The Annals of Applied Statistics
2013, Vol. 7, No. 1, 418–442
DOI: 10.1214/12-AOAS592

VARIABLE SELECTION FOR SPARSE DIRICHLET-MULTINOMIAL REGRESSION WITH AN APPLICATION TO MICROBIOME DATA ANALYSIS[1]

BY JUN CHEN AND HONGZHE LI

University of Pennsylvania

With the development of next generation sequencing technology, researchers have now been able to study the microbiome composition using direct sequencing, whose output are bacterial taxa counts for each microbiome sample. One goal of microbiome study is to associate the microbiome composition with environmental covariates. We propose to model the taxa counts using a Dirichlet-multinomial (DM) regression model in order to account for overdispersion of observed counts. The DM regression model can be used for testing the association between taxa composition and covariates using the likelihood ratio test. However, when the number of covariates is large, multiple testing can lead to loss of power. To address the high dimensionality of the problem, we develop a penalized likelihood approach to estimate the regression parameters and to select the variables by imposing a sparse group ℓ_1 penalty to encourage both group-level and within-group sparsity. Such a variable selection procedure can lead to selection of the relevant covariates and their associated bacterial taxa. An efficient block-coordinate descent algorithm is developed to solve the optimization problem. We present extensive simulations to demonstrate that the sparse DM regression can result in better identification of the microbiome-associated covariates than models that ignore overdispersion or only consider the proportions. We demonstrate the power of our method in an analysis of a data set evaluating the effects of nutrient intake on human gut microbiome composition. Our results have clearly shown that the nutrient intake is strongly associated with the human gut microbiome.

1. Introduction. The human body is inhabited by complex microbial communities, called microbiomes. It is estimated that the number of microbial cells associated with the human body is 10 times the total number of human cells. The collective genomes of these microbes constitute an extended human genome that provides us with genetic and metabolic capabilities that we do not inherently possess [Bäckhed et al. (2005)]. With the development of next generation sequencing technology such as the 454 pyrosequencing and Illumina Solexa sequencing, microbiome composition can now be determined by direct DNA sequencing without laborious cultivation. Typically, instead of sequencing all bacterial genomic DNA

Received November 2011; revised August 2012.
[1]Supported in part by NIH Grants CA127334, GM097505 and UH2DK083981.
Key words and phrases. Coordinate descent, counts data, overdispersion, regularized likelihood, sparse group penalty.

as in a shotgun metagenomic approach, only the 16S rRNA gene, which is ubiquitous in the bacteria kingdom and has variable regions, is sequenced. Since each bacterial cell is assumed to have the same number of copies of this gene, the basic idea is to isolate from all the bacteria the DNA strands corresponding to some variable region of the gene, to count different versions of the sequences, and then to identify to which bacteria the versions correspond. The types and abundances of different bacteria in a sample can therefore be determined. After preprocessing of the raw sequences, the 16S sequences are either mapped to an existing phylogenetic tree in a taxonomic dependent way [e.g., Matsen, Kodner and Armbrust (2010)] or clustered into operational taxonomic units (OTUs) at a certain similarity level in a taxonomic independent way [e.g., Caporaso et al. (2010), Schloss et al. (2009)]. At 97% similarity level, these OTUs are used to approximate the taxonomic rank *species*. The OTU based approach is most commonly used in 16S based microbiome studies. Each OTU is characterized by a representative DNA sequence and can be assigned a taxonomic lineage by comparing to a known bacterial 16S rRNA database. Most OTUs are in extremely low abundances, with a large proportion being simply singletons (possibly due to sequencing error). We can further aggregate OTUs from the same genus and perform analysis on the abundances at the genus level, which is more robust to sequencing error and can reduce the number of variables significantly. Either way we finally obtain the taxa counts for each sample.

Recent studies have linked the microbiome with human diseases including obesity and inflammatory bowel disease [Virgin and Todd (2011)]. It is therefore important to understand how genetic or environmental factors shape the human microbiome in order to gain insight into etiology of many microbiome-related diseases and to develop therapeutic measures to modulate the microbiome composition. Benson et al. (2010) demonstrated that genetic variants are associated with the mouse gut microbiome. Wu et al. (2011) showed that dietary nutrients are associated with the human gut microbiome. Both studies have considered a large number of genetic loci or nutrients and aimed to identify the genetic variants or nutrients that are associated with the gut microbiome. When there are numerous possible covariates affecting the microbiome composition, variable selection becomes necessary. Variable selection cannot only increase biological interpretability but also provide researchers with a short list of top candidates for biological validation. The methods we develop in this paper are particularly motivated by an ongoing study at the University of Pennsylvania to link the nutrient intake to the human gut microbiome. In this study, gut microbiome data were collected on 98 normal volunteers. In addition, food frequency questionnaire (FFQ) were filled out by these individuals. The questionnaires were scored and the quantitative measurements of 214 micronutrients were obtained. Details of the study and the data set can be found in Section 6 and in Wu et al. (2011). Our goal is to identify the nutrients that are associated with the gut microbiome and also their associated bacterial taxa.

Most of the microbiome studies used distance-based methods to link the microbiome and environmental covariates, where a distance metric was defined between two microbiome samples and statistical analysis was then performed using the distances. However, the choice of distance metric is sometimes subjective and different distances vary in their power of identifying relevant environmental factors. Another limitation of distance-based methods is its inefficiency for detecting subtle changes since distances summarize the overall relationship. In addition, such distance-based approaches do not provide information on how covariates affect the microbiome compositions and which taxa are affected. Therefore, it is desirable to model the counts directly instead of summarizing the data as distances. One way of testing for covariate effects is by performing a multivariate multiple regression (called redundancy analysis in ecology) after appropriate transformation of the count data such as converting into proportions [Legendre and Legendre (2002)]. A pseudo-F statistic is then calculated and the significance is then evaluated by permutation test. Alternatively, one can define a distance between the samples and then use a PERMANOVA procedure to test for covariate effects [McArdle (2001)]. It is easy to show that when the distance is Euclidean, these two procedures are equivalent.

In this paper, we consider the sparse Dirichlet-multinomial (DM) regression [Mosimann (1962)] to link high-dimensional covariates to bacterial taxa counts from microbiome data. The DM regression model is chosen to model the overdispersed taxa counts. The observed taxa count variance is much larger than that predicted by a multinomial model that assumes fixed underlying taxa proportions, an assumption that is hardly met for real microbiome data. Uncontrollable sources of variation such as individual-to-individual variability, day-to-day variability, sampling location variability or even technical variability such as sample preparation lead to enormous variability in the underlying proportions. In contrast, the DM model assumes that the underlying taxa proportions come from a Dirichlet distribution. We use a log-linear link function to associate the mean taxa proportions with covariates. In this DM modeling framework, the effects of the covariates on taxa proportions can be tested using the likelihood ratio test.

When the number of the covariates is large, we propose a sparse group ℓ_1 penalized likelihood approach for variable selection and parameter estimation. The sparse group ℓ_1 penalty function [Friedman, Hastie and Tibshirani (2010)] consists of a group ℓ_1 penalty and an overall ℓ_1 penalty, which induce both group-level sparsity and within-group sparsity. This is particularly relevant in our setting. For the nutrient-microbiome association example, we have p nutrients and q taxa, so the fully parameterized model has $(p + 1) \times q$ coefficients including the intercepts, since each nutrient-taxon association is characterized by one coefficient. The q coefficients for each nutrient constitute a group. If we assume many nutrients have no or ignorable effects on the microbiome composition, the groups of coefficients associated with these irrelevant nutrients should be zero altogether, which is a group-level sparsity that is achieved by imposing a group ℓ_1 penalty.

However, the group ℓ_1 penalty does not perform within-group selection, wherein if one group is selected, all the coefficients in that group are nonzeros. In the case of nutrient-microbiome association, we are also interested in knowing which taxa are associated with a selected nutrient. By imposing an overall ℓ_1 penalty, within-group selection becomes possible. Therefore, we impose a sparse group ℓ_1 penalty not only to select these important nutrients but also to recover relevant nutrient-taxon associations.

Section 2 reviews the Dirichlet-multinomial model for count data. Section 3 introduces the Dirichlet-multinomial regression framework for incorporating covariate effects and proposes a likelihood ratio statistic for testing the covariate effect. Section 4 proposes a sparse group ℓ_1 penalized likelihood procedure for variable selection for the DM models followed by a detailed description of a block-coordinate descent algorithm in Section 4.1. Section 5 shows simulation results and Section 6 demonstrates the proposed method on a real human gut microbiome data set to associate the nutrient intake with the human gut microbiome composition.

