Published in final edited form as:

Biometrics. 2017 September ; 73(3): 792–801. doi:10.1111/biom.12654.

A Dirichlet-Tree Multinomial Regression Model for Associating Dietary Nutrients with Gut Microorganisms

Tao Wang^{1,3,*} and Hongyu Zhao^{2,3}

¹Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China

²Department of Biostatistics, Yale University, New Haven, Connecticut, U.S.A

³SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai, China

Summary

Understanding the factors that alter the composition of the human microbiota may help personalized healthcare strategies and therapeutic drug targets. In many sequencing studies, microbial communities are characterized by a list of taxa, their counts, and their evolutionary relationships represented by a phylogenetic tree. In this article, we consider an extension of the Dirichlet multinomial distribution, called the Dirichlet-tree multinomial distribution, for multivariate, over-dispersed, and tree-structured count data. To address the relationships between these counts and a set of covariates, we propose the Dirichlet-tree multinomial regression model for which we develop a penalized likelihood method for estimating parameters and selecting covariates. For efficient optimization, we adopt the accelerated proximal gradient approach. Simulation studies are presented to demonstrate the good performance of the proposed procedure. An analysis of a data set relating dietary nutrients with bacterial counts is used to show that the incorporation of the tree structure into the model helps increase the prediction power.

Keywords

Dirichlet distributions; Over-dispersion; Sparse group lasso; Tree-structured learning

1. Introduction

Next generation sequencing of DNA extracted from microbial communities has advanced our knowledge of the role of microbiota in human health and disease (Cho and Blaser, 2012). Researchers have found that changes in the human microbiota are associated with a number of pathological states such as obesity and inflammatory bowel disease (Clemente et al., 2012). It has also been found that microbiota populations are sensitive to genetic and environmental influences (Spor et al., 2011). Despite these progresses, we are still at the very beginning of fully understanding the factors regulating our microbiota and the impacts of microbiota on our health. In particular, our statistical methods for extracting meaningful

* neowangtao@sjtu.edu.cn.

Supplementary Materials

Web Appendices A and B referenced in Sections 5 and 6, simulated data, and computer code are available with this article at the *Biometrics* website on Wiley Online Library.

information from microbiome studies have not been developed as quickly as experimental techniques.

To process raw data from either targeted amplicon or metagenomic sequencing studies (Kuczynski et al., 2011), Operational Taxonomic Units (OTUs), which are microbial genomic sequences clustered by sequence similarity, are commonly used to partition sequences into discrete groups in place of traditional taxonomic units. For each OTU, its abundance or count is defined as the number of sequences belonging to this OTU. Typically, the most abundant sequence in an OTU is chosen as the representative sequence, and representative sequences from all the OTUs are used to construct a phylogenetic tree among all the OTUs, and/or to map OTUs to a taxonomic reference database (Schloss et al., 2009; Navas-Molina et al., 2013). After these processing steps, a microbial community can be characterized by a list of OTUs and their counts, their phylogenetic relationships, and/or their taxonomy.

The data that motivated the statistical method in this article were part of a human gut microbiome study conducted at the University of Pennsylvania (Wu et al., 2011), where the investigators aimed to assess the effect of diet on gut microbiome composition. For this study, both gut microbiome data and nutrient intake data were available (see Section 6 for details). Using a distance-based analysis procedure, the authors identified a few gut microbiome-associated nutrients that are biologically interpretable. As noted by Chen and Li (2013), however, this method is based on distances between microbiome samples, and thus is unable to provide information on how dietary nutrients affect bacterial taxa. In order to identify key nutrients as well as the taxa they affect, Chen and Li (2013) adopted a regression-based approach, where they treated OTU-abundance data as multivariate count responses, and nutrients as covariates.

Let $Y = (Y_1, \dots, Y_K)^T$ denote the random vector of counts on K bacterial taxa or OTUs. The simplest distribution for multivariate count data is the multinomial distribution. Suppose $y = (y_1, \dots, y_K)^T$ is a draw from this distribution. Given $\sum_{k=1}^K Y_k = \sum_{k=1}^K y_k$, the probability mass of Y at y is

$$f_M(y; p) = \frac{\Gamma\left(\sum_{k=1}^K y_k + 1\right)}{\prod_{k=1}^K \Gamma(y_k + 1)} \prod_{k=1}^K p_k^{y_k},$$

where $\Gamma(\cdot)$ is the gamma function, and $p = (p_1, \dots, p_K)^T$ is the vector of probabilities such that $\sum_{j=1}^K p_j = 1$. In the multinomial regression model, p is linked to the covariates via the multinomial-Poisson transformation.

Bacterial count data in microbiome studies are usually over-dispersed. The multinomial distribution is inappropriate, however, when more variation is observed than expected. In order to overcome this problem, Chen and Li (2013) assumed that p is random with some prior distribution. Denote by $\mathcal{S}^d$ the $(d-1)$ -dimensional simplex. Since the support of p is

$\mathcal{S}^{K-1}$, they used the Dirichlet distribution whose density at $u = (u_1, \dots, u_K) \in \mathcal{S}^{K-1}$, parameterized by K positive values $\alpha_1 > 0, \dots, \alpha_K > 0$, is given by

$$f_D(u; \alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K u_k^{\alpha_k - 1},$$

where $\alpha = (\alpha_1, \dots, \alpha_K)^\top$. The Dirichlet distribution is popular mainly because, as a prior distribution, it is conjugate to the multinomial distribution, leading to the Dirichlet multinomial distribution (Mosimann, 1962), also known as the Dirichlet compound multinomial distribution, with probability mass function

$$f_{DM}(y; \alpha) = \int_{u \in \mathcal{S}^{K-1}} f_M(y; u) f_D(u; \alpha) du = \frac{\Gamma\left(\sum_{k=1}^K y_k + 1\right) \Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K y_k + \sum_{k=1}^K \alpha_k\right)} \prod_{k=1}^K \frac{\Gamma(y_k + \alpha_k)}{\Gamma(y_k + 1) \Gamma(\alpha_k)}.$$

Using the Dirichlet multinomial distribution, Chen and Li (2013) considered a Dirichlet-multinomial regression model, and applied it to the analysis of gut microbiome and nutrient intake data.

