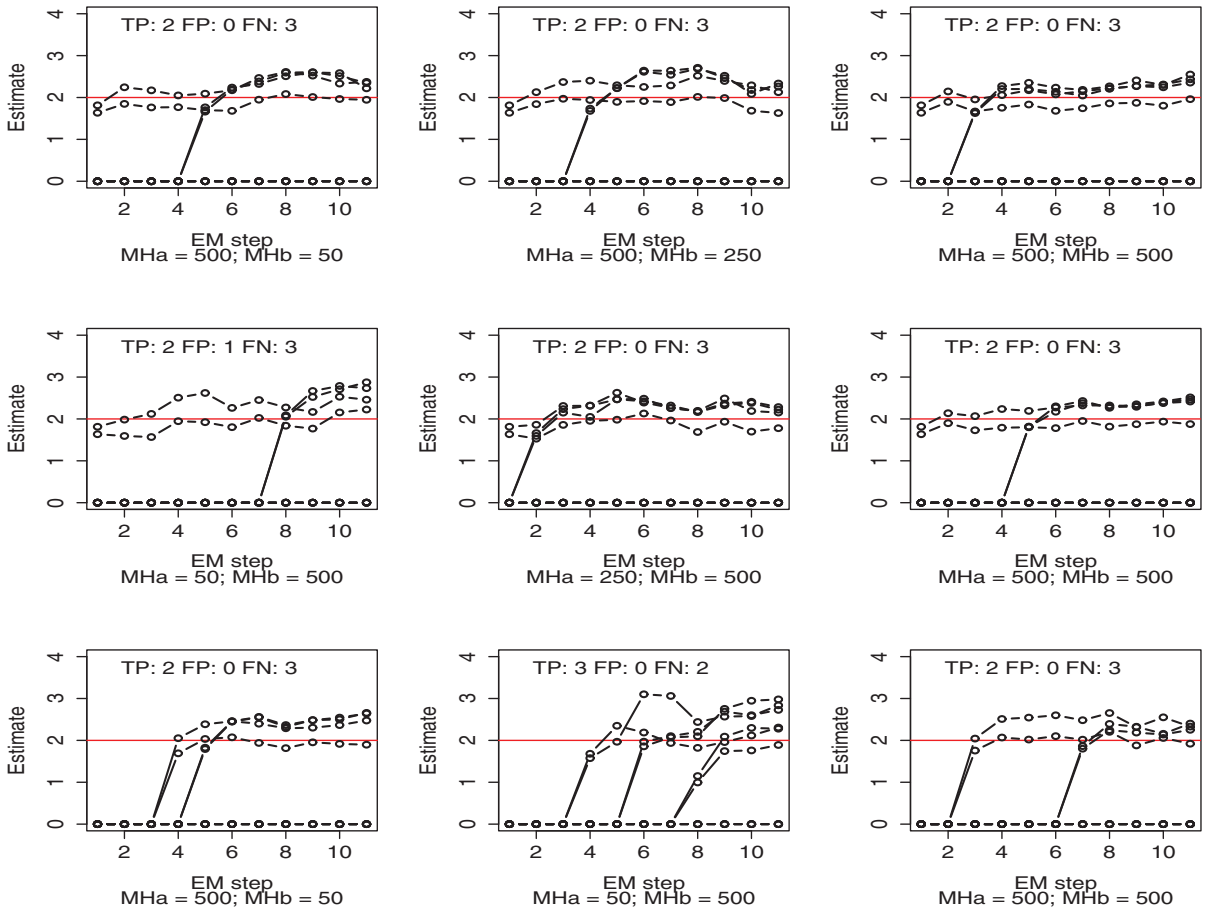

Figure 2. Simulation study to evaluate sensitivity to the HM samples and convergence of the MCEM. Top: effect of the number of burn-in (MHb = 50, 250, 500 from left to right) in the MCEM algorithm when the number of MH samples used in the E-step is fixed at 500; middle: effect of the number of MH samples used in the E-step (MHa = 50, 250, 500 from left to right) in the MCEM algorithm when the number of burn-in is fixed at 500; bottom: parameter estimates when zeros are used as the initial values.

parameter $\lambda = 36$. To obtain the averaged acceptance ratio around 30–40%, the MH step size is set to 1.5. Figure 2 shows the estimates of the parameters of the MCEM algorithm using different numbers of the burn-in and different numbers of the MH runs in approximating the expected values in the EM algorithm. We observe that the parameter estimates stabilize within a few EM iterations and the estimates are not very sensitive to the number of the burn-ins or the number of the MH samples used in the E-step. Similar results were also observed for other models we considered. Finally, we also exam the sensitivity of the parameter values when the initial β are set to different values. The parameter estimates are not sensitive to the initial values (see Figure 2).