2. Dirichlet-multinomial model for microbiome composition data. Suppose we have q bacterial taxa and their counts $Y = (Y_1, Y_2, \ldots, Y_q)$ are random variables. Denote $\mathbf{y} = (y_1, y_2, \ldots, y_q)$ as the observed counts. The simplest model for count data is the multinomial model and its probability function is given as

$$f_M(y_1, y_2, \ldots, y_q; \boldsymbol{\phi}) = \binom{y_+}{\mathbf{y}} \prod_{j=1}^{q} \phi_j^{y_j},$$

where $y_+ = \sum_{j=1}^{q} y_j$ and $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_q)$ are underlying species proportions with $\sum_{j=1}^{q} \phi_j = 1$. Here the total taxa count y_+ is determined by the sequencing depth and is treated as an ancillary statistic since its distribution does not depend on the parameters in the model. The mean and variance of the multinomial component Y_j $(j = 1, \ldots, q)$ are

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

For microbiome composition data, the actual variation is usually larger than what would be predicted by the multinomial model, which assumes fixed underlying proportions. This increased variation is due to the heterogeneity of the microbiome samples and the underlying proportions vary among samples. To account for the extra variation or overdispersion, we assume the underlying proportions $(\phi_1, \phi_2, \ldots, \phi_q)$ are themselves positive random variables $(\Phi_1, \Phi_2, \ldots, \Phi_q)$ subject to the constraint $\sum_{j=1}^{q} \Phi_j = 1$. One commonly used distribution is the Dirichlet distribution [Mosimann (1962)] with the probability function given by

$$f_D(\phi_1, \phi_2, \ldots, \phi_q; \boldsymbol{\gamma}) = \frac{\Gamma(\gamma_+)}{\prod_{j=1}^{q} \Gamma(\gamma_j)} \prod_{j=1}^{q} \phi_j^{\gamma_j - 1},$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)$ are positive parameters, $\gamma_+ = \sum_{j=1}^{q} \gamma_j$ and $\Gamma(\cdot)$ is the Gamma function. The mean and variance of the Dirichlet component Φ_j $(j = 1, \ldots, q)$ are

$$\mathrm{E}(\Phi_j) = \frac{\gamma_j}{\gamma_+}, \qquad \mathrm{Var}(\Phi_j) = \frac{\gamma_j(\gamma_+ - \gamma_j)}{(1+\gamma_+)\gamma_+^2}.$$

The mean is proportional to γ_j and the variance is controlled by γ_+, which can be regarded as a "precision parameter." As γ_+ becomes larger, the proportions are more concentrated around the means.

The Dirichlet-multinomial (DM) distribution [Mosimann (1962)] results from a compound multinomial distribution with weights from the Dirichlet distribution (parametrization I):

$$\begin{aligned} f_{\mathrm{DM}}(y_1, y_2, \ldots, y_q; \boldsymbol{\gamma}) &= \int f_M(y_1, y_2, \ldots, y_q; \boldsymbol{\phi}) f_D(\boldsymbol{\phi}; \boldsymbol{\gamma})\, d\boldsymbol{\phi} \\ &= \binom{y_+}{\mathbf{y}} \frac{\Gamma(y_+ + 1)\Gamma(\gamma_+)}{\Gamma(y_+ + \gamma_+)} \prod_{j=1}^{q} \frac{\Gamma(y_j + \gamma_j)}{\Gamma(\gamma_j)\Gamma(y_j + 1)}. \end{aligned} \tag{2}$$

The mean and variance of the DM distribution for each component Y_j $(j = 1, \ldots, q)$ is given by

$$\mathrm{E}(Y_j) = y_+ \mathrm{E}(\Phi_j), \qquad \mathrm{Var}(Y_j) = y_+ \mathrm{E}(\Phi_j)\{1 - \mathrm{E}(\Phi_j)\}\left(\frac{y_+ + \gamma_+}{1+\gamma_+}\right). \tag{3}$$

Comparing (3) with (1), we see that the variation of the DM component is increased by a factor of $(y_+ + \gamma_+)/(1+\gamma_+)$, where γ_+ controls the degree of overdispersion with a larger value indicating less overdispersion. Using an alternative parameterization, the probability function can be written as (parameterization II)

$$f^*_{\mathrm{DM}}(y_1, y_2, \ldots, y_q; \boldsymbol{\phi}, \theta) = \binom{y_+}{\mathbf{y}} \frac{\prod_{j=1}^{q} \prod_{k=1}^{y_j} \{\phi_j(1-\theta) + (k-1)\theta\}}{\prod_{k=1}^{y_+} \{1 - \theta + (k-1)\theta\}}, \tag{4}$$

where $\phi_j = \gamma_j/\gamma_+$ is the mean and $\theta = 1/(1+\gamma_+)$ is the dispersion parameter. When $\theta = 0$, it is easy to verify (4) is reduced to the multinomial distribution.

3. Dirichlet-multinomial regression for incorporating the covariate effects. When there is no covariate effect, the DM model can be used to produce more accurate estimates of taxa proportions of a given microbiome sample than the simple multinomial model, due to its ability to model the overdispersion. Beyond proportion estimation, microbial ecologists are more interested in associating the microbiome composition with some environmental covariates. Suppose we have n microbiome samples and q species. Let $\mathbf{Y} = (y_{ij})_{n\times q}$ be the observed count matrix for the n samples. Let $\mathbf{X} = (x_{ij})_{n\times p}$ be the design matrix of p covariates

for n samples. We assume the parameters γ_j $(j = 1, \ldots, q)$ in the DM model (parametrization I) depend on the covariate via the following log-linear model,

$$\gamma_j(\mathbf{x}^i) = \exp\left(\alpha_j + \sum_{k=1}^{p} \beta_{jk} x_{ik}\right), \tag{5}$$

where $\mathbf{x}^i$ is the ith row vector of $\mathbf{X}$ and β_{jk} is the coefficient for the jth taxon with respect to kth covariate, whose sign and magnitude measure the effect of the kth covariate on the jth taxon. From (3), we see that $\mathrm{E}(Y_{ij}) \propto \exp(\alpha_j) \prod_{k=1}^{p} \exp(\beta_{jk} x_{ik})$, where $\exp(\alpha_j)$ can be interpreted as the baseline abundance level for species j and the coefficient β_{jk} indicates the magnitude of the kth covariate effect on species j. Though the log-linear link is assumed mainly for ease of computation, it is biologically consistent, in that microorganisms usually exhibit exponential growth in a favorable environment.

For notational simplicity, we denote β_{j0} as α_j and augment $\mathbf{X}$ with an n-vector of 1's as its first column. We number the columns from 0 to p. The link function becomes

$$\gamma_j(\mathbf{x}^i) = \exp\left(\sum_{k=0}^{p} \beta_{jk} x_{ik}\right). \tag{6}$$

Let $\boldsymbol{\beta}$ be the $q \times (p+1)$ regression coefficient matrix, $\boldsymbol{\beta}^j = (\beta_{j0}, \ldots, \beta_{jp})^T$ be the vector of coefficients for the jth taxon $(j = 1, \ldots, q)$ and $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{qk})^T$ be the vector of coefficients for the kth covariate $(k = 0, \ldots, p)$. We also use $\boldsymbol{\beta}$ to denote the $q(p+1)$ vector that contains all the coefficients. Substituting (5) into DM probability function (2) and ignoring the part that does not involve the parameters, the log-likelihood function given the covariates is given by

$$\begin{aligned} l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} \Bigg[& \tilde{\Gamma}\left(\sum_{j=1}^{q} \gamma_j(\mathbf{x}^i; \boldsymbol{\beta}^j)\right) - \tilde{\Gamma}\left(\sum_{j=1}^{q} y_{ij} + \sum_{j=1}^{q} \gamma_j(\mathbf{x}^i; \boldsymbol{\beta}^j)\right) \\ & + \sum_{j=1}^{q} \{\tilde{\Gamma}(y_{ij} + \gamma_j(\mathbf{x}^i; \boldsymbol{\beta}^j)) - \tilde{\Gamma}(\gamma_j(\mathbf{x}^i; \boldsymbol{\beta}^j))\} \Bigg], \end{aligned} \tag{7}$$

where $\tilde{\Gamma}(\cdot)$ is the log-gamma function.

Based on the likelihood function (7), one can test the effect of a given covariate or the joint effects of all covariates on the microbiome composition using the standard likelihood ratio test (LRT). To solve the maximization problem, we implemented the Newton–Raphson algorithm, since the gradient and Hessian matrix of the log-likelihood can be calculated analytically. Alternatively, we can use the general-purpose optimization algorithm such as *nlm* in R, which computes the gradient and Hessian numerically. By selecting an appropriate starting point (e.g., $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{0}$), for moderate-size problems in the dimensions p and q, the algorithm converges to a stationary point sufficiently fast.

With a large number of covariates in the DM regression model, direct maximization of the likelihood function becomes infeasible or unstable. When each covariate is tested separately using the LRT, adjustment for multiple testing is required. In addition, when the number of taxa q is large, the null distribution of the LRT has large degrees of freedom and therefore reduced power. It is also desirable to select the relevant covariates that are associated with the microbiome composition. Although one can test the null hypothesis $H_0 : \beta_{jk} = 0$ for each (j, k) pair by the LRT, adjustment of multiple comparisons can lead to a loss of power. In the next section we present a sparse group ℓ_1 penalized estimation for variable selection and parameter estimation for sparse DM regression models.

4. Variable selection for sparse Dirichlet-multinomial regression. To perform variable selection, we estimate the regression coefficient vector $\boldsymbol{\beta}$ in model (6) by minimizing the following sparse group ℓ_1 penalized negative log-likelihood function,

$$pl(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}, \lambda_1, \lambda_2) = -l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) + \lambda_1 \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_2 + \lambda_2 \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_1, \tag{8}$$

where $l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X})$ is the log-likelihood function defined as in (7), λ_1 and λ_2 are the tuning parameters and $\|\boldsymbol{\beta}_k\|_1 = \sum_{j=1}^{q} |\beta_{ik}|$ is the ℓ_1 norm and $\|\boldsymbol{\beta}_k\|_2 = \sqrt{\sum_{j=1}^{q} \beta_{ik}^2}$ is the group ℓ_1 norm of the coefficient vector $\boldsymbol{\beta}_k$, respectively. We do not penalize the intercept vector $\boldsymbol{\beta}_0$. The first part of the sparse group ℓ_1 penalty is the group ℓ_1 penalty that induces group-level sparsity, which facilitates selection of the covariates that are associated with taxa proportions. The second ℓ_1 penalty on all the coefficients facilitates the within-group selection, which is important for interpretability of the resulting model. A similar penalty involving both group ℓ_1 and ℓ_1 terms is discussed in Peng et al. (2010) and Friedman, Hastie and Tibshirani (2010) for regularized multivariate linear regression. When $\lambda_2 = 0$, criterion (8) reduces to the group lasso.