The Dirichlet (multinomial) distribution has three limitations. First, all components must share a common variance parameter. Second, the components are mutually independent, up to the constraint that they must sum up to one (Mosimann, 1962). Finally, the distribution fails to take into account the special and inherent property of microbiome count data, namely, microbial taxa are not independent of each other, but are related evolutionarily in a phylogenetic tree. To address these limitations, we consider a generalization of the Dirichlet (multinomial) distribution. Assuming that the relationships among components of the count vector can be represented as a tree, the generalized distribution encodes the tree structure node-by-node as detailed in the next section. By construction, each component has an independent variance, and components are correlated at subtree levels. Based on this distribution, we propose a regression model to include the effects of covariates, and develop a regularized method for selecting covariates (e.g., nutrients) that are associated with the count responses (e.g., OTUs).

To circumvent the first two issues with the Dirichlet (multinomial) distribution, Billheimer et al. (2001) proposed an alternative distribution for over-dispersed multivariate count data, called the logistic normal multinomial distribution, by replacing the Dirichlet distribution for p with Aitchison's logistic normal distribution (Aitchison, 1982). The corresponding regression model, which relates covariates to the count vector, was studied by Billheimer et al. (2001), and was recently applied to link dietary nutrients with bacterial counts by Xia et al. (2013). However, like the Dirichlet multinomial distribution, the logistic normal multinomial distribution cannot exploit the tree structure information, and unlike the Dirichlet multinomial distribution and the generalized version considered here, its density does not have a closed-form expression, making the regression estimation computationally intensive.

The rest of the article is organized as follows. We review the Dirichlet-tree multinomial distribution and its properties in Section 2. We then propose the Dirichlet-tree multinomial regression model in Section 3. Regularized likelihood estimation, including the algorithm and tuning, are considered in Section 4. Simulation studies and an application to a human gut microbiome data set, respectively in Sections 5 and 6, are used to evaluate the performance of our method. Finally, Section 7 contains some remarks.

2. Dirichlet-Tree Multinomial Distributions

Data on bacterial community structure can be compiled in a matrix of counts, where samples are represented as rows and taxa as columns. As in Chen and Li (2013) and Xia et al. (2013), we treat the total number of counts, which is determined by the sequencing depth, as an ancillary statistic, and conduct the analysis conditioning on this number. Now suppose we have available a tree T representing the hierarchical structure over the count responses (e.g., a phylogenetic tree of bacterial taxa). We seek to incorporate this structural information into the modeling process.

Denote by $\mathcal{L}$ the set of leaf nodes and $\mathcal{V}$ the set of internal nodes of T , respectively. For each internal node $v \in \mathcal{V}$, let $\mathcal{C}_v$ be the set of child nodes of v . For each v and $c \in \mathcal{C}_v$, define $\delta_{vc}(l)$ to be 1, if the branch from v to c leads to $l \in \mathcal{L}$, and 0, otherwise. For simplicity, write $\mathcal{L} = \{1, \dots, K\}$. Denote by $y_{vc} = \sum_{l \in \mathcal{L}} \delta_{vc}(l) y_l$ and $p_{vc} = \sum_{l \in \mathcal{L}} \delta_{vc}(l) p_l$ respectively, the count and probability in the subtree indexed by $c \in \mathcal{C}_v$.

Let v_0 be the root node of T . For each leaf node $l \in \mathcal{L}$, denote by

$$v_0 \equiv v_0^l \rightarrow v_1^l \rightarrow \dots \rightarrow v_{D_l-1}^l \rightarrow v_{D_l}^l \equiv l$$

the path from v_0 to l , where $v_d^l \in \mathcal{V}$ for $d = 0, \dots, D_l - 1$, and $D_l - 1$ is the number of branches in this path. Let $b_{vc} = p_{vc} / \sum_{s \in \mathcal{C}_v} p_{vs}$. Then $\sum_{c \in \mathcal{C}_v} b_{vc} = 1$, hence each branch of T is assigned a probability. A simple calculation shows that

$$p_l = b_{v_0 v_1^l} \times b_{v_1^l v_2^l} \times \dots \times b_{v_{D_l-1}^l v_{D_l}^l} = \prod_{v \in \mathcal{V}} \prod_{c \in \mathcal{C}_v} b_{vc}^{\delta_{vc}(l)}.$$

That is, p_l is the product of branch probabilities as we traverse from v_0 to l . See Figure 1 for an illustration. Furthermore, we can write

$$f_M(y; p) = f_M(y; b_v, v \in \mathcal{V}) = \prod_{v \in \mathcal{V}} \frac{\Gamma(\sum_{c \in \mathcal{C}_v} y_{vc} + 1)}{\prod_{c \in \mathcal{C}_v} \Gamma(y_{vc} + 1)} \prod_{c \in \mathcal{C}_v} b_{vc}^{y_{vc}},$$

where $b_v = \{b_{vc}, c \in \mathcal{C}_v\}$.

