The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1019–1040
DOI: 10.1214/16-AOAS928

REGRESSION ANALYSIS FOR MICROBIOME COMPOSITIONAL DATA

BY PIXU SHI*,[1], ANRU ZHANG† AND HONGZHE LI*[1]

University of Pennsylvania and University of Wisconsin-Madison†*

One important problem in microbiome analysis is to identify the bacterial taxa that are associated with a response, where the microbiome data are summarized as the composition of the bacterial taxa at different taxonomic levels. This paper considers regression analysis with such compositional data as covariates. In order to satisfy the subcompositional coherence of the results, linear models with a set of linear constraints on the regression coefficients are introduced. Such models allow regression analysis for subcompositions and include the log-contrast model for compositional covariates as a special case. A penalized estimation procedure for estimating the regression coefficients and for selecting variables under the linear constraints is developed. A method is also proposed to obtain debiased estimates of the regression coefficients that are asymptotically unbiased and have a joint asymptotic multivariate normal distribution. This provides valid confidence intervals of the regression coefficients and can be used to obtain the p-values. Simulation results show the validity of the confidence intervals and smaller variances of the debiased estimates when the linear constraints are imposed. The proposed methods are applied to a gut microbiome data set and identify four bacterial genera that are associated with the body mass index after adjusting for the total fat and caloric intakes.

1. Introduction.

The human microbiome includes all microorganisms in and on the human body. These microbes play important roles in human metabolism, nutrient intake and energy generation, and thus are essential in human health. The gut microbiome has been shown to be associated with many human diseases such as obesity, diabetes and inflammatory bowel disease [Manichanh et al. (2012), Qin et al. (2012), Turnbaugh et al. (2006)]. Next generation sequencing technologies make it possible to study the microbial compositions without the need for culturing the bacterial species. There are, in general, two approaches to quantify the relative abundances of bacteria in a community. One approach is based on sequencing the 16S ribosomal RNA (rRNA) gene, which is ubiquitous in all bacterial genomes. The resulting sequencing reads provide information about the bacterial taxonomic composition. Another approach is based on shotgun metagenomic sequencing, which sequences all the microbial genomes presented in the sample rather than

Received June 2015; revised January 2016.

[1] Supported in part by NIH Grants CA127334 and GM097505.

Key words and phrases. Compositional coherence, coordinate descent method of multipliers, high dimension, log-contrast model, model selection, regularization.

just one marker gene. Both 16S rRNA and shotgun sequencing approaches provide bacterial taxonomic composition information and have been widely applied to human microbiome studies, including the Human Microbiome Project (HMP) [Turnbaugh et al. (2007)] and the Metagenomics of the Human Intestinal Tract (MetaHIT) project [Qin et al. (2010)].

Several methods are available for quantifying the microbial relative abundances based on the sequencing data, which typically involve aligning the reads to some known database [Segata et al. (2012)]. Since the DNA yielding materials are different across different samples, the resulting numbers of sequencing reads vary greatly from sample to sample. In order to make the microbial abundance comparable across samples, the abundances in read counts are usually normalized to the relative abundances of all bacteria observed. This results in high-dimensional compositional data with a unit sum. Some of the most widely used metagenomic processing softwares such as MEGAN [Huson et al. (2007)] and MetaPhlAn [Segata et al. (2012)] only output the relative abundances of the bacterial taxa at different taxonomic levels.