4.1. *A block-coordinate gradient descent algorithm for sparse group ℓ_1 penalized DM regression.* The sparse group ℓ_1 estimates of $\boldsymbol{\beta}$ can be obtained by minimizing the penalized negative log-likelihood function (8):

$$\hat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2} = \arg\min_{\boldsymbol{\beta}} \left\{ -l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) + \lambda_1 \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_2 + \lambda_2 \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_1 \right\}.$$

Using the general block coordinate gradient descent algorithm of Tseng and Yun (2008), we develop in the following an efficient algorithm to solve this optimization problem. Meier, van de Geer and Bühlmann (2008) present a block coordinate gradient descent algorithm for group lasso for logistic regression that includes only the group ℓ_1 penalty (i.e., $\lambda_2 = 0$). In contrast, our optimization problem (8) has two nondifferentiable parts, both at the individual β_{jk} and at the group $\boldsymbol{\beta}_k$ levels.

The key idea of the algorithm is to combine a quadratic approximation of the log-likelihood function with an additional line search. First we expand (7) at current estimate $\hat{\boldsymbol{\beta}}^{(t)}$ to a second-order Taylor series. The Hessian matrix is then replaced by a suitable matrix $\mathbf{H}^{(t)}$. We define

$$l_Q^{(t)}(\mathbf{d}) = l\big(\hat{\boldsymbol{\beta}}^{(t)}\big) + \mathbf{d}^T \nabla l\big(\hat{\boldsymbol{\beta}}^{(t)}\big) + \tfrac{1}{2}\mathbf{d}^T \mathbf{H}^{(t)} \mathbf{d}, \tag{9}$$

where $\mathbf{d} \in \mathbb{R}^{q(p+1)}$. Also denote $\nabla l(\hat{\boldsymbol{\beta}}^{(t)})_k$ and $\mathbf{d}_k$ the gradient and increment with respect to $\hat{\boldsymbol{\beta}}_k^{(t)}$ for the kth group, and $\nabla l(\hat{\boldsymbol{\beta}}^{(t)})_{sk}$ and $\mathbf{d}_{sk}$ with respect to $\hat{\beta}_{sk}^{(t)}$. We then minimize the following function $pl_Q^{(t)}(\mathbf{d})$ with respect to the kth penalized parameter group:

$$\begin{aligned} pl_Q^{(t)}(\mathbf{d}) &= -l_Q^{(t)}(\mathbf{d}) + \lambda_1 \sum_{k=1}^{p} \big\|\hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k\big\|_2 + \lambda_2 \sum_{k=1}^{p} \big\|\hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k\big\|_1 \\ &\approx pl\big(\hat{\boldsymbol{\beta}}^{(t)} + \mathbf{d}; \mathbf{Y}, \mathbf{X}, \lambda_1, \lambda_2\big). \end{aligned} \tag{10}$$

We restrict ourselves to vectors $\mathbf{d}$ with $\mathbf{d}_j = \mathbf{0}$ for $j \neq k$ and the corresponding $q \times q$ submatrix $\mathbf{H}_{kk}^{(t)}$ for the kth group is a diagonal matrix of the form $\mathbf{H}_{kk}^{(t)} = h_k^{(t)} \mathbf{I}_q$ for some scalar $h_k^{(t)} \in \mathbb{R}$.

The solution to the general optimization problem of the form (10) is given by Theorem 1 and its corollary in the Appendix. Let $S = \{s | |\nabla l(\hat{\boldsymbol{\beta}}^{(t)})_{sk} - h_k^{(t)} \hat{\beta}_{sk}^{(t)}| < \lambda_2\}$ and $\bar{S}$ be the set $\{1, \ldots, q\} \setminus S$. Denote $\mathbf{d}_{Sk}$ the subvector of $\mathbf{d}_k$ with indices in S and $\mathbf{d}_{\bar{S}k}$ in $\bar{S}$. The minimizer of (10) can be decomposed into two parts: The first part $\mathbf{d}_{Sk}^{(t)}$ can be obtained by

$$\mathbf{d}_{Sk}^{(t)} = -\hat{\boldsymbol{\beta}}_{Sk}^{(t)}.$$

The second part $\mathbf{d}_{\bar{S}k}^{(t)}$ can be computed by minimizing

$$f^{(t)}(\mathbf{d}_k) = -\big\{\mathbf{d}_k^T \mathbf{u}_k^{(t)} + \tfrac{1}{2}\mathbf{d}_k^T \mathbf{H}_{kk}^{(t)} \mathbf{d}_k\big\} + \lambda_1 \big\|\hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k\big\|_2 \tag{11}$$

with respect to $\mathbf{d}_{\bar{S}k}$ (set components other than $\mathbf{d}_{\bar{S}k}$ to be 0), where

$$\mathbf{u}_k^{(t)} = \big[\nabla l\big(\hat{\boldsymbol{\beta}}^{(t)}\big)_k - \lambda_2 \operatorname{sgn}\big\{\nabla l\big(\hat{\boldsymbol{\beta}}^{(t)}\big)_k - h_k^{(t)} \hat{\boldsymbol{\beta}}_k^{(t)}\big\}\big]$$

and $\operatorname{sgn}(\cdot)$ is the sign function.

Minimization of (11) with respective to $\mathbf{d}_{\bar{S}k}$ can be performed in a similar fashion as in Meier, van de Geer and Bühlmann (2008) for the group ℓ_1 penalty. Specifically, if $\|\mathbf{u}_{\bar{S}k}^{(t)} - h_k^{(t)} \boldsymbol{\beta}_{\bar{S}k}^{(t)}\|_2 < \lambda_1$, the minimizer of equation (11) for $\mathbf{d}_{\bar{S}k}$ is

$$\mathbf{d}_{\bar{S}k}^{(t)} = -\hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)}.$$

Otherwise

$$\mathbf{d}_{\bar{S}k}^{(t)} = -\frac{1}{h_k^{(t)}} \left\{ \mathbf{u}_{\bar{S}k}^{(t)} - \lambda_1 \frac{\mathbf{u}_{\bar{S}k}^{(t)} - h_k^{(t)} \hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)}}{\|\mathbf{u}_{\bar{S}k}^{(t)} - h_k^{(t)} \hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)}\|_2} \right\}.$$

For the unpenalized intercept, the solution can be directly computed:

$$\mathbf{d}_0^{(t)} = -\frac{1}{h_0^{(t)}} \nabla l(\hat{\boldsymbol{\beta}}^{(t)})_0.$$

If $\mathbf{d}^{(t)} \neq \mathbf{0}$, an inexact line search using the Armijo rule will be performed. Let $\alpha^{(t)}$ be the largest value in $\{\alpha_0 \delta^l\}_{l \geq 0}$ such that

$$pl(\hat{\boldsymbol{\beta}}^{(t)} + \alpha^{(t)} \mathbf{d}^{(t)}) - pl(\hat{\boldsymbol{\beta}}^{(t)}) \leq \alpha^{(t)} \sigma \Delta^{(t)},$$

where $0 < \delta < 1, 0 < \sigma < 1, \alpha_0 > 0$, and $\Delta^{(t)}$ is the improvement in the objective function $pl(\boldsymbol{\beta})$ using a linear approximation, that is,

$$\Delta^{(t)} = -\mathbf{d}^{(t)T} \nabla l(\hat{\boldsymbol{\beta}}^{(t)}) + \lambda_1 \left\{ \sum_{k=1}^{p} \|\hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k^{(t)}\|_2 - \sum_{k=1}^{p} \|\hat{\boldsymbol{\beta}}_k^{(t)}\|_2 \right\}$$

$$+ \lambda_2 \left\{ \sum_{k=1}^{p} \|\hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k^{(t)}\|_1 - \sum_{k=1}^{p} \|\hat{\boldsymbol{\beta}}_k^{(t)}\|_1 \right\}.$$

Finally, we update the current estimate by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \alpha^{(t)} \mathbf{d}^{(t)}.$$

For $\mathbf{H}_{kk}^{(t)}$, we use the same choice as in Meier, van de Geer and Bühlmann (2008), that is,

$$h_k^{(t)} = -\max[\operatorname{diag}\{-\nabla^2 l(\hat{\boldsymbol{\beta}}^{(t)})_{kk}\}, c^*],$$

where $c^* > 0$ is a lower bound to ensure convergence. In this paper, we use the standard choices for the parameters, $\alpha_0 = 1, \delta = 0.5, \sigma = 0.1$ and $c^* = 0.001$ [Tseng and Yun (2008)], in the block coordinate descent algorithm to ensure the convergence of the algorithm.

REMARK. In each iteration of the algorithm detailed above, when estimating the kth column of the $q \times p$ coefficient matrix $\boldsymbol{\beta}$ with all other columns fixed, the algorithm first identifies the coefficients with zero estimates, denoted by set S in the algorithm. For the coefficients in set S, $d_{Sk}^{(t)} = -\hat{\boldsymbol{\beta}}_{Sk}^{(t)}$ and, therefore, when $\alpha^t = 1$, $\hat{\boldsymbol{\beta}}_{Sk}^{(t+1)} = \hat{\boldsymbol{\beta}}_{Sk}^{(t)} + \alpha^t d_{Sk}^{(t)} = 0$ and the coefficients in S are shrunk to zero. Based on its definition, the set S depends on the turning parameter λ_2 and a larger value of λ_2 leads to fewer nonzero coefficients. The algorithm then performs a group shrinkage of the nonzero estimates of the coefficients in the complementary set $\bar{S}$. These nonzero coefficients can further be shrunk to zero as a group if the condition $\|\mathbf{u}_{\bar{S}k}^{(t)} - h_k^{(t)} \boldsymbol{\beta}^{(t)}\|_2 < \lambda_1$ is met, in which case $d_{\bar{S}k}^{(t)} = -\hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)}$ and, therefore, $\hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t+1)} = \hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)} + d_{\bar{S}k}^{(t)} = 0$. Clearly, this group shrinkage depends on the tuning parameter λ_1. Thus, with careful choice of the tuning parameters λ_1 and λ_2, some column group coefficients are set to zero and the within-group sparsity is achieved by the plain ℓ_1 penalty.