5. Associating Nutrients With the Human Gut Microbiome Composition

We applied the LNM model to a cross-sectional study of the association between diet and stool microbiome composition (Wu et al., 2011). The goal of the study was to investigate the effects of the long-term diet on human gut microbiome composition. In this study, stool samples from 98 healthy volunteers were collected and the DNA samples were analyzed by the 454/Roche pyrosequencing of 16S rRNA gene segments of the V1–V2 region. The pyrosequences were analyzed by the QIIME pipeline (Caporaso et al., 2010) where 3608 OTUs were defined. These OTUs can be further combined into 11 phyla and 127 genera. In addition, information on habitual long-term diet of these 98 subjects was also collected using the food frequency questionnaire, which provided quantitative measurements of 214 micro-nutrients. The micro-nutrient measurements were highly correlated and were further grouped together if the correlation coefficient was greater than 0.90. For each nutrient group, one representative nutrient was selected, leading to a total of $p = 117$ covariates in our analysis. Grouping these micro-nutrients also improves the interpretability of the results. For example, we grouped potassium and potassium without supplement as a potassium group, and proanthocyanidin trimers, proanthocyanidin 4–6mers and proanthocyanidin, 7–10mers together as a proanthocyanidin group.

We focused on studying the effects of the micro-nutrients on the composition of three bacterial genera, Bacteroides, Prevetella, and Ruminococcus. These three bacterial genera were shown by Arumugam et al. (2011) to cluster individual

Table 1
Additive logistic transformed estimates of the LNM model applied to analyze the composition of three genera that define the microbiome enterotype

	Bacteroides	Prevotella	Ruminococcus	Magnitude
Acrylamide	0.144	0.756	0.100	1.525
Palmitelaidic trans fatty acid	0.480	0.078	0.442	1.456
Maltose	0.125	0.735	0.141	1.401
Vitamin C	0.556	0.135	0.309	1.004
Beta cryptoxanthin	0.179	0.581	0.240	0.867
Added germ from wheats[a,b]	0.186	0.572	0.242	0.832
Sucrose[a]	0.205	0.394	0.401	0.542
Vitamin E, food fortification	0.251	0.486	0.263	0.522
Proline[a,b]	0.362	0.245	0.393	0.359
Total choline[b]	0.352	0.248	0.400	0.352

For each selected nutrient, $\phi^{-1}(\boldsymbol{\beta}_{l\cdot})$ and $||\phi^{-1}(\boldsymbol{\beta}_{l\cdot})||_2$ are presented based on the estimated regression coefficients. Nutrients that are marked with [a] and [b] are also selected by the ML model and the DM model, respectively.

samples into three clusters, and were therefore used to defined so-called enterotypes. We similarly observed such clusters in our data set using a multi-dimensional scaling and cluster analysis (Wu et al., 2011). Among these three genera, the Bacteroides was observed in all 98 samples with high abundance, Prevetella was observed in only 36 samples, and Ruminococcus was observed in 73 samples. Among the samples that have Bacteroides, the count ranged from 216 to 11,890 with a mean count of 3687. For the 36 samples with Prevotella, the count ranged from 1 to 9623 with a mean count of 1786. For the 73 samples with Ruminococcus, the count ranged from 2 to 728 with a mean count of 109. Clearly, these counts varied greatly from sample to sample and from genus to genus.