This paper considers regression analysis of microbiome compositional data, where the goal is to identify the bacterial taxa that are associated with a continuous response such as the body mass index (BMI). Compositional data are strictly positive and multivariate that are constrained to have a unit sum. Such data are also referred to as mixture data [Aitchison and Bacon-Shone (1984), Cornell (2002), Snee (1973)]. Regression analysis with compositional covariates needs to account for the intrinsic multivariate nature and the inherent interrelated structure of such data. For compositional data, it is impossible to alter one proportion without altering at least one of the other proportions. A linear log-contrast model [Aitchison and Bacon-Shone (1984)] has been proposed for compositional data regression where logarithmic-transformed proportions are treated as covariates in a linear regression model with the constraint of the sum of the regression coefficients being zero. Lin et al. (2014) proposed a variable selection procedure for such models in high-dimensional settings and derived the weak oracle property of the resulting estimates. In analysis of microbiome data, it is also of biological interest to study the subcompositions of bacteria taxa within higher taxonomic levels, such as subcompositions of species under a given genus or phylum, or subcompositions of genera within a phylum. In subcompositional data, the proportions of species have been calculated relative to total proportions of the species under a given genus; that is, the values in the subcomposition have been "reclosed" to add up to 1. Regression analysis of such subcompositional data is also considered in this paper.

One of the founding principles of compositional data analysis is that of subcompositional coherence [Aitchison (1982)]: any compositional data analysis should be done in a way that we obtain the same results in a subcomposition, regardless of whether we analyze only that subcomposition or a larger composition containing other parts. This is especially relevant in high-dimensional regression analysis with compositional covariates, where the goal is to select the bacteria whose compositions are associated with the response. Once such bacteria are identified, it is desirable to recalculate the subcomposition only within those identified. However, these subcompositions have different values from those calculated based on a larger set of bacterial taxa. The log-contrast model of Aitchison and Bacon-Shone (1984) and Lin et al. (2014) satisfies this principal by imposing a linear constraint on the regression coefficients. This paper extends this model for analysis of microbiome subcompositions, where multiple linear constraints are imposed in order to achieve the subcompositional coherence.

Penalized and constrained regression, including constrained Lasso regression, has been studied by James, Paulson and Rusmevichientong (2015), where the regression coefficients are subject to a set of linear constraints. A computational algorithm through reformulating the problem as an unconstrained optimization problem was proposed and nonasymptotic error bounds of the estimates were derived. Different from James, Paulson and Rusmevichientong (2015), this paper presents an efficient computational algorithm based on the coordinate descent method of multipliers and the augmented Lagrange of the optimization problem. Since the resulting estimates are often biased due to the ℓ_1 penalty imposed on the coefficients, variance estimation and statistical inference of the resulting estimates are difficult to derive. In order to make the statistical inference on the regression coefficients and to obtain the confidence intervals, asymptoticly unbiased estimates of the regression coefficients are first obtained through a debiased procedure and their joint asymptotic distribution is derived. The proposed debiased procedure extends that of Javanmard and Montanari (2014) to take into account the linear constraints on regression coefficients. However, due to the linear constraints on the regression coefficients, the theoretical developments are different from Javanmard and Montanari (2014).

Section 2 presents linear regression models with linear constraints for compositional covariates. Section 3 presents an efficient coordinate descent method of multipliers to implement the penalized estimation of the regression coefficients under linear constraints. Section 4 provides an algorithm to obtain debiased estimates of the coefficients and derives their joint asymptotic distribution. Section 5 presents results from an analysis of a gut microbiome data set in order to identify the bacterial genera that are associated with BMI. Methods are evaluated in Section 6 through simulations.

2. Regression models for compositional data.

2.1. *Linear log-contrast model.* A linear log-contrast model [Aitchison and Bacon-Shone (1984)] has been proposed for compositional data regression. Specifically, suppose an $n \times p$ matrix $\mathbf{X}$ consists of n samples of the composition of a mixture with p components, and suppose Y is a response variable depending on $\mathbf{X}$. The nature of the composition makes each row of $\mathbf{X}$ lie in a $(p-1)$-dimensional positive simplex $S^{p-1} = \{(x_1, \dots, x_p) : x_j > 0, j = 1, \dots, p \text{ and } \sum_{j=1}^{p} x_j = 1\}$. Based on this nature, Aitchison and Bacon-Shone (1984) introduced a linear log-contrast model as follows:

$$Y = \mathbf{Z}^p \beta_{\backslash p} + \varepsilon, \tag{1}$$

where $\mathbf{Z}^p = \{\log(x_{ij}/x_{ip})\}$ is an $n \times (p-1)$ log-ratio matrix with the pth component as the reference component, $\beta_{\backslash p} = (\beta_1, \dots, \beta_{p-1})$ is the regression coefficient vector, and noise ε is independently distributed as $N(0, \sigma^2)$. An intercept term is not included in the model since it can be eliminated by centering the response and predictor variables.