4.2. *Tuning parameter selection*. Two tuning parameters λ_1 and λ_2 in the penalized likelihood estimation need to be tuned with data by v-fold cross-validation or a BIC criterion. To facilitate computation, we reparameterize λ_1 and λ_2 as $\lambda_1 = c\lambda\sqrt{q}$ and $\lambda_2 = (1-c)\lambda$. The multiplier $\sqrt{q}$ in the group penalty is used so that the group ℓ_1 penalty and overall ℓ_1 penalty are on a similar scale. Here we use λ to control the overall sparsity level and use $c \in [0, 1]$ to control the proportion of group ℓ_1 in the composite sparse group penalty. When $c = 0$, the penalty is reduced to the lasso; when $c = 1$, it is reduced to a group lasso. We consider the tuning parameter c from the set $\{0, 0.05, 0.1, 0.2, 0.4\}$. For each c, to search for the best tuning parameter value, we run the algorithm from $\lambda_{\max}$ so that it produces the sparsest model with the intercepts $\boldsymbol{\beta}_0$ only. The value $\lambda_{\max}$ can be roughly determined by using the starting value $\boldsymbol{\beta}^{(0)}$ with components $\boldsymbol{\beta}_j^{(0)} = \mathbf{0}$ $(j \neq 0)$ and $\boldsymbol{\beta}_0^{(0)}$ the MLE of (7) without covariates, and choosing the smallest value of λ so that the iteration converges in the first iteration, that is, $\boldsymbol{\beta}^{(0)}$ is a stationary point. We then decrease the λ value and use the estimate of $\boldsymbol{\beta}$ from the last λ as a warm start. The grid of λ can be chosen to be equally spaced on a log-scale, for example, $\lambda_j = 0.96^j \lambda_{\max}$ $(j = 1, \ldots, m)$, where m is set so that $\lambda_{\min} = 0.2\lambda_{\max}$ or, alternatively, we could terminate the loop until the model receives more than the maximum number of nonzero coefficients allowed.

5. Simulation studies.

5.1. *Simulation strategies*. We simulate n microbiome samples, p nutrients and q bacterial taxa to mimic the real data set that we analyze in Section 6. The nutrient intake vector is simulated using a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix $\Sigma_{i,j} = \rho^{|i-j|}$. We simulate p_r relevant nutrients with each nutrient being associated with q_r taxa. For each nutrient, the association coefficients β_{ij} for the q_r taxa are equally spaced over the interval $[0.6f, 0.9f]$ with alternative signs, where f controls the association strength. We consider two growth models to relate the taxa abundances to the covariates. In the exponential growth model, the proportion of the jth taxon of the ith sample is determined as

$$\phi_{ij} = \frac{\exp(\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik})}{\sum_{j=1}^{q} \exp(\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik})}. \tag{12}$$

The intercepts $\boldsymbol{\beta}_0$, which determine the base abundances of the taxa, are taken from a uniform distribution over $(-2.3, 2.3)$ so that the base taxa abundances can differ up to 100 folds. The exponential growth model is a common model for bacteria growth in response to environmental stimuli. We also consider a linear growth model, in which the proportion of the jth taxon of the ith sample is determined as

$$\phi_{ij} = \frac{\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik}}{\sum_{j=1}^{q} (\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik})}.$$

The intercepts $\boldsymbol{\beta}_0$ are now drawn from a uniform distribution over $(0.02, 2)$ so that the base taxa abundances can also differ up to 100 folds. To deal with possible negative $\sum_{k=0}^{p} \beta_{jk} x_{ik}$, we add a small constant to make it positive.

We then generate the count data using the DM model of parametrization II (4) with a common dispersion θ. The number of individuals (sequence reads) for the ith sample m_i is generated from a uniform distribution over $(m, 2m)$. Note that the data are not generated exactly according to our model assumptions, which are based on parametrization I (2) and link (6). This can further demonstrate the robustness of our proposed model.

5.2. *Evaluation of the penalized likelihood approach for selecting covariates affecting the microbiome composition.* To evaluate the variable selection performance of the proposed sparse penalized likelihood approach with group ℓ_1 penalty, we first simulate the count data using the exponential growth model with $n = 100$, $p = 100$, $p_r = 4$, $q = 40$, $q_r = 4$, $m = 1000$, $\theta = 0.025$, and $\rho = 0.4$, totaling 4000 variables. We compare the results to the corresponding penalized estimation of the DM model using only the ℓ_1 penalty function and two other sparse group ℓ_1 estimations based on multinomial or Dirichlet regression. In sparse multinomial regression, we use the multinomial model for count data and the link function is given by (12). We set $\beta_{10} = 0$ to make the coefficients identifiable. In sparse Dirichlet regression, instead of modeling the counts directly, we model the proportions using the Dirichlet distribution and the link function is the same as that of the DM regression. Since the count data contain zeros, we add 0.5 to the cells with 0 counts. We also include results from the LRT based univariate testing procedure for group selection controlling the false discovery rate (FDR) at 0.05.

We measure the selection performance using

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

where TP, FN and FP are true positives, false negatives and false positives, respectively, and F_1 is an overall measure, which weights the precision and recall equally. The averages of these measures are reported based on 100 replications.

To select the best tuning parameter values, we simulate an independent test data set of $n/2$ samples. We then run the penalized procedure over the training data set and re-estimate the selected coefficients using an unpenalized procedure ("nlm" function in R). The log-likelihood of the test data set is calculated based on the re-estimated coefficients and the tuning parameter is selected to maximize the log-likelihood over the test data set. We choose the tuning parameter c from the set $\{0, 0.05, 0.1, 0.2, 0.4\}$. Figure 1 shows that a small c is sufficient to identify the groups efficiently, while further increase of c only improves the group selection marginally. On the other hand, within-group selection exhibits a unimode pattern indicating slight grouping could lead to better identification of within-group elements. In the following simulations, we tune both c and λ to achieve the maximum likelihood values in the test data sets.

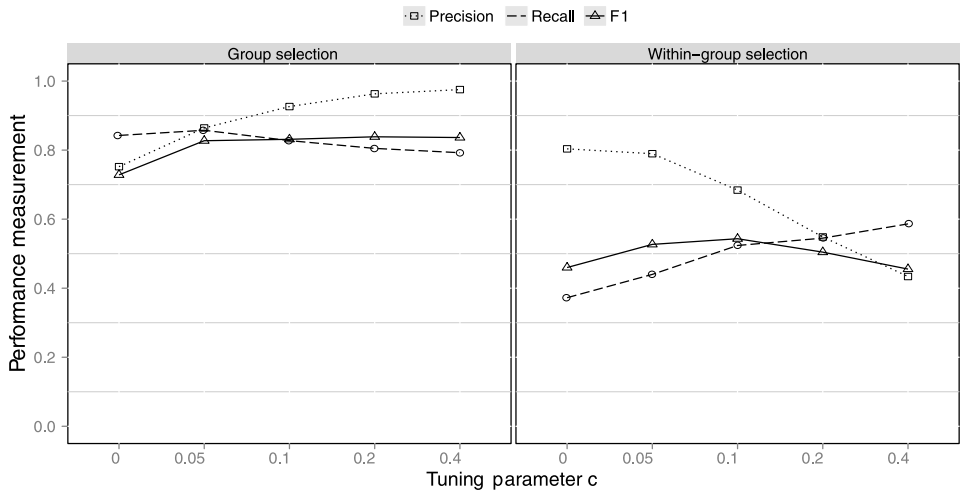

FIG. 1. *Effect of the tuning parameter c on variable selection. The tuning parameter c is varied from* 0 *to* 0.4. *Under each value of c, the best* λ *value, which maximizes the likelihood of the test data set, is selected to generate the sparse model. Group* (*left*) *and within-group* (*right*) *selection performances are then evaluated using measures of recall, precision and* F_1 *based on* 100 *replications. Simulation setting*: $n = 100$, $p = 100$, $p_r = 4$, $q = 40$, $q_r = 4$, $m = 500$, $\theta = 0.025$, $\rho = 0.4$.

Table 1 shows the simulation results. The sparse group ℓ_1 penalized DM regression has a much higher precision rate in group selection than the corresponding ℓ_1 penalized procedure, while both achieve similar recall rates, demonstrating the gain from including the group ℓ_1 penalty in the regularization. Interestingly, the sparse group penalized DM regression also performs better in within-group selection, as shown by a higher recall rate and F_1, indicating better group selection could also facilitate better overall variable selection. Compared to models based on the sparse Dirichlet regression and multinomial regression, the DM model performs better in variable selection, especially for within-group selection, suggesting the DM model is more appropriate than multinomial or Dirichlet models when the counts exhibit overdispersion. The Dirichlet model performs slightly better than the multinomial model. At 5% FDR, the LRT based univariate testing procedure selects far more variables than these penalized procedures, yielding a higher recall rate but a much worse precision rate.