Instead of assuming one Dirichlet distribution for p , we can specify a separate Dirichlet distribution for b_v , for each internal node $v \in \mathcal{V}$. Specifically, the joint density function of $(b_v, v \in \mathcal{V})$ has the form

$$\prod_{v \in \mathcal{V}} f_D(u_v; \alpha_v) = \prod_{v \in \mathcal{V}} \frac{\Gamma(\sum_{c \in \mathcal{C}_v} \alpha_{vc})}{\prod_{c \in \mathcal{C}_v} \Gamma(\alpha_{vc})} \prod_{c \in \mathcal{C}_v} u_{vc}^{\alpha_{vc}-1},$$

where $u_v = (u_{vc}; c \in \mathcal{C}_v) \in \mathcal{S}^{K_v-1}$, K_v is the number of children of v , and $\alpha_v = (\alpha_{vc} > 0, c \in \mathcal{C}_v)$ is a vector of positive scalars. This gives the Dirichlet-tree multinomial distribution (Dennis, 1996; Minka, 2004), which is a compound multinomial distribution with probability mass function

$$\begin{aligned} f_{\text{DTM}}(y; \alpha_v, v \in \mathcal{V}) &= \prod_{v \in \mathcal{V}} \int_{u_v \in \mathcal{S}^{K_v-1}} f_M(y; u_v, v \in \mathcal{V}) f_D(u_v; \alpha_v) du_v \\ &= \prod_{v \in \mathcal{V}} \frac{\Gamma(\sum_{c \in \mathcal{C}_v} y_{vc} + 1) \Gamma(\sum_{c \in \mathcal{C}_v} \alpha_{vc})}{\Gamma(\sum_{c \in \mathcal{C}_v} y_{vc} + \sum_{c \in \mathcal{C}_v} \alpha_{vc})} \\ &\quad \times \prod_{c \in \mathcal{C}_v} \frac{\Gamma(y_{vc} + \alpha_{vc})}{\Gamma(y_{vc} + 1) \Gamma(\alpha_{vc})}. \end{aligned} \quad (1)$$

Note that each component in the product $\prod_{v \in \mathcal{V}}$ corresponds to an internal node in the tree, which is described by a Dirichlet multinomial distribution based on the accumulated counts along branches of that node. In the special case where v_0 is the only internal node, that is, $\mathcal{V} = \{v_0\}$, this distribution reduces to the Dirichlet multinomial distribution.

It is not difficult to show that

$$E(p_l) = \prod_{v \in \mathcal{V}} \prod_{c \in \mathcal{C}_v} \{E(b_{vc})\}^{\delta_{vc}(l)} = \prod_{v \in \mathcal{V}} \prod_{c \in \mathcal{C}_v} \left(\frac{\alpha_{vc}}{\sum_{c \in \mathcal{C}_v} \alpha_{vc}} \right)^{\delta_{vc}(l)}.$$

Thus, given $\sum_{k=1}^K Y_k = \sum_{k=1}^K y_k$, $E(Y_l) = E(p_l) \sum_{k=1}^K y_k$. Other moments of the Dirichlet-tree distribution can be found in Dennis (1991). In particular, Dennis showed that when $\mathcal{V} = \{v_0\}$, each component p_l has an independent variance, and two components corresponding to the subtree indexed by $v \in \mathcal{V}$ are correlated, due to their dependence on the ancestors of v .

3. A Dirichlet-Tree Multinomial Regression Model

The Dirichlet-tree distribution is often used as a prior in a Bayesian analysis (see, e.g., Tam and Schultz, 2007; Haffari and Teh, 2009). On the other hand, to our knowledge the Dirichlet-tree (multinomial) distribution is not thought of as a response distribution, while in practice it is frequent to find multivariate count data (such as counts of bacterial taxa) with observed covariates (such as measurements of dietary nutrients). In this section, we use the Dirichlet-tree multinomial distribution to develop a model for regressing the count vector Y on p covariates $X_1, \dots, X_p$.

To model the effects of covariates, for each $v \in \mathcal{V}$ and $c \in \mathcal{C}_v$ we express $\log(\alpha_{vc})$ as a linear combination of the covariates. Specifically, we assume that

$$\log(\alpha_{vc}) = X^\top \beta_{vc}, \quad (2)$$

where $X = (X_1, \dots, X_p)^\top$, and $\beta_{vc} = (\beta_{vc1}, \dots, \beta_{vcp})^\top$ is the vector of coefficients.

Let $\beta_v = (\beta_{vc}, c \in \mathcal{C}_v)$ and $\beta = (\beta_v, v \in \mathcal{V})$. The standard way of estimating β is to use maximum likelihood. Let $y_i = (y_{i1}, \dots, y_{iK})^\top$ denote the observed vector of counts and $x_i = (x_{i1}, \dots, x_{ip})^\top$ the observed vector of covariates, for $i = 1, \dots, n$. Subject to a constant, the log-likelihood function for the observed data is

$$l_{\text{DTM}}(\beta) = \sum_{v \in \mathcal{V}} l_v(\beta_v) \quad (3)$$

with

$$l_v(\beta_v) = \sum_{i=1}^n \left[\tilde{\Gamma} \left(\sum_{c \in \mathcal{C}_v} \alpha_{ivc} \right) - \tilde{\Gamma} \left(\sum_{c \in \mathcal{C}_v} y_{ivc} + \sum_{c \in \mathcal{C}_v} \alpha_{ivc} \right) + \sum_{c \in \mathcal{C}_v} \{ \tilde{\Gamma}(y_{ivc} + \alpha_{ivc}) - \tilde{\Gamma}(\alpha_{ivc}) \} \right],$$

where $\tilde{\Gamma}(\cdot) = \log\{\Gamma(\cdot)\}$, $\alpha_{ivc} = \exp(x_i^\top \beta_{vc})$, and $y_{ivc} = \sum_{j \in \mathcal{L}} \delta_{vc}(j) y_{ij}$.