The selection of a reference component is crucial to analysis, especially in high-dimensional settings. To avoid choosing an arbitrary reference component, Lin et al. (2014) reformulated model (1) as a regression problem with a linear constraint on the coefficients by letting $\beta_p = -\sum_{j=1}^{p-1} \beta_j$,

$$Y = \mathbf{Z}\beta + \varepsilon, \qquad 1_p^\top \beta = 0, \tag{2}$$

where $1_p = (1, \dots, 1)^\top \in \mathbb{R}^p$, $\mathbf{Z} = (z_1, \dots, z_p) = (\log x_{ij}) \in \mathbb{R}^{n \times p}$, and $\beta = (\beta_1, \dots, \beta_p)^\top$.

2.2. *Subcompositional regression model.* In analysis of microbiome data, the relative abundances of taxa are often obtained at different taxonomic ranks, including species, genus, family, class and phylum. It is of interest to study whether the composition of taxa that belong to a given taxon at a higher rank is associated with the response in which case subcompositions of taxa (e.g., all the genera that belong to a given phylum) are calculated. Suppose r taxa at a given rank are considered with m_g taxa at the lower rank that belong to taxon g. Let X_{gs} be the relative abundance of the sth taxon that belong to the gth taxon at a higher rank, for $g = 1, \dots, r$, $s = 1, \dots, m_g$ such that

$$\sum_{s=1}^{m_g} X_{gs} = 1, \qquad \text{for } g = 1, \dots, r.$$

Let $n \times m_g$ matrix $\mathbf{X}_g$ represent n samples of the subcomposition of m_g taxa. The following model can be used to link the subcompositions to a response Y,

$$Y = \sum_{g=1}^{r} \mathbf{Z}_g \beta_g + \varepsilon, \tag{3}$$

where $\mathbf{Z}_g = (Z_{g1}, \dots, Z_{gm_g}) = (\log X_{g1}, \dots, \log X_{gm_g}) \in \mathbb{R}^{n \times m_g}$, and $\beta_g = (\beta_{g1}, \dots, \beta_{gm_g})^\top$. To make the model subcompositional coherence, the following r linear constraints are imposed:

$$\mathbf{1}_{m_g}^\top \beta_g = \sum_{s=1}^{m_g} \beta_{gs} = 0 \qquad \text{for } g = 1, \dots, r.$$

This set of linear constraints can be written as $\mathbf{C}^\top\beta = 0$, where $\beta = (\beta_1^\top, \ldots, \beta_r^\top)^\top$, and

$$\mathbf{C}^\top = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}_{r\times p}.$$

Models (2) and (3) belong to a more general high-dimensional linear model with r linear constraints on the coefficients,

$$Y = \mathbf{Z}\beta + \varepsilon, \qquad \mathbf{C}^\top\beta = 0, \tag{4}$$

where the rows of $\mathbf{Z} \in \mathbb{R}^p$ are independently and identically distributed with mean zero, $\mathbf{C}$ is a $p \times r$ matrix of the constraint coefficients, $\beta = (\beta_1, \ldots, \beta_p)^\top$, and $\varepsilon \sim N_n(0, \sigma^2\mathbf{I})$. Without loss of generality, $\mathbf{C} = (c_1, \ldots, c_r)$ is assumed to be orthonormal. In high-dimensional settings, β is assumed to be s-sparse, where $s = \#\{i : \beta_i \neq 0\}$ and $s = o(\sqrt{n}/\log p)$.