5.3. *Effects of overdispersion and model misspecification.* We further investigate the effect of overdispersion and simulate the count data with different degrees of overdispersion and present the results in Figure 2. We observe that larger overdispersion makes the selection more difficult for all three models, as shown by smaller F_1 values. When the data have slight overdispersion ($\theta = 0.005$), the selection performances of the three models are similar. On the other hand, when the data have much overdispersion ($\theta = 0.1$), DM performs much better than the other

TABLE 1

Comparison of sparse group ℓ_1 and ℓ_1 penalized procedures for variable selection under Dirichlet-multinomial (DM), Dirichlet (D) and multinomial (M) regression models. The selection performance, both group selection and within-group selection, is evaluated using recall rate (R), precision rate (P) and F_1 (F), all averaged over 100 runs (standard deviation in parenthesis). The selection based on a univariate likelihood ratio test (LRT) at FDR = 0.05 is also indicated

	Sparse group ℓ_1 penalization						**ℓ_1 penalization**					
	Within-group			**Group**			**Within-group**			**Group**		
Model	**R**	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
	Exponential growth, $p = 100, q_r = 4, \theta = 0.025$											
DM	0.59	0.70	0.59	0.86	0.92	0.87	0.42	0.76	0.48	0.88	0.68	0.70
	(0.23)	(0.23)	(0.18)	(0.23)	(0.16)	(0.18)	(0.21)	(0.23)	(0.18)	(0.22)	(0.29)	(0.22)
D	0.48	0.73	0.52	0.83	0.89	0.82	0.36	0.82	0.45	0.82	0.77	0.72
	(0.23)	(0.23)	(0.20)	(0.26)	(0.18)	(0.21)	(0.20)	(0.21)	(0.19)	(0.26)	(0.27)	(0.23)
M	0.46	0.72	0.50	0.82	0.85	0.79	0.36	0.76	0.44	0.84	0.70	0.69
	(0.23)	(0.26)	(0.21)	(0.27)	(0.24)	(0.25)	(0.19)	(0.24)	(0.18)	(0.26)	(0.28)	(0.24)
LRT	–	–	–	0.96	0.54	0.66	–	–	–	0.96	0.54	0.66
	–	–	–	(0.11)	(0.21)	(0.16)	–	–	–	(0.11)	(0.21)	(0.16)

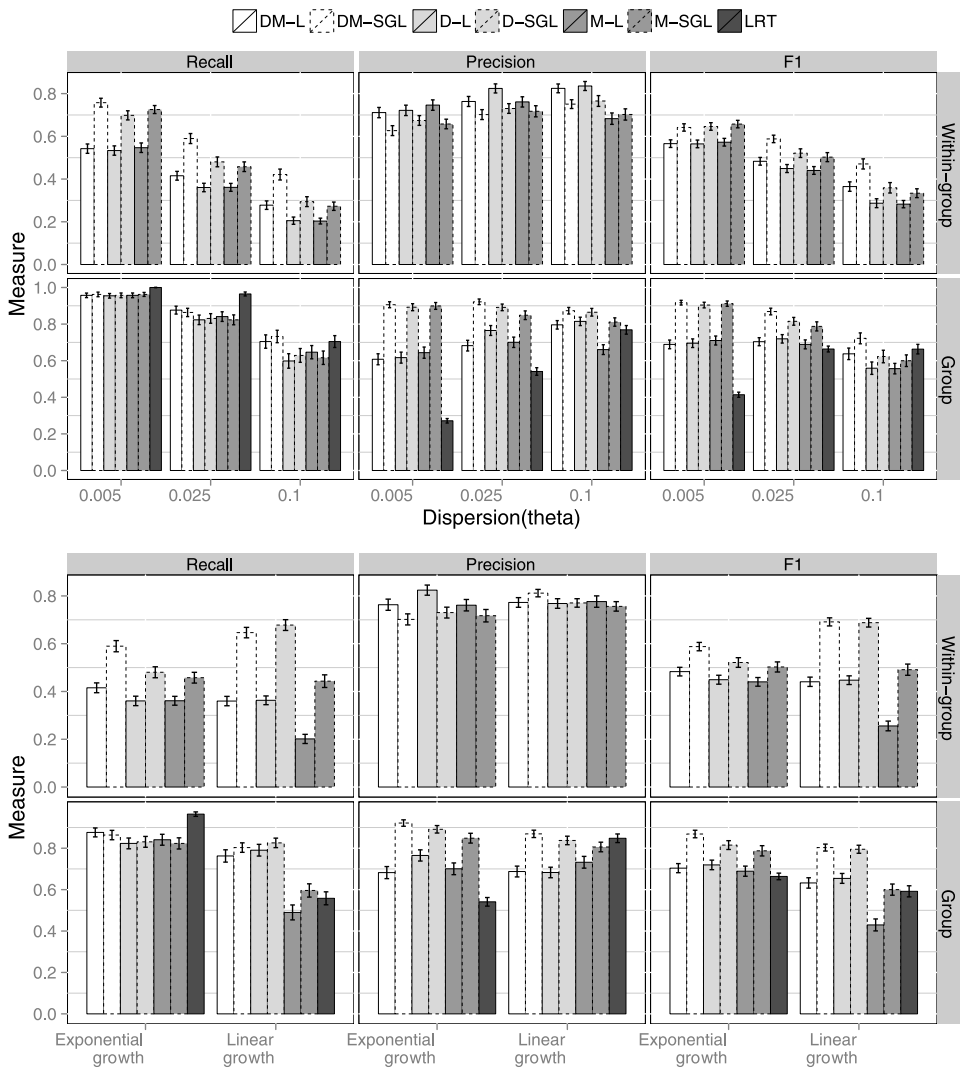

FIG. 2. *Effects of overdispersion* (*top panel*) *and model-misspecification* (*bottom panel*) *on the performance of three different models and methods*. *DM-SGL*: *sparse group* ℓ_1 *penalized Dirichlet-multinomial model*; *DM-L*: ℓ_1 *penalized Dirichlet-multinomial model*; *M-SGL*: *sparse group* ℓ_1 *penalized multinomial model*; *M-L*: ℓ_1 *penalized multinomial model*; *D-SGL*: *sparse group* ℓ_1 *penalized Dirichlet model*; *D-L*: ℓ_1 *penalized Dirichlet model*. *For each bar, mean* $\pm$ *standard error is presented based on* 100 *replications*.

two models in terms of both group selection and within-group selection. Therefore, modeling overdispersion can lead to power gains in identifying relevant variables if the data are overdispersed.

To assess the sensitivity to model misspecification, we simulate the counts using the linear growth model instead and compare the results with the exponential growth model (see Figure 2). Interestingly, both the Dirichlet and DM model are very robust to model misspecification and their selection performances do not decrease significantly. On the other hand, the multinomial model suffers a large performance loss with the F_1 measure for group selection decreasing from 0.79 to 0.56. We also study the effect of the total counts for each sample (data not shown). Even increasing the total count by 10 folds, the DM model is still better than the proportion based Dirichlet model. Therefore, even though we have much deeper sequencing of the microbiome that results in larger counts for each sample, using the DM model can still lead to improved performance over the model that considers only the proportions.

5.4. *Effects of the number of the covariates and the relevant taxa.* We next study the effect of the number of relevant taxa in each group on the performance of different models and present the results in Figure 3. When each relevant group contains only one relevant taxon, the grouping is not very helpful, so the sparse group regularized DM model and ℓ_1 regularized DM model do not differ much in selecting the relevant groups. When the relevant group contains 8 relevant taxa, variable grouping becomes much more important and the sparse group regularized DM model performs much better than the ℓ_1 penalized DM. The group penalized multinomial and Dirichlet regression models, on the other hand, select groups as well as the DM regression model, since the grouping effect is much stronger.

Figure 3 also shows the results when we increase the dimension of covariates to 400 (16,000 variables in total). Increase of the dimension does not deteriorate the variable selection performance, demonstrating the efficiency of our method in handling high-dimensional data.

6. Associating nutrient intake with the human gut microbiome composition. Diet strongly affects the human health, partly by modulating gut microbial community composition. Wu et al. (2011) studied the habitual diet effect on the human gut microbiome, where a cross-section of 98 healthy volunteers were enrolled in the study. Diet information was collected using a food frequency questionnaire (FFQ) and was then converted to nutrient intake values of 214 micronutrients. Nutrient intake was further normalized using the residual method to adjust for caloric intake and was standardized to have mean 0 and standard deviation 1. Since some nutrient measurements were almost identical, we used only one representative for these highly correlated nutrients (correlation $\rho > 0.9$), resulting in 118 representative nutrients. Stool samples were collected and DNA samples were analyzed by the 454/Roche pyrosequencing of 16S rDNA gene segments of the V1–V2 region. The pyrosequences were denoised prior to taxonomic assignment, yielding an average of 9265 ± 3864 (SD) reads per sample. The denoised sequences were