4. Regularized Likelihood Estimation

One can show that the number of parameters in the Dirichlet-tree multinomial regression model equals the number of covariates, p , times the number of branches, $\sum_{v \in \mathcal{V}} K_v$, and thus it increases rapidly both as the dimension p grows and as the tree T expands. This makes the maximum likelihood estimation unappealing in terms of both accuracy and interpretability. In order to obtain a reliable and interpretable model, we propose a penalized (negative) log-likelihood method that estimates parameters and selects covariates simultaneously. Specifically, we consider the following objective function

$$pl_{\text{DTM}}(\beta) \equiv pl_{\text{DTM}}(\beta; \lambda_1, \lambda_2) = -l_{\text{DTM}}(\beta) + \lambda_1 \sum_{v \in \mathcal{V}} \sum_{c \in \mathcal{C}_v} \|\beta_{vc}\|_1 + \lambda_2 \sum_{v \in \mathcal{V}} \sum_{c \in \mathcal{C}_v} \|\beta_{vc}\|_2, \quad (4)$$

where λ_1 and λ_2 are tuning parameters, and $\|\cdot\|_1$ and $\|\cdot\|_2$ are the usual l_1 and l_2 norms.

4.1. Algorithm

Since the penalty function is non-smooth, we adopt the accelerated proximal gradient method (see, e.g., Beck and Teboulle, 2009) to minimize $pl_{\text{DTM}}(\beta)$. The algorithm alternates between two sequences $\{\beta^{(t)}\}$ and $\{\eta^{(t)}\}$, where t is the index of iteration. At the $(t+1)$ -th iteration, we approximate $-l_{\text{DTM}}(\beta)$ at $\eta^{(t)}$ by

$$-\tilde{l}_{\text{DTM}}(\beta; \eta^{(t)}) = -l_{\text{DTM}}(\eta^{(t)}) - \langle \beta - \eta^{(t)}, \nabla l_{\text{DTM}}(\eta^{(t)}) \rangle + \frac{C}{2} \|\beta - \eta^{(t)}\|_2^2,$$

where $\langle \cdot, \cdot \rangle$ is the trace operator, $\nabla l_{\text{DTM}}(\cdot)$ is the gradient of $l_{\text{DTM}}(\cdot)$, and $C > 0$ is a constant. Let

$$p\tilde{l}_{\text{DTM}}(\beta) = -\tilde{l}_{\text{DTM}}(\beta; \eta^{(t)}) + \lambda_1 \sum_{v \in \mathcal{V}} \sum_{c \in \mathcal{C}_v} \|\beta_{vc}\|_1 + \lambda_2 \sum_{v \in \mathcal{V}} \sum_{c \in \mathcal{C}_v} \|\beta_{vc}\|_2. \quad (5)$$

We update $\beta^{(t+1)}$ by solving

$$\beta^{(t+1)} = \arg \min_{\beta} \{p\tilde{l}_{\text{DTM}}(\beta)\}. \quad (6)$$

We then update $\eta^{(t+1)}$ as

$$\eta^{(t+1)} = \beta^{(t+1)} + \frac{1 - a_t}{a_t} a_{t+1} (\beta^{(t+1)} - \beta^{(t)}). \quad (7)$$

Here, $a_t = 2/(t+2)$. A simple calculation shows that the optimization problem (6) can be decomposed into a set of subproblems

$$\beta_{vc}^{(t+1)} = \arg \min_{\beta_{vc}} \left[\frac{1}{2} \left\| \beta_{vc} - \left\{ \eta_{vc}^{(t)} - \frac{1}{C} \nabla l_{vc}(\eta_v^{(t)}) \right\} \right\|_2^2 + \frac{\lambda_1}{C} \|\beta_{vc}\|_1 + \frac{\lambda_2}{C} \|\beta_{vc}\|_2 \right]$$

for $v \in \mathcal{V}$ and $c \in \mathcal{C}_v$, where $\nabla l_{vc}(\eta_v)$ is the gradient of $l_v(\eta_v)$ with respect to η_{vc} . Let

$$\tilde{\beta}_{vc}^{(t+1)} = \text{sgn} \left\{ \eta_{vc}^{(t)} - \frac{1}{C} \nabla l_{vc}(\eta_v^{(t)}) \right\} \max \left\{ 0, \left| \eta_{vc}^{(t)} - \frac{1}{C} \nabla l_{vc}(\eta_v^{(t)}) \right| - \frac{\lambda_1}{C} \right\},$$

where sgn is the sign function, and the equality holds element-wise. The solution of $\beta_{vc}^{(t+1)}$ is easily shown to be

$$\beta_{vc}^{(t+1)} = \frac{\tilde{\beta}_{vc}^{(t+1)}}{\|\tilde{\beta}_{vc}^{(t+1)}\|_2} \max \left\{ 0, \|\tilde{\beta}_{vc}^{(t+1)}\|_2 - \frac{\lambda_2}{C} \right\}. \quad (8)$$

Instead of using a constant step size C , we can use the backtracking rule to choose a suitable C at each iteration. In summary, the algorithm for each pair of (λ_1, λ_2) proceeds as follows.

S1 Initialize $C = \sum_{v \in \mathcal{V}} K_v \times p$, $a_0 = 1$, β_0 , and $\eta_0 = \beta_0$.