We applied the group ℓ_1 penalized LNM regression to this data set taking Prevetella as the base component in the log-ratio transformation. After initial burn-ins, 2000 MH samples were used in each of the E-steps. For each tuning parameter, the parameter estimates stabilized within 10 EM iterations. We used fivefold cross validation to select the final tuning parameter. A total of 10 micro-nutrients were selected in the final model (see Table 1 for the names of these nutrients). To assess the effects of these nutrients on the enterotype, we applied the additive logistic transformation ϕ^{-1} to the estimates of the non-zero regression coefficients to obtain the estimates of the perturbations from the selected nutrients and present the results in Table 1. These results largely agree with what were reported in Wu et al. (2011) using a simple Spearman's correlation analysis. For example, we observed that Prevotella was positively associated while Bacteroides was negatively associated with maltose, sucrose, added germ from wheats and vitamin E/food fortification. In contrast, inverse associations were observed for fats and amino acid/choline nutrients including palmitelaidic trans fatty acid, proline and total choline, although the magnitudes of the effects of proline and total choline were not too large.

In addition to the associated nutrients mentioned above, the LNM model identified three other nutrients that are worth further investigation because of their large perturbation effects on the composition of these three genera. We observed that acrylamide and beta cryptoxanthin were positively associated with Prevotella, while vitamin C was positively associated with Bacteroides. Acrylamide is a natural byproduct of carbohydrate-rich foods and it significantly increases the level of Prevotella. Acrylamide forms from sugars and an amino acid (asparagine) during certain types of high-temperature cooking, such as frying, roasting, and baking. Beta cryptoxanthin and vitamin C are the most common nutrients in gut healing supplements (Claesson et al., 2012).

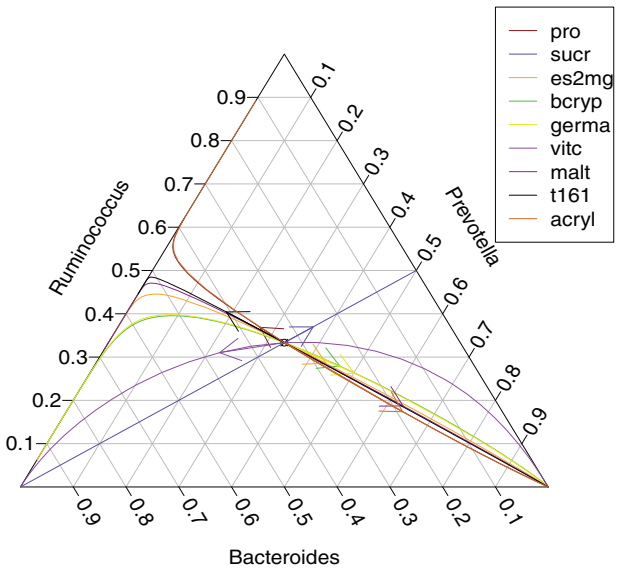

Figure 3. Ternary diagram of perturbations from nine nutrients identified by the LNM model. The curves shown are $\phi^{-1}(\mathbf{J}_2) \oplus \left\{\phi^{-1}(\beta_{l\cdot})\right\}^{\triangle x_l}$ for the lth nutrient with $-50 \leq \triangle x_p \leq 50$ denoting the units of changes in nutrient l. The arrows indicate direction and magnitude as x increase by 2 units. For the purpose of clear demonstration, we assume the baseline composition for the three enterotypes is at the center of the simplex. pro, proline; sucr, sucrose; es2mg, vitamin E/food fortification; bcryp, beta cryptoxanthin; germa, added germ from wheats; vitc, vitamin C; malt,maltose; t161, palmitelaidic trans fatty acid; acryl, acrylamide. This figure appears in color in the electronic version of this article.