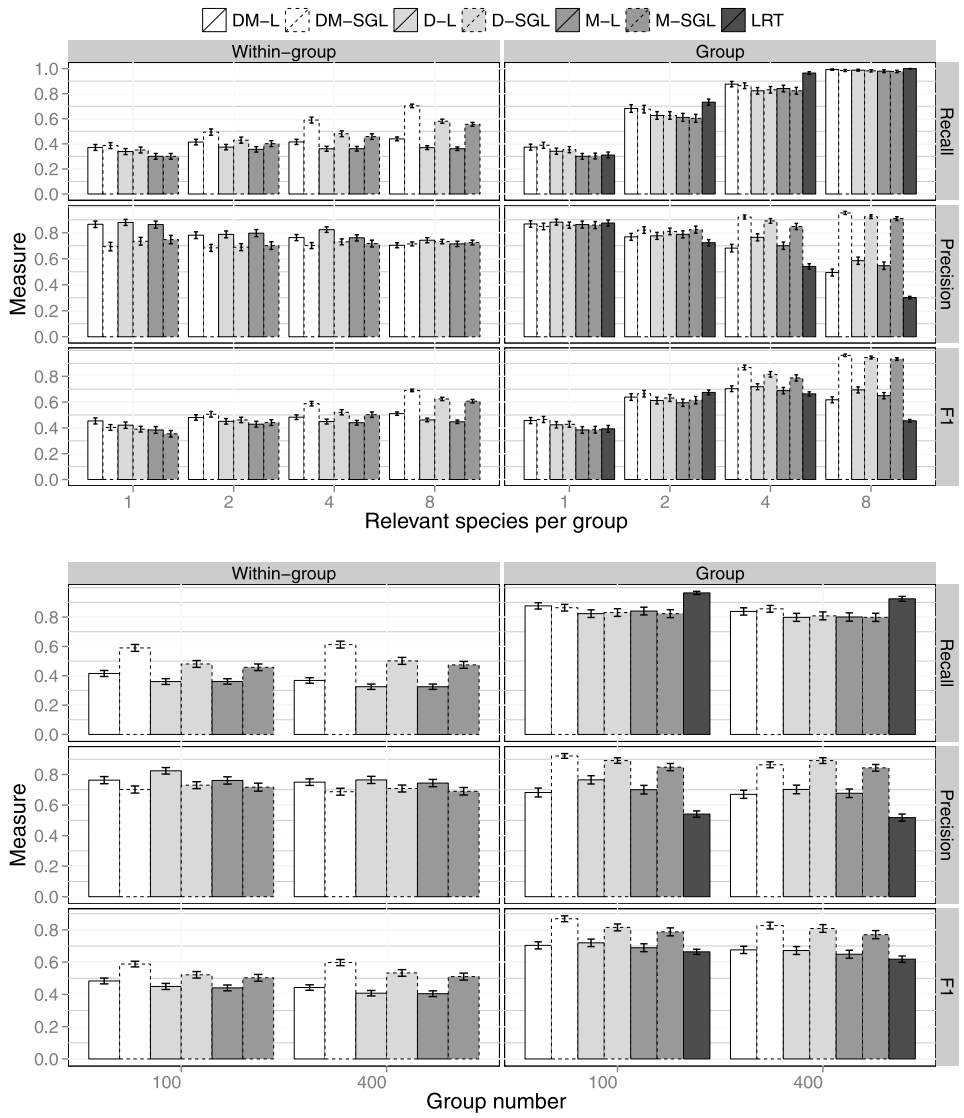

FIG. 3. *Effects of the number of relevant taxa* (*top panel*) *and the number of the covariates* (*bottom panel*) *on the performances of several models and methods. DM-SGL*: *sparse group* ℓ_1 *penalized Dirichlet-multinomial model*; *DM-L*: ℓ_1 *penalized Dirichlet-multinomial model*; *M-SGL*: *sparse group* ℓ_1 *penalized multinomial model*; *M-L*: ℓ_1 *penalized multinomial model*; *D-SGL*: *sparse group* ℓ_1 *penalized Dirichlet model*; *D-L*: ℓ_1 *penalized Dirichlet model. For each bar, mean* $\pm$ *standard error is presented based on* 100 *replications*.

then analyzed by the QIIME pipeline [Caporaso et al. (2010)] with the default parameter settings. The OTU table contained 3068 OTUs (excluding the singletons) and these OTUs can be further combined into 127 genera. We studied 30 relatively

common genera that appeared in at least 25 subjects. Finally, we had the count matrix $\mathbf{Y}_{98\times 30}$ and covariate matrix $X_{98\times 118}$. Our goal is to identify the micronutrients that are associated with the gut microbiomes and the specific genera that the selected nutrients affect.

We applied the sparse group ℓ_1 penalized DM regression to this data set. We used the BIC to select the tuning parameters. The final DM model selected 11 nutrients and 13 associated genera. We refit the DM regression model using the selected variables and obtained the maximum likelihood estimates of the coefficients. We compared the fitted counts (total count $\times$ fitted proportion) against the observed counts in Figure 4 (top panel). The model fits the data quite well with $r^2 = 0.79$. Table 2 shows the MLEs of the regression coefficients for the selected nutrients and genera. Except for Methionine (second column), the coefficients are not too small. Since the nutrient measurements are standardized, the exponentiation of a given coefficient can be interpreted as the factor of change in proportion of a taxon when a given nutrient changes by one unit while other nutrients remain constant. The marginal p-value based on the LRT for each of the selected nutrients is also shown in this table. Except for Vitamin E and Eriodictyol, these selected nutrients all showed a significant marginal association with the gut microbiome.

To further assess the relevance of the nutrients selected, we used the bootstrap to analyze the stability of the selected nutrients [Bach (2008)]. Specifically, we took 100 bootstrap samples and for each sample we ran our algorithm to select the nutrients. Since some nutrients are highly correlated, we expect that highly correlated nutrients (if the correlation is greater than 0.75) can be selected in different bootstrap samples; we define the bootstrap selection probability of a given nutrient as the number of times that this nutrient or its correlated nutrients were selected. Table 2 shows the bootstrap probabilities of the nutrients that were selected by the sparse DM regression, indicating quite stable selection of most of the selected microbiome-associated nutrients. Vitamin E had the least stable selection over the 100 bootstrap samples.

The identified nutrient-taxon associations are visualized in a bipartite graph shown in Figure 5, where the genera and nutrients are depicted with circles and hexagons, respectively. These results further confirmed the findings of Wu et al. (2011), where they found the human gut microbiome can be clustered into two enterotypes characterized by Prevotella and Bacteroides, respectively, and the Prevotella enterotype is associated with a high carbohydrate diet while the Bacteroides enterotype is associated with a high protein/fat/choline diet. Figure 5 shows that two carbohydrates, Maltose and Sucrose, are positively associated with Prevotella and negatively associated with Bacteroides, while animal proteins are positively associated with Bacteroides, Parabacteroides and Alistipes, the three genera mostly enriched in the Bacteroides enterotype. Choline is positively associated with Bacteroides and negatively associated with Prevotella. Polyunsaturated fat is strongly associated with Alistipes, Odoribacter, Barnesiella and Parasutterella, indicating the large effect of fat on the human microbiome.

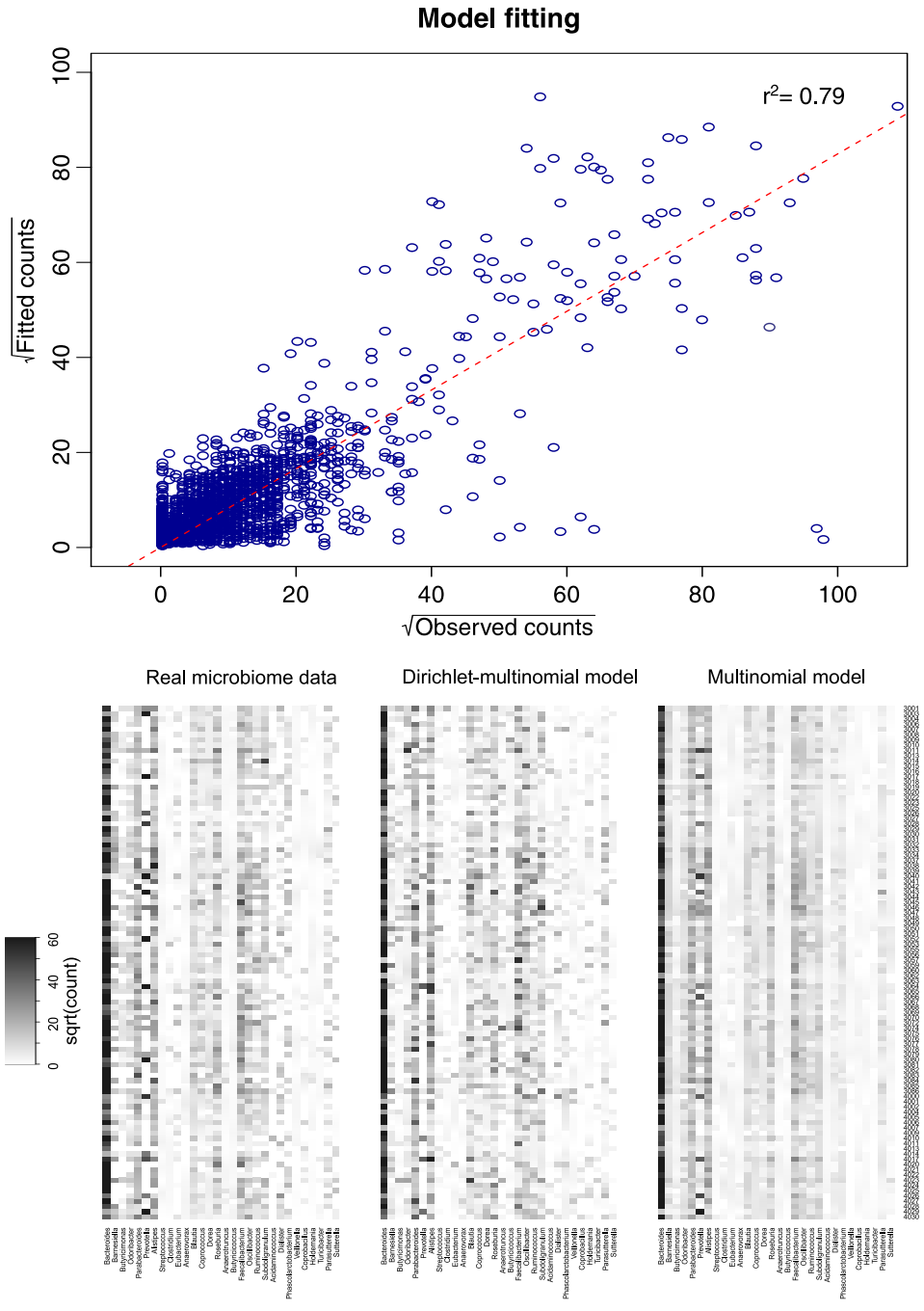

FIG. 4. *Model fit using the variables selected by the sparse group l_1 penalized DM model. Top plot*: *square root of the fitted counts versus square root of the observed counts based on the DM model with the selected nutrients*; *bottom plots*: *observed counts and simulated counts produced by the fitted sparse DM model and multinomial model.*