S2 Iterate until $pl_{DTM}(\beta^{(t+1)}) < \tilde{pl}_{DTM}(\beta^{(t+1)}; \eta^{(t)})$.

S2.1 For all $v \in \mathcal{V}$ and $c \in \mathcal{C}_v$, compute $\nabla l_{vc}(\eta_v^{(t)})$ and

$$\tilde{\beta}_{vc}^{(t+1)} = \text{sgn} \left\{ \eta_{vc}^{(t)} - \frac{1}{C} \nabla l_{vc}(\eta_v^{(t)}) \right\} \times \max \left\{ 0, \left| \eta_{vc}^{(t)} - \frac{1}{C} \nabla l_{vc}(\eta_v^{(t)}) \right| - \frac{\lambda_1}{C} \right\}.$$

S2.2 Update

$$\beta_{vc}^{(t+1)} = \frac{\tilde{\beta}_{vc}^{(t+1)}}{\|\tilde{\beta}_{vc}^{(t+1)}\|_2} \max \left\{ 0, \|\tilde{\beta}_{vc}^{(t+1)}\|_2 - \frac{\lambda_2}{C} \right\}.$$

S2.3 Set C to be $2C$.

S3 Update $a_{t+1} = 2/(t+3)$ and

$$\eta^{(t+1)} = \beta^{(t+1)} + \frac{1-a_t}{a_t} a_{t+1} (\beta^{(t+1)} - \beta^{(t)}).$$

S4 If the objective value converges, stop the algorithm. Otherwise, set t to be $t+1$ and go to S2.

Three things are noteworthy. First, each of the two sequences is computed analytically, hence the algorithm is easy to implement. Second, the complexity of the algorithm scales well with both the number of covariates and the number of branches. Third, as $l_{DTM}(\beta)$ is non-convex, convergence is not guaranteed. Generally, the algorithm converges to a local minimizer, since the objective value decreases over iterations. Empirically, based on our own experience in simulations, the algorithm is fast and efficient.

4.2. Tuning and Implementation

A key issue in the above penalized likelihood estimation is the choice of appropriate values for the tuning parameters. To this end, we consider a re-parametrization of λ_1 and λ_2 as follows. Let $\lambda = \lambda_1 + \lambda_2$ and $\gamma = \lambda_2/(\lambda_1 + \lambda_2)$. We can write

$$pl_{DTM}(\beta; \lambda_1, \lambda_2) = pl_{DTM}(\beta; \lambda, \gamma) = -l_{DTM}(\beta) + \lambda \left\{ (1-\gamma) \sum_{v \in \mathcal{V}} \sum_{c \in \mathcal{C}_v} \|\beta_{vc}\|_1 + \gamma \sum_{v \in \mathcal{V}} \sum_{c \in \mathcal{C}_v} \|\beta_{vc}\|_2 \right\}.$$

Denote by $\hat{\beta}(\lambda, \gamma)$ the minimizer of $pl_{DTM}(\beta; \lambda, \gamma)$. Two special cases are the lasso estimate $\hat{\beta}(\lambda, 0)$ and the group lasso estimate $\hat{\beta}(\lambda, 1)$. When $0 < \gamma < 1$, we call $\hat{\beta}(\lambda, \gamma)$ a sparse group lasso estimate.

For each given γ , we use the Bayesian information criterion (Schwarz, 1978) to select λ . To be specific, we define

$$\text{BIC}(\lambda, \gamma) = -2 \times l_{\text{DTM}}\{\hat{\beta}(\lambda, \gamma)\} + \log(n) \times \text{df}\{\hat{\beta}(\lambda, \gamma)\},$$

where $\text{df}\{\hat{\beta}(\lambda, \gamma)\}$ is the effective degrees of freedom implied by $\hat{\beta}(\lambda, \gamma)$. In this article, we approximate $\text{df}\{\hat{\beta}(\lambda, \gamma)\}$ by the number of nonzero components in $\hat{\beta}(\lambda, \gamma)$. We then select λ by minimizing $\text{BIC}(\lambda, \gamma)$ over a set of candidate values. Note that cross validation is an alternative to BIC, but it tends to over-select spurious features (Wasserman and Roeder, 2009). Also, cross validation is computationally too intensive on large data sets.

It is now essential to specify an interval for the values of λ , for each given γ . Clearly, $\hat{\beta}(\lambda, \gamma)$ is exactly zero when λ is sufficiently large. To determine such a value, denoted by $\lambda_{\max}$, we search for the smallest λ so that zero is a minimizer of

$$\frac{1}{2} \left\| \beta_{vc} + \frac{1}{C} \nabla l_{vc}(0) \right\|_2^2 + \lambda \left(\frac{1-\gamma}{C} \|\beta_{vc}\|_1 + \frac{\gamma}{C} \|\beta_{vc}\|_2 \right)$$

for all $v \in \mathcal{V}$ and $c \in \mathcal{C}_v$. To obtain estimates over a grid of λ values, we start with $\lambda_{\max}$ for which $\hat{\beta}(\lambda, \gamma) = 0$, decrease it on a log scale, and run the proposed algorithm to compute $\hat{\beta}(\lambda, \gamma)$. Then λ is decreased again and the process is repeated. To reduce the computing time, we use the estimate from the previous λ as a warm start, and to avoid over-fitting, we stop the process until the proportion of nonzero components in $\hat{\beta}(\lambda, \gamma)$ exceeds a certain threshold (e.g., 80%).