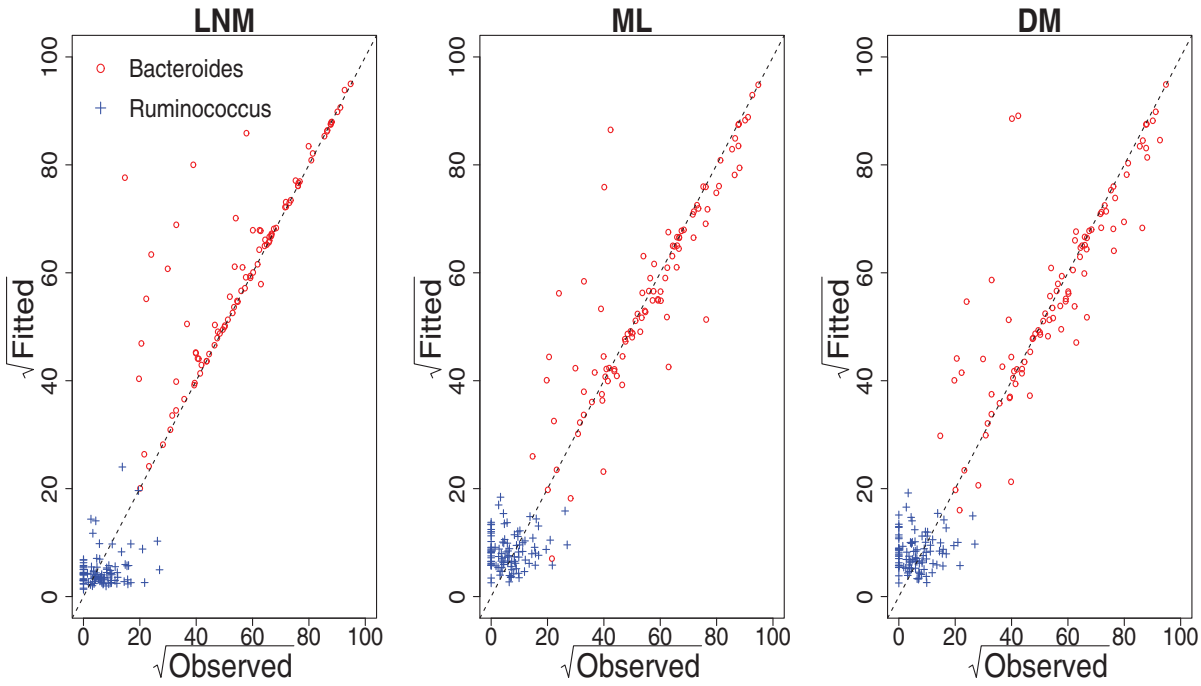

Figure 4. Comparison of model fits for the gut enterotype data, fitted versus observed counts for Bacteroides and Ruminococcus. LNM, logistic normal multinomial regression model; ML, multinomial logistic regression model; DM, Dirichlet-multinomial regression model. This figure appears in color in the electronic version of this article.

To further demonstrate the effects of nutrients on gut enterotype, we used a ternary diagram shown in Figure 3 to visualize the perturbation pattern from nutrients by the regression curves on simplex. The ternary diagram has long been used for displaying the compositional data as a plotting region for three-part compositions (e.g., proportions of Bacteroides, Prevetella, and Ruminococcus). It may be viewed as the plane in the positive orthant satisfying the summation constraint of a three-part composition (Aitchison, 1986). For the purpose of clear demonstration, we assumed the baseline composition for the three enterotypes is at the center of the simplex and only the first nine nutrients with a large magnitude of perturbation were plotted. This plot clearly shows that the composition of Ruminococcus is quite stable with respect to changes in nutrients, while the relative abundances of Bacteroides and Prevotella change dramatically according to the changes of micro-nutrients.

As a comparison, we also applied the ML model and DM model to this data set. The ML model selected five nutrients, sucrose, proline, added germ from wheat, cholesterol, and vitamin B12, while the DM model selected six nutrients, including sucrose, proline, added germ from wheat, choline (phosphatidylcholine), cholesterol, and methionine. Except for cholesterol, vitamin B12, and methionine, the LNM model selected all these nutrients. However, the three nutrients that perturbed the enterotype composition the most, acrylamide, maltose and palmitelaidic trans fatty acid were not selected by either ML or DM model. Figure 4 shows the fitted counts of Bacteroides and Ruminococcus by the LMN, ML, and DM models and the observed counts. All three models fit the majority of the data well, except for about 10 samples. LMN model seems to fit the Bacteroides counts tighter. For the very low abundant Ruminococcus, it is difficult to obtain precise fits from any of these three models due to small numbers of the counts observed.