This paper considers estimation and inference of Model (4) under the general linear constraints. Lin et al. (2014) proposed a procedure for variable selection and estimation for Model (2) and derived the weak oracle property of the resulting estimates. James, Paulson and Rusmevichientong (2015) considered a more general model and provided nonasymptotic bounds on estimation errors. However, variances of the estimates and statistical inference are lacking. In this paper, an algorithm to perform variable selection for Model (4) based on ℓ_1 penalized estimation is first proposed based on a coordinate descent method of multipliers. An inference procedure for the penalized estimator of the regression coefficients is then introduced. The proposed approach parallels that of Javanmard and Montanari (2014) by first obtaining debiased estimates of the coefficients for a high-dimensional linear model with linear constraints, $\hat{\beta}^u$, which are shown to be asymptotically Gaussian, with mean β and covariance $\sigma^2(\widetilde{\mathbf{M}}\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{M}})/n$, where $\widehat{\mathbf{\Sigma}}$ is the empirical covariance and $\widetilde{\mathbf{M}}$ is determined by solving a convex program. Based on this asymptotic result, the corresponding confidence intervals and p-values are constructed and used for statistical inference.

3. Penalized estimation.

In the following presentation, for a matrix $\mathbf{A}_{m\times n}$, $\|\mathbf{A}\|_p$ is the ℓ_p operator norm defined as $\|\mathbf{A}\|_p = \sup_{\|x\|_p=1}\|\mathbf{A}x\|_p$, where $\|v\|_p$ is the standard ℓ_p norm of a vector v. In particular, $\|\mathbf{A}\|_\infty = \max_{1\le i\le m}\sum_{j=1}^n |a_{ij}|$. We also define $|\mathbf{A}|_\infty = \max_{i,j}|a_{ij}|$.

Consider model (4). Define $\mathbf{P}_\mathbf{C} = \mathbf{C}\mathbf{C}^\top$ as the projection onto the space spanned by the columns of $\mathbf{C}$. Two basic regularity conditions on $\mathbf{C}$ are assumed:

CONDITION 1. $\|\mathbf{I}_p - \mathbf{P}_\mathbf{C}\|_\infty \le k_0$ for a constant k_0 that is free of p.

CONDITION 2. The diagonal elements of $\mathbf{I}_p - \mathbf{P}_\mathbf{C}$ are greater than zero.

Condition 1 is equivalent to that $\|c_j\|_1\|c_j\|_\infty$, $j = 1, \dots, r$ are all bounded by a constant that is free of p. Condition 2 means that the group of constraints do not indicate simple constraint such as $\beta_j = 0$. If $(\mathbf{I}_p - \mathbf{P}_\mathbf{C})_{j,j} = 0$, then the jth row and column of $\mathbf{I}_p - \mathbf{P}_\mathbf{C}$ are all zeros, and thus $(\mathbf{I}_p - \mathbf{P}_\mathbf{C})e_j = 0$, which means that e_j lies in the space spanned by the columns of $\mathbf{C}$. It is easy to verify that the constraint matrix $\mathbf{C}$ in the log-contrast model (2) or the subcompositional model (3) satisfies both conditions. For example, in the log-contrast model (2), $k_0 = 2$ for $\mathbf{C} = 1_p/\sqrt{p}$ since

$$\|(\mathbf{I}_p - 1_p 1_p^\top/p)a\|_\infty = \left\| a - \frac{1}{p}\sum_{j=1}^p a_j 1 \right\|_\infty \le \|a\|_\infty + \left| \frac{1}{p}\sum_{j=1}^p a_j \right| \le 2\|a\|_\infty.$$

Define $\widetilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{I}_p - \mathbf{P}_\mathbf{C})$. Since $\mathbf{P}_\mathbf{C}\beta = 0$, model (4) can be rewritten as

$$Y = \widetilde{\mathbf{Z}}\beta + \varepsilon, \qquad \mathbf{C}^\top\beta = 0. \tag{5}$$

The regression coefficients can be estimated using ℓ_1 penalized estimation with linear constraints,

$$\widehat{\beta}^n = \underset{\beta}{\operatorname{argmin}}\left(\frac{1}{2n}\|Y - \widetilde{\mathbf{Z}}\beta\|_2^2 + \lambda\|\beta\|_1\right) \qquad \text{subject to } \mathbf{C}^\top\beta = 0, \tag{6}$$

where λ is a tuning parameter.