It might be cumbersome to interpret potentially tens of thousands of parameters in $\hat{\beta}(\lambda, \gamma)$. To partly overcome this problem, as we will do in the data analysis, we can proceed as follows. Note that for each leaf node l in the tree, which represents a count response, there is a path from the root node v_0 to l . To see how a covariate (e.g., a dietary nutrient) affects this count response (e.g., a bacterial taxon), we can check whether this nutrient appears in the path from v_0 to l . In other words, we look at the problem from the viewpoint of covariate selection and take advantage of the tree topology to summarize the information in $\hat{\beta}(\lambda, \gamma)$.

5. Simulation Studies

In this section, we use simulated data to examine the performance of our method. To mimic the association study between nutrient intake and OTU abundance that will be described in Section 6, we generate $n = 75$ samples, each with $K = 28$ count responses and $p = 100$ covariates, and we use the tree in Figure 2 to represent the relationship among the responses. For simplicity, the first 10 covariates are assumed to be relevant (i.e., each of them is associated with at least one response) and the rest redundant.

We first draw the covariate vector $X = (X_1, \dots, X_p)^\top$ from a multivariate normal distribution with mean vector zero and covariance matrix $\Sigma = (0.5^{|i-j|}) \in \mathbb{R}^{p \times p}$. We then generate the count vector $Y = (Y_1, \dots, Y_K)^\top$ using the Dirichlet-tree multinomial model with parameters specified in Figure 3. Here, we choose nonzero coefficients to be of the same magnitude just for simplicity and for ease of illustration. The conclusions are qualitatively similar if they vary in magnitude.

We generated 100 data sets in this way, and for each data set, we applied our method with the tuning parameter λ selected by BIC. Throughout this section, we considered four instances of γ : 0, 0.25, 0.5, and 1. To see how each competitor selected the covariates, we calculated the true positive rate and false positive rate at each tuning parameter value λ , and summarized them by the Receiver Operating Characteristic (ROC) curve that resulted from a grid of λ values. The average ROC curves over 100 data replications are shown in Figure 4. We can see that, for this simulation setup, the sparse group lasso ($\gamma = 0.25$ or $\gamma = 0.5$) performs better than the lasso ($\gamma = 0$) and the group lasso ($\gamma = 1$).

For each data set, we further generated a separate set of size $n^* = 25$, and computed the prediction error using the following cost function

$$\sum_{i=1}^{n^*} \frac{1}{y_i^*} \sum_{k=1}^K \frac{(y_{ik}^* - \hat{y}_{ik}^*)^2}{\hat{y}_{ik}^*},$$

where $\{(y_{i1}^*, \dots, y_{iK}^*)^\top, i=1, \dots, n^*\}$ are test samples, $y_i^* = \sum_{k=1}^K y_{ik}^*$, and $\hat{y}_{ik}^*$ denotes a predicted value of y_{ik}^* . For comparison, we also evaluated the performance of the method of Chen and Li (2013), which was based on the Dirichlet multinomial model. The results are reported in Table 1. Our method had higher prediction accuracy than theirs. Furthermore, for this simulation setup, the performance of the sparse group lasso was better than that of the (group) lasso.

Finally, we increased the sample size from $n = 75$ to $n = 100$. From Figure 4 and Table 1 we see that the performance of our method improved, but the conclusions remain qualitatively similar.

Note that the Dirichlet-tree distribution reduces to the Dirichlet distribution, when the parameters in the former are chosen carefully. However, this nesting structure is lost in regression models: in the presence of predictors, the Dirichlet multinomial regression is not a sub-model of the Dirichlet tree multinomial regression. This makes the direct comparison between these two models difficult, and is the reason for our use of prediction accuracy as a measure of performance. To make feasible the comparison of ROC curves between our method and that of Chen and Li (2013), we can calculate the true positive rate and false positive rate in terms of how each method identifies the relevant covariates. For the simulation example considered here, our method was shown to be superior to Chen and Li (2013)'s; see Web Appendix A for details.

6. Data Analysis

The human gut harbors trillions of microbes that play a fundamental role in human health and disease (Clemente et al., 2012). Some recent studies have suggested that diet is an important factor that underlies the gut microbiota composition (De Filippo et al., 2010; Wu et al., 2011; David et al., 2014). One important goal of microbiome studies is thus to identify which food constituents specifically promote/inhibit growth and functionality of which microorganisms in the human gut.

The data we consider come from a cross-sectional study associating nutrient intake with the human gut microbiota composition (Wu et al., 2011). In this study, the investigators used 16S rRNA sequencing to characterize fecal samples from 98 healthy volunteers. After processing the generated sequences using the QIIME pipeline with the default parameter settings (Navas-Molina et al., 2013), more than 3000 species-level OTUs were obtained, and the microbiome data was summarized by an OTU table (i.e., a matrix of OTU counts) including the taxonomy of each OTU, and a phylogenetic tree relating these OTUs. To link long-term dietary patterns with the microbiome composition, diet information of each subject was collected using a food frequency questionnaire, and was converted into nutrient intake profiles for more than 200 micro-nutrients.

In our analysis, we combined the species-level OTUs at the genus level, and considered 28 relatively prevalent genera that occurred in more than 25 of the 98 samples. Further, we considered a subset of 97 dietary nutrients (Wu et al., 2011) that were obtained by screening the nutrients using a distance-based nonparametric test (McArdle and Anderson, 2001). Our final data consisted of an OTU-count matrix $Y \in \mathbb{R}^{98 \times 28}$, a phylogenetic tree T among 28 genus-levels OTUs, as shown in Figure 2, and a nutrient-intake matrix $X \in \mathbb{R}^{98 \times 97}$. Each nutrient was standardized to have zero mean and unit variance.