Finally, to assess the sensitivity of the results to the correlation cutoff used for grouping the nutrients, we also performed a similar analysis by grouping the micro-nutrients into $p = 140$ nutrient groups using a correlation cutoff of 0.95. The results were consistent with those when the correlation cutoff of 0.90 was used. The LNM model selected seven nutrient groups, six of these were also selected as before when the correlation cutoff of 0.90 was used, including palmitelaidic trans fatty acid, acrylamide, vitamin C, maltose, sucrose, added gem from wheats. The LNM also selected the nutrient group betaine and choline, which overlaps with the nutrient group of total choline selected by LNM when the correlation cutoff of 0.90 was used.

6. Discussion

In this article, we have proposed a group ℓ_1 penalized estimation for the LNM regression model to select covariates associated with bacterial composition. With the coefficients corresponding to each of the covariates being treated as a group, the model combines information from the taxa data to select the relevant covariates that perturb the overall location of bacterial composition on simplex. Compared to the commonly used Dirichlet-multinomial regression model for count data, the LNM model provides a more flexible way

of modeling the dependency of the bacterial composition. One particular reason that we chose the LNM model is that it allows for the rich structure of the logistic normal distribution to describe inter-bacterial covariances. More generally, our modeling framework may be viewed as a random effects model for multinomial data. This approach allows extra-multinomial variability that is typical of data in microbiome studies. We estimate the model using a MCEM algorithm with a penalized estimation in the M-step. Extensive simulations have been conducted to compare the proposed model to other regression models to select the relevant covariates. The results have shown that the proposed method has an overall better variable selection performance. We have applied the proposed model to analyze human gut microbiome data and identified nutrients that are associated with the composition of three genera that define the human gut microbiome enterotype.

The excessive number of zero values can lead to some difficulty in analyzing microbiome compositional data. These zeros can represent either components that are truly absent from the community (called structured zeros), or rare components that are not present in the sample drawn from the community. Without additional knowledge, these two possibilities are indistinguishable. Since the multinomial distribution allows zero counts, our LMN model can handle zero observations, as we have seen in the analysis of the gut microbiome data. Depending on the goal of the analysis, the researcher should decide how to interpret these zero counts and choose analysis methods accordingly. In our analysis of the microbiome data, since only three genera were considered, the problem of having zero observations is not as severe. However, when hundreds or thousands of bacterial taxa are considered, we expect to observe excessive zeros in the data set since many bacterial taxa are very rare and sparse. Although the proposed MCEM algorithm can in principal be applied to analyze such data sets, it is computationally more challenging. In addition, how to better handle such sparse compositional data requires further studies. To deal with a large number of taxa considered, one possible solution is to consider a sparse group ℓ_1 penalty function that induces sparsity in an overall regression coefficient matrix. This approach has been employed for sparse Dirichlet multinomial regression in Chen and Li (2013).

7. Supplementary Materials

We have implemented the method as a R package "PenLNM_1.0," which is now available on CRAN (http://cran.r-project.org/web/packages/PenLNM/index.html). The zip file PenLNM_1.1.tr.gz is available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

This research is supported by NIH grants CA127334 and GM097505 and Hong Kong RGC Research Grant (766511M). We thank the reviewers and the AE for many very helpful comments.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B* **44**, 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, UK: Chapman & Hall, Ltd.

here is the complete list of all authors:

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J.; MetaHIT Consortium, Antolín, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D., and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature* **4**, 550–553.

Billheimer, D., Guttorp, P., and Fagan, W. F. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association* **96**, 1205–1214.

Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E., et al. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336.

Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research* **20**, 947–59.

Chen, J. and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Annals of Applied Statistics*, **7**, 418–442.

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., et al. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature* **11319**, in press.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., et al. (2009). The ribosomal database project: Improved alignments and new tools for rrna analysis. *Nucleic Acids Research* **37**, 141–145.

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, et al. (2012). Experimental and analytical tools for studying the human microbiome. *Nature Review Genetics* **13**, 47–58.

Meier, L., van de Geer, S., and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of The Royal Statistical Society Series B* **70**, 53–71.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., et al. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics* **4**, 53–77.

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214.

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

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Received August 2012. *Revised June* 2013. *Accepted June* 2013.