TABLE 2

Estimated regression coefficients from the sparse group ℓ_1 penalized DM regression for the diet-gut microbiome data. The exponentiation of a given coefficient can be interpreted as the factor of change in proportion of a taxon when a given nutrient changes by one unit while other nutrients remain constant. Columns 1–11 *represent the selected nutrients*: *Polyunsaturated fat*, *Methionine*, *Sucrose*, *Animal Protein*, *Vitamin E-Food Fortification*, *Maltose*, *Added Germ from wheats*, *Choline-Phosphatidylcholine*, *Taurine*, *Naringenin-flavanone and Eriodictyol-flavonone. Rows* 1–13 *represent the selected bacteria taxa*: *Bacteroides*, *Barnesiella*, *Odoribacter*, *Parabacteroides*, *Prevotella*, *Alistipes*, *Coprococcus*, *Faecalibacterium*, *Oscillibacter*, *Ruminococcus*, *Subdoligranulum*, *Phascolarctobacterium and Parasutterella. The marginal p-value based on the LRT and the bootstrap selection probability of each of the selected nutrients are also shown*

Row: taxon; column: nutrient										
–	−0.03	−0.08	0.09	−0.08	−0.10	−0.02	0.02	0.10	–	−0.03
−0.32	–	−0.33	–	–	–	0.22	–	–	–	–
−0.38	–	–	–	–	–	–	–	–	−0.29	–
–	−0.01	−0.08	0.13	−0.07	–	–	–	0.02	−0.23	–
–	–	0.23	–	–	0.36	0.63	−0.72	–	–	–
−0.19	−0.04	–	0.16	–	–	–	–	–	–	0.05
–	–	–	–	–	–	–	–	–	–	0.16
–	–	–	–	–	−0.08	–	–	–	0.07	–
–	−0.02	–	–	–	–	–	–	−0.10	–	–
–	–	0.19	–	–	–	–	–	–	–	–
–	0.02	–	–	–	–	–	–	−0.12	−0.12	0.14
–	–	–	–	−0.35	–	–	–	–	–	–
−0.26	–	−0.29	–	–	–	–	–	–	–	–
Marginal *p*-value										
4.5×10^{-3}	2.2×10^{-4}	8.4×10^{-4}	3.6×10^{-4}	1.1×10^{-1}	6.0×10^{-3}	9.5×10^{-6}	2.7×10^{-3}	5.9×10^{-3}	5.8×10^{-2}	5.2×10^{-3}
Bootstrap selection probability										
0.50	0.93	0.72	0.94	0.35	0.67	0.43	0.58	0.92	0.61	0.60

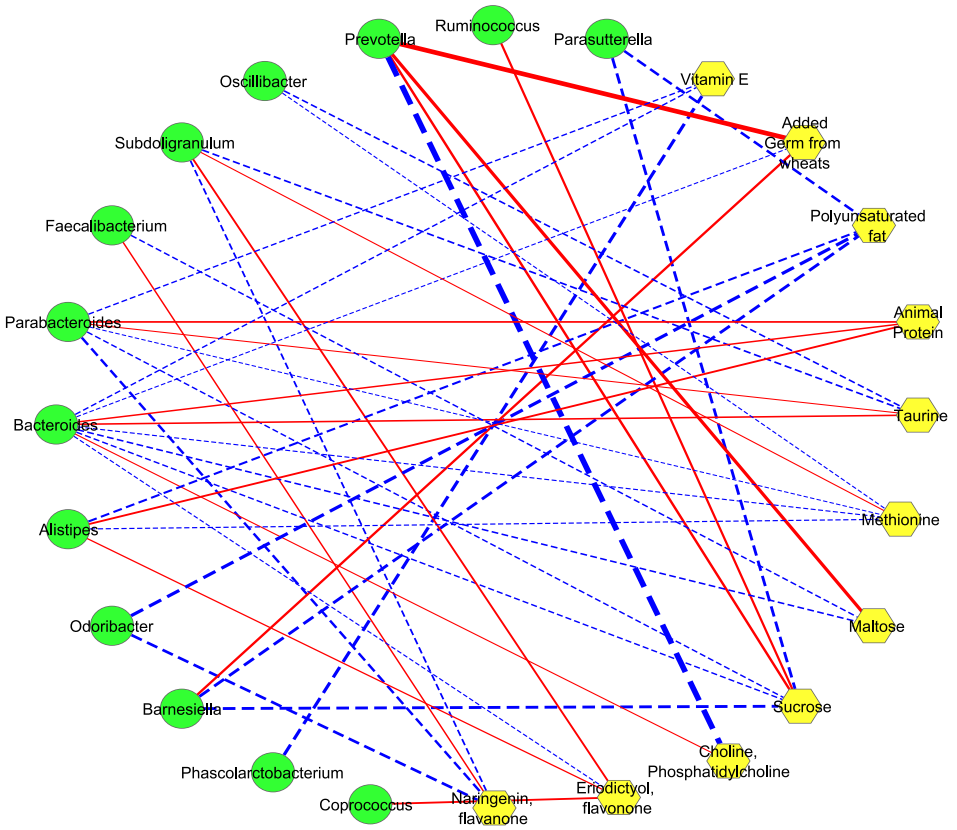

FIG. 5. *Association of nutrients with human gut microbial taxa identified by the sparse group ℓ_1 regularized DM model. We use a bipartite graph to visualize the selected nutrients and their associated genera based on sparse group ℓ_1 penalized DM regression. Circle*: *genus*; *hexagon*: *nutrient*; *solid line*: *positive correlation*; *dashed line*: *negative correlation. The thickness of the line represents the association strength.*

The DM model also identified several other associations that are worth further investigation. For example, we found that Naringenin (flavanone) was positively associated with Faecalibacterium, an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn's disease patients [Sokol et al. (2008)]. If the association is validated, diet with high Naringenin (e.g., Orange, Grapefruit) can be beneficial for patients with Crohn's disease.

As a comparison, we also ran the sparse group ℓ_1 penalized multinomial or Dirichlet regression models and the identified nutrient-genus associations showed significant overlap with those from the DM regression model. However, the interpretability of the DM regression model was the best. To further demonstrate the advantage of the DM model, we simulated taxa counts for each individual based on the fitted models and the observed total taxa counts. The bottom plot of Figure 4

shows that the simulated counts produced by the fitted sparse DM model resemble the observed counts better than those from the sparse multinomial model, where the simulated counts are apparently over-smoothed. This indicates the importance of considering the overdispersion in modeling the gut microbiome data. We also performed the LRT based univariate testing procedure. At FDR $= 0.05$, the LRT identified 13 nutrients, 8 of which are also identified or highly correlated with the nutrients identified by the sparse group ℓ_1 penalized DM model.

7. Discussion. We have proposed a sparse group ℓ_1 penalized estimation for the DM regression in order to select covariates associated with the microbiome composition. The sparse group ℓ_1 penalty encourages both group-level and within-group sparsity, with which we can select the relevant taxa associated with the selected covariates. We have performed extensive simulations to evaluate our proposed penalized estimation procedure for both group and within-group selections. We demonstrated the procedure with a real data set on associating nutrient intakes with gut microbiome composition and confirmed the major findings in Wu et al. (2011).

In our penalized likelihood estimation of the DM model, we use a combination of group ℓ_1 and individual ℓ_1 penalties, which result in a convex and separable (in groups of parameters) penalty function. This property facilitates the application of the general coordinate gradient descent method of Tseng and Yun (2008) to implement an efficient optimization algorithm. In each iteration, we have a closed form solution for a block update. For a given set of the sparsity tuning parameters, our algorithm is fully automatic and does not require the specification of an algorithmic tuning parameter to ensure convergence. For example, it took about 3 minutes on a standard laptop (Core i5, 2G memory) to finish the analysis of the real data set using an R implementation of the algorithm (available at http://statgene.med.upenn.edu/). Besides the sparse l_1 group penalty, other group penalty functions such as the sup-norm penalty in Zhang et al. (2008) and the composite absolute penalties in Zhao, Rocha and Yu (2009) can also be used in the setup of the Dirichlet multinomial regression. However, efficient implementation of the optimization problems with these penalty functions is challenging.

In microbiome data analysis literature, one commonly used approach is to normalize the counts into proportions and perform statistical analysis using the proportions. However, by converting into the proportions, the variation associated with the multinomial sampling process is lost. In 16S rRNA sequencing, the sequencing depths (total counts) for samples can vary up to 10-fold. Obviously, the accuracy of the proportion estimates under sequencing depth of 500 reads is very different from that of 10,000 reads. As shown in our simulations, modeling counts directly can result in gain of power in selecting relevant variables even when the number of sequence reads is very large. Another problem associated with proportions is the existence of numerous zeros in the taxa count data. Many proportion

based approaches require taking logarithms of the proportions, which is problematic for the zero proportions. To circumvent this problem, either a pseudo count (e.g., 0.5) is added to these zero counts before converting into proportions or an arbitrary small proportion is substituted for these zero proportions. The effects of creating pseudo counts have not been evaluated thoroughly when the data contain excessive zeros.

Besides overdispersion, the taxa count data can also exhibit zero-inflation [Barry and Welsh (2002)], where the count data contain more zeros than expected from the DM model. How to model the microbiome count data that allows overdispersion, zero-inflation and possibly the phylogenetic correlations among the taxa is an important future research topic. The multilevel zero-inflated DM regression model for overdispersed count data with extra zeros [Lee et al. (2006), Moghimbeigi et al. (2008)] can potentially provide a solution to this problem. Another problem associated with the DM model is its inflexibility in modeling the covariance structure among the taxa counts. The multinomial model for counts compounded by a logistic normal model [Aitchison (1982)] for proportions provides a possible solution. This needs to be investigated further.