In order to evaluate the performance, we randomly split the 98 samples into balanced training and test sets. Specifically, we set approximately 75% of the observations as training samples, and the rest as test samples. Our method was first applied to the training set, and its performance was then assessed by the test set using the same cost function defined in the previous section. We also compared our method with the method of Chen and Li (2013), which was developed for the Dirichlet multinomial model and thus failed to exploit the tree information. Three values of γ were explored to favor sparse solutions, 0, 0.25, and 0.5.

To reduce the variability, the splitting into training and test sets was repeated 100 times and the error estimates were averaged. The results are summarized in Table 2. We can see that the incorporation of the tree structure helped improve the prediction accuracy.

To see how dietary nutrients affected bacterial taxa, we set $\gamma = 0.25$ and concentrated on two genera, *Bacteroides* and *Prevotella*, that were previously used to define enterotypes of the human gut microbiome (Arumugam et al., 2011; Wu et al., 2011). For each genus-nutrient pair, we checked whether the nutrient appeared in the path from the root node to the leaf node corresponding to the genus. The top 10 nutrients, based on 100 random splits of the total samples into training and test sets, are listed in Table 3. The results are largely consistent with the findings of Wu et al. (2011), where they reported that the *Bacteroides* enterotype was associated with a number of amino acids and animal protein, while the *Prevotella* enterotype was associated with a carbohydrate-based diet. Comparing our method with that of Chen and Li (2013), the two subsets of identified nutrients were different. Nevertheless, the conclusions were similar, largely because some nutrients are highly correlated. In Web Appendix B, we provide a heatmap of selection frequencies of all dietary nutrients across 100 random splits.

7. Discussion

We have considered a generalization of the Dirichlet multinomial distribution for describing multivariate, over-dispersed, and tree-structured count data, called the Dirichlet-tree multinomial distribution, and its associated regression model for relating a set of covariates to these counts. By introducing the $l_1 + l_2$ regularization, we have proposed a method for simultaneous parameter estimation and covariate selection. We have conducted simulation studies to demonstrate the efficacy of the new approach. When applied to analyze the motivating data set, our method was shown to improve the accuracy of prediction.

We have implemented the proposed method in R (R Core Team, 2014). Our experience suggests that the procedure works reasonably fast. For instance, for the simulation example in Section 5 where $n = 100$, $K = 28$, and $p = 100$, the average computing time (including tuning by BIC) was less than 1.5 minutes on the Yale Louise high performance cluster (Dell m620 system, 8-core processor, 48G of memory) using R version 3.1.0 and a single processor. Due to the incorporation of tree structure, our method is computationally more intensive than that of Chen and Li (2013). However, the complexity of the proposed algorithm increases linearly with the number of branches, and in the numerical studies our method was 2–4 times slower than Chen and Li's. We also tried the block-coordinate descent algorithm of Chen and Li (2013), and found that this algorithm was much slower than ours, and that it occasionally broke down for a range of tuning parameters. We believe our method of coupling l_1 and l_2 regularizations with the Dirichlet-tree multinomial model would provide a useful data analysis tool for associating covariates (such as dietary nutrients) with multivariate, over-dispersed, and tree-structured count data (such as taxa counts).

In the literature, several schemes have been proposed to take advantage of the tree structure over either the covariates or responses (see, for example, Zhao et al., 2009; Jenatton et al., 2011; Kim and Xing, 2012; Garcia et al., 2014). However, these methods are mainly for univariate or multivariate continuous responses, while our method is for multivariate count responses. More importantly, our way of incorporating the tree information into the model is inherently different from the imposition of tree-guided penalties on the regression parameters for achieving structured sparsity. Indeed, our proposed Dirichlet-tree multinomial model is more flexible than the Dirichlet multinomial model in the sense that it allows the rich dependence structure among the count responses.

It might be reasonable to expect that non-zero regression coefficients are similar for phylogenetically related OTUs, and in this case some tree-based fused lasso regularization is preferable to sparse group lasso regularization. But, the resulting optimization is expected to be more challenging. Also, in reality it is possible that sister taxa can have very different metabolic pathways, ecologies, and interactions with the host. A comparison of our method and a tree-based fused lasso on multiple data sets would be an interesting area for future research.

In this article, we assume that the tree structure is known a priori. In the real data example, the tree is learned from 16S rRNA sequences using a fast and approximate maximum

likelihood method (Price et al., 2010). Note that the Dirichlet-tree multinomial regression reduces to the Dirichlet multinomial regression, when the tree is a “comb” tree, that is, the root node of the tree is the only internal node (see Figure 1 for illustration). This, together with the simulation results in Section 5, implies that the correct specification of the tree structure is important. Indeed, tree mis-specification could be a major risk for shotgun sequencing data (Matsen et al., 2010). Another issue is that we exploit the topology of the phylogenetic tree, but ignores the branch lengths. Since each component I_v of the log-likelihood function (3) corresponds to one internal node v , direct incorporation of branch lengths is cumbersome. For each internal node v , the distances of its child nodes from it (i.e., branch lengths) may be summarized and then used as the weight for I_v . But, robustness against tree mis-specification could be a concern, since both tree topology and branch lengths are exploited. Also, we need to choose a summary statistic that is biologically meaningful. Alternatively, we may use a new model with new inference algorithms. These are beyond the scope of the present article.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to the Editor, the Associate Editor, and three anonymous referees for their helpful comments. We thank Hongzhe Li and Jun Chen for providing the data. This work was supported by National Natural Science Foundation of China (11601326), and NIH grant R01 GM59507.