A coordinate descent method of multipliers can be used to implement the constrained optimization problem (6). First, the augmented Lagrange of optimization problem (6) [Bertsekas (1996)] is formed as

$$L_\mu(\beta, \eta) = \frac{1}{2n}\|y - \widetilde{\mathbf{Z}}\beta\|_2^2 + \lambda\|\beta\|_1 + \eta^\top\mathbf{C}^\top\beta + \frac{\mu}{2}\|\mathbf{C}^\top\beta\|_2^2,$$

where $\eta \in \mathbb{R}^r$ is the Lagrange multiplier, and $\mu > 0$ is a penalty parameter. Problem (6) can be solved by iterations

$$\beta^{k+1} \leftarrow \underset{\beta}{\operatorname{argmin}}\, L_\mu(\beta, \eta^k), \qquad \eta^{k+1} \leftarrow \eta^k + \mu\mathbf{C}^\top\beta^{k+1}.$$

Defining $\xi = \eta/\mu$, the iterations become

$$\beta^{k+1} \leftarrow \underset{\beta}{\operatorname{argmin}}\left\{\frac{1}{2n}\|y - \widetilde{\mathbf{Z}}\beta\|_2^2 + \lambda\|\beta\|_1 + \frac{\mu}{2}\|\mathbf{C}^\top\beta + \xi^k\|_2^2\right\}, \tag{7}$$

$$\xi^{k+1} \leftarrow \xi^k + \mathbf{C}^\top\beta^{k+1}. \tag{8}$$

The iteration of β can be further detailed as

$$\begin{aligned} \beta_j^{k+1} \leftarrow \frac{1}{\frac{\|\tilde{z}_j\|_2^2}{n} + \mu\|C_j\|_2^2} S_\lambda\Bigg[& \frac{1}{n}\tilde{z}_j^\top\Big(y - \sum_{i \ne j}\beta_i^{k+1}\tilde{z}_i\Big) \\ & - \mu\Big(\sum_{i \ne j}\beta_i^{k+1}C_i^\top C_j + C_j^\top\xi^k\Big)\Bigg], \end{aligned} \tag{9}$$

Algorithm 1: Coordinate descent method of multipliers for solving problem (6)

```
Input: $Y$, $\widetilde{\mathbf{Z}}$, and $\lambda$.
Output: $\widehat{\beta}^n$
Initialize $\beta^0$ with 0 or a warm start, $\xi^0 = 0$, $\mu > 0$ and $k = 0$.
For $j = 1, \dots, p, 1, \dots, p, \dots$, update $\beta_j^{k+1}$ by (9) until convergence.
Update $\xi^{k+1}$ by (8).
$k \leftarrow k + 1$ and repeat the two steps above until convergence.
```

where $C_i, i = 1, \dots, p$ are the rows of $\mathbf{C}$, $\tilde{z}_i, i = 1, \dots, p$ are columns of $\widetilde{\mathbf{Z}}$, and $S_\lambda(t) = \operatorname{sgn}(t)(|t| - \lambda)_+$. Combining (7)–(9) yields the following algorithm for solving problem (6).

The penalty parameter μ that is needed to enforce the zero-sum constraints does not affect the convergence of Algorithm 1 as long as $\mu > 0$. It can, however, affect the convergence rate of the algorithm. In this paper, $\mu = 1$ is taken in all the computations.