APPENDIX

THEOREM 1. *Letting* $\mathbf{b}, \mathbf{x} \in \mathbb{R}^n$, λ_1, λ_2, c *are nonnegative constants and* $\mathbf{x}^0$ *is the minimizer of the following function*:

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{x} + \mathbf{b}^T\mathbf{x} + c + \lambda_1\|\mathbf{x}\|_2 + \lambda_2\|\mathbf{x}\|_1, \tag{13}$$

then $\mathbf{x}^0_S = \mathbf{0}$ *and*

$$\mathbf{x}^0_{\bar{S}} = \arg\min_{\mathbf{x}_{\bar{S}}}\big\{\tfrac{1}{2}\mathbf{x}^T_{\bar{S}}\mathbf{x}_{\bar{S}} + \big(\mathbf{b}_{\bar{S}} - \lambda_2\,\mathrm{sgn}(\mathbf{b}_{\bar{S}})\big)^T\mathbf{x}_{\bar{S}} + c + \lambda_1\|\mathbf{x}_{\bar{S}}\|_2\big\},$$

where $S = \{i \in \{1,\ldots,n\} | |b_i| < \lambda_2\}$ *and* $\bar{S} = \{1,\ldots,n\} \setminus S$ *and* $\mathrm{sgn}(\cdot)$ *is the sign function.*

PROOF. We prove $\mathbf{x}^0_S = \mathbf{0}$ by contradiction. If $x^0_i \neq 0$ $(i \in S)$, then we can construct a new $\mathbf{x}^1$ with $x^1_i = 0$ and other components being the same as $\mathbf{x}^0$. Clearly, $\frac{1}{2}\mathbf{x}^{1T}\mathbf{x}^1 + \mathbf{b}^T\mathbf{x}^1 + c + \lambda_2\|\mathbf{x}^1\|_1 < \frac{1}{2}\mathbf{x}^{0T}\mathbf{x}^0 + \mathbf{b}^T\mathbf{x}^0 + c + \lambda_2\|\mathbf{x}^0\|_1$ and $\lambda_1\|\mathbf{x}^1\|_2 < \lambda_1\|\mathbf{x}^0\|_2$. The former is due to the fact that $\frac{1}{2}(x^0_i)^2 + b_i x^0_i + \lambda_2|x^0_i| > 0$ for $|b_i| < \lambda_2$. Hence, $\mathbf{x}^0$ is not the minimizer of $f(\mathbf{x})$, which is contradictory. Therefore, $\mathbf{x}^0_S = \mathbf{0}$.

To prove the second part, we note that x^0_i must be either 0 or have an opposite sign of b_i for $i \in \{1,\ldots,n\}$. So the minimization of $\mathbf{f}(\mathbf{x})$ is equivalent to minimiz-

ing

$$f^*(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{x} + \big(\mathbf{b} - \lambda_2 \operatorname{sgn}(\mathbf{b})\big)^T\mathbf{x} + c + \lambda_1\|\mathbf{x}\|_2,$$

subject to

$$\operatorname{sgn}(x_i) = -\operatorname{sgn}(b_i) \quad \text{or} \quad x_i = 0.$$

Since $\mathbf{x}^0_S = \mathbf{0}$, we can restrict the minimization over only $\mathbf{x}_{\bar{S}}$,

$$f^*(\mathbf{x}_{\bar{S}}) = \tfrac{1}{2}\mathbf{x}^T_{\bar{S}}\mathbf{x}_{\bar{S}} + \big(\mathbf{b}_{\bar{S}} - \lambda_2 \operatorname{sgn}(\mathbf{b}_{\bar{S}})\big)^T\mathbf{x}_{\bar{S}} + c + \lambda_1\|\mathbf{x}_{\bar{S}}\|_2, \tag{14}$$

subject to

$$\operatorname{sgn}(x_i) = -\operatorname{sgn}(b_i) \quad \text{or} \quad x_i = 0 \qquad (i \in \bar{S}).$$

Since $\mathbf{x}^0_{\bar{S}}$ is the minimizer of $f^*(\mathbf{x}_{\bar{S}})$ without the constraint, the sign of $\mathbf{x}^0_{\bar{S}}$ should be the opposite of the sign of $(\mathbf{b}_{\bar{S}} - \lambda_2 \operatorname{sgn}(\mathbf{b}_{\bar{S}}))$. Because $|b_i| \geq \lambda_2$ for $i \in \bar{S}$, the sign of $(\mathbf{b}_{\bar{S}} - \lambda_2 \operatorname{sgn}(\mathbf{b}_{\bar{S}}))$ is the same as $\mathbf{b}_{\bar{S}}$. So the sign of $\mathbf{x}^0_{\bar{S}}$ is the opposite of that of $\mathbf{b}_{\bar{S}}$. Therefore, $\mathbf{x}^0_{\bar{S}}$ satisfies the constraint. □

Using simple variable substitution, we have the following corollary.

COROLLARY 1. *Letting* $\mathbf{b}, \boldsymbol{\beta}, \mathbf{d} \in \mathbb{R}^n$, λ_1, λ_2, c *are nonnegative constants and* $\mathbf{d}^0$ *is the minimizer of the following function*,

$$f(\mathbf{d}) = \tfrac{1}{2}\mathbf{d}^T\mathbf{d} + \mathbf{b}^T\mathbf{d} + c + \lambda_1\|\boldsymbol{\beta} + \mathbf{d}\|_2 + \lambda_2\|\boldsymbol{\beta} + \mathbf{d}\|_1, \tag{15}$$

then $\mathbf{d}^0_S = -\boldsymbol{\beta}_S$ *and*

$$\mathbf{d}^0_{\bar{S}} = \arg\min_{\mathbf{d}_{\bar{S}}}\left\{\frac{1}{2}\mathbf{d}^T_{\bar{S}}\mathbf{d}_{\bar{S}} + \big(\mathbf{b}_{\bar{S}} - \lambda_2 \operatorname{sgn}(\mathbf{b}_{\bar{S}} - \boldsymbol{\beta}_{\bar{S}})\big)^T\mathbf{d}_{\bar{S}} + c + \lambda_1\|\mathbf{d}_{\bar{S}} + \boldsymbol{\beta}_{\bar{S}}\|_2\right\},$$

where $S = \{i \in \{1, \ldots, n\} | |b_i - \beta_i| < \lambda_2\}$, $\bar{S} = \{1, \ldots, n\} \setminus S$ *and* $\operatorname{sgn}(\cdot)$ *is the sign function*.

Acknowledgments. We thank Doctors Rick Bushman, James Lewis and Gary Wu for providing the data and for many insightful discussions. We also thank Professor Karen Kafadar, an Associate Editor and two reviewers for many helpful comments.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **44** 139–177. MR0676206

BACH, F. R. (2008). Bolasso: Model consistent Lasso estimation through the bootstrap. In *ICML'08: Proceedings of the 25th International Conference on Machine Learning* 33–40. ACM, New York.

BÄCKHED, F., LEY, R. E., SONNENBURG, J. L., PETERSON, D. A. and GORDON, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science* **307** 1915–1920.

BARRY, S. and WELSH, A. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling* **157** 179–188.

BENSON, A. K., KELLY, S. A., LEGGE, R., MA, F., LOW, S. J., KIM, J., ZHANG, M., OH, P. L., NEHRENBERG, D., HUA, K. et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. USA* **107** 18933–18938.

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PEÑA, A. G., GOODRICH, J. K., GORDON, J. I. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7** 335–336.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). A note on the group lasso and a sparse group lasso. Preprint. Available at arXiv:1001.0736.

LEE, A. H., WANG, K., SCOTT, J. A., YAU, K. K. W. and MCLACHLAN, G. J. (2006). Multilevel zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Stat. Methods Med. Res.* **15** 47–61. MR2225145

LEGENDRE, P. and LEGENDRE, L. (2002). *Numerical Ecology*, 2nd ed. Elsevier, Amsterdam.

MATSEN, F. A., KODNER, R. B. and ARMBRUST, E. V. (2010). pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11** 538.

MCARDLE, B. H. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* **82** 290–297.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

MOGHIMBEIGI, A., ESHRAGHIAN, M. R., MOHAMMAD, K. and MCARDLE, B. (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *J. Appl. Stat.* **35** 1193–1202. MR2522147

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49** 65–82. MR0143299

PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D.-Y., POLLACK, J. R. and WANG, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4** 53–77. MR2758084

SCHLOSS, P. D., WESTCOTT, S. L., RYABIN, T., HALL, J. R., HARTMANN, M., HOLLISTER, E. B., LESNIEWSKI, R. A., OAKLEY, B. B., PARKS, D. H., ROBINSON, C. J. et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75** 7537–7541.

SOKOL, H., PIGNEUR, B., WATTERLOT, L., LAKHDARI, O., BERMÚDEZ-HUMARÁN, L. G., GRATADOUX, J. J., BLUGEON, S., BRIDONNEAU, C., FURET, J. P., CORTHIER, G. et al. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. USA* **105** 16731–16736.

TSENG, P. and YUN, S. (2008). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.* **117** 387–423. MR2421312

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., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R. et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334** 105–108.

ZHANG, H. H., LIU, Y., WU, Y. and ZHU, J. (2008). Variable selection for the multicategory SVM via adaptive sup-norm regularization. *Electron. J. Stat.* **2** 149–167. MR2386091

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566

DEPARTMENT OF BIOSTATISTICS
AND EPIDEMIOLOGY
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6021
USA
E-MAIL: chenjun@mail.med.upenn.edu
hongzhe@upenn.edu