References

- Aitchison J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B.* 1982; 44:139–177.
- Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. *Nature.* 2011; 473:174–180. [PubMed: 21508958]
- Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences.* 2009; 2:183–202.
- Billheimer D, Guttorm P, Fagan WF. Statistical interpretation of species composition. *Journal of the American statistical Association.* 2001; 96:1205–1214.
- Chen J, Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics.* 2013; 7:418–442.
- Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics.* 2012; 13:260–270.
- Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. *Cell.* 2012; 148:1258–1270. [PubMed: 22424233]
- David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature.* 2014; 505:559–563. [PubMed: 24336217]
- De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences.* 2010; 107:14691–14696.
- Dennis SY. On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Communications in Statistics-Theory and Methods.* 1991; 20:4069–4081.
- Dennis SY. A Bayesian analysis of tree-structured statistical decision problems. *Journal of Statistical Planning and Inference.* 1996; 53:323–344.

- Garcia TP, Müller S, Carroll RJ, Walzem RL. Identification of important regressor groups, subgroups and individuals via regularization methods: Application to gut microbiome data. *Bioinformatics*. 2014; 30:831–837. [PubMed: 24162467]
- Haffari, G., Teh, YW. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics; Stroudsburg, PA, USA: 2009. Hierarchical Dirichlet trees for information retrieval; p. 173–181.
- Jenatton R, Mairal J, Obozinski G, Bach F. Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research*. 2011; 12:2297–2334.
- Kim S, Xing EP. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *The Annals of Applied Statistics*. 2012; 6:1095–1117.
- Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, et al. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*. 2011; 13:47–58.
- Matsen FA, Kodner RB, Armbrust EV. pplacer: Linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*. 2010; 11:538. [PubMed: 21034504]
- McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*. 2001; 82:290–297.
- Minka, T. The Dirichlet-tree distribution. Technical report. 2004. Paper available online at: <http://msr-waypoint.com/en-us/um/people/minka/papers/dirichlet/minka-dirtree.pdf>
- Mosimann JE. On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions. *Biometrika*. 1962; 49:65–82.
- Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, et al. Advancing our understanding of the human microbiome using QIIME. *Methods in Enzymology*. 2013; 531:371–444. [PubMed: 24060131]
- Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 2010; 5:e9490. [PubMed: 20224823]
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*. 2009; 75:7537–7541. [PubMed: 19801464]
- Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978; 6:461–464.
- Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology*. 2011; 9:279–290. [PubMed: 21407244]
- Tam, Y-C., Schultz, T. *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, Volume 4, IV–41*. IEEE Piscataway; New Jersey, US: 2007. Correlated latent semantic model for unsupervised LM adaptation.
- Wasserman L, Roeder K. High dimensional variable selection. *The Annals of Statistics*. 2009; 37:2178–2201. [PubMed: 19784398]
- Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*. 2011; 334:105–108. [PubMed: 21885731]
- Xia F, Chen J, Fung WK, Li H. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*. 2013; 69:1053–1063. [PubMed: 24128059]
- Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*. 2009; 37:3468–3497.

Figure 1.
Two trees, each with four leaf nodes. Shown are branch and leaf probabilities.

Figure 2.
The phylogenetic tree. Leaves (i.e., OTUs) are labeled as 1, ..., 28, and internal nodes are labeled as 29, ..., 55 and indicated by circles.

Figure 3. Coefficients of the 10 relevant covariates used in the Dirichlet-tree multinomial model. This figure appears in color in the electronic version of this article.

Figure 4. Simulation studies. ROC curves for the four regularized estimators, all based on the Dirichlet-tree multinomial model. The colored triangles show the true positive rate (TPR) and false positive rate (FPR), when the tuning parameter λ was selected by the BIC criterion. (a) $n = 75$, and (b) $n = 100$.

Table 1

Means and standard deviations (in parentheses) of the prediction error, over 100 replications, are reported for the four regularized estimators based on the Dirichlet-tree multinomial (DTM) model, and those based on the Dirichlet multinomial (DM) model

		$n = 75$	$n = 100$
DTM	$\gamma = 0$	80 (28)	65 (24)
	$\gamma = 0.25$	77 (24)	63 (20)
	$\gamma = 0.5$	79 (23)	64 (19)
	$\gamma = 1$	138 (34)	123 (33)
DM	$\gamma = 0$	315 (653)	154 (160)
	$\gamma = 0.25$	145 (95)	113 (35)
	$\gamma = 0.5$	141 (34)	115 (25)
	$\gamma = 1$	212 (37)	175 (51)

Table 2

Means and standard deviations (in parentheses) of the prediction error, based on 100 random splits of 98 samples into training sets (75 samples) and test sets (23 samples), are reported for the three regularized estimators based on the Dirichlet-tree multinomial (DTM) model, and those based on the Dirichlet multinomial (DM) model.

DTM	$\gamma = 0$	170 (88)
	$\gamma = 0.25$	162 (77)
	$\gamma = 0.5$	149 (69)
DM	$\gamma = 0$	189 (112)
	$\gamma = 0.25$	169 (95)
	$\gamma = 0.5$	154 (76)

Table 3

The 10 most frequently appeared nutrients, respectively for *Bacteroides* and *Prevotella*, are reported based on 100 random splits of the total samples

Bacteroides	Prevotella
Cholesterol	Catechin, flavan-3-ol
Added Germ from wheats	Total Trans/Cis Trans Linoleic
Sucrose	Alcohol
Valine	Added Germ from wheats
Proline	Naringenin, flavanone
Methionine	Cholesterol
Maltose	Glycemic Index
Animal Protein	Sucrose
Serine	Maltose
Tyrosine	Valine