4. A debiased estimator and its asymptotic distribution.

4.1. *A debiased estimator.* The asymptotic distribution of the ℓ_1 regularized estimator $\widehat{\beta}^n$ is not manageable and $\widehat{\beta}^n$ is biased due to regularization. Javanmard and Montanari (2014) proposed a procedure to construct a debiased version of the unconstrained LASSO estimator that has a tractable asymptotic distribution, which can be used to obtain the confidence intervals of the regression coefficients. Similar debiased procedures were also developed by Zhang and Zhang (2014) and Bühlmann (2013).

Adapting the debiased procedure of Javanmard and Montanari (2014), the following algorithm can be used to obtain debiased estimates of the regression coefficients, $\widehat{\beta}^u$.

To solve problem (10), Matlab package CVX is used for specifying and solving convex programs [Grant and Boyd (2013)]. To briefly explain the logic behind this algorithm, denote $\mathbf{\Sigma} = \mathbb{E}\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}/n$, and suppose that $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ is the eigenvalue/eigenvector decomposition of $\mathbf{\Sigma}$, where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_{p-r})$. Note that $(\mathbf{V}, \mathbf{C})$ is full rank and orthonormal, and

$$\mathbf{\Sigma} = (\mathbf{V}, \mathbf{C}) \begin{pmatrix} \mathbf{\Lambda} & 0 \\ 0 & 0 \end{pmatrix} (\mathbf{V}, \mathbf{C})^\top.$$

Define

$$\mathbf{\Omega} = (\mathbf{V}, \mathbf{C}) \begin{pmatrix} \mathbf{\Lambda}^{-1} & 0 \\ 0 & 0 \end{pmatrix} (\mathbf{V}, \mathbf{C})^\top,$$

and then

$$\mathbf{\Sigma\Omega} = (\mathbf{V}, \mathbf{C}) \begin{pmatrix} \mathbf{I}_{p-r} & 0 \\ 0 & 0 \end{pmatrix} (\mathbf{V}, \mathbf{C})^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_p - \mathbf{P}_\mathbf{C},$$

where $\mathbf{\Omega}$ is the inverse of $\mathbf{\Sigma}$ in the perpendicular space of the column space of $\mathbf{C}$. The debiased algorithm first finds an approximation of $\mathbf{\Omega}$ by rows, denoted by $\widetilde{\mathbf{M}}$, and then corrects the bias based on $\widetilde{\mathbf{M}}$. At the last step of this algorithm, $\widehat{\beta}^u$ is the debiased version of $\widehat{\beta}^n$. It is easy to check that $\mathbf{C}^\top \widehat{\beta}^u = 0$, which is guaranteed by (11).

The feasibility of the optimization (10) is presented in Lemma 1 under the following assumptions on matrix $\widetilde{\mathbf{Z}} = (\widetilde{Z}_1, \dots, \widetilde{Z}_n)^\top$.

CONDITION 3. There exist uniform constants $C_{\min}, C_{\max}$ such that $0 < C_{\min} \le \sigma_{\min}(\mathbf{\Sigma}) \le \sigma_{\max}(\mathbf{\Sigma}) \le C_{\max} < \infty$, where $\sigma_{\max}(\mathbf{A})(\sigma_{\min}(\mathbf{A}))$ is the largest (smallest) nonzero eigenvalue of matrix $\mathbf{A}$.

CONDITION 4. There exists a uniform constant $\kappa \in (0, \infty)$ such that the rows of $\widetilde{\mathbf{Z}}\mathbf{\Omega}^{1/2}$ are sub-Gaussian with $\|\mathbf{\Omega}^{1/2}\widetilde{Z}_1\|_{\psi_2} \le \kappa$, where the sub-Gaussian norm of a random vector $Z \in \mathbb{R}^n$ is defined as

$$\|Z\|_{\psi_2} = \sup\{\|Z^\top x\|_{\psi_2} : x \in \mathbb{R}^n, \|x\|_2 = 1\},$$

with $\|X\|_{\psi_2}$ defined as $\|X\|_{\psi_2} = \sup_{q\ge 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q}$ for a random variable X.

These two conditions are imposed on $\widetilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{I}_p - \mathbf{P}_\mathbf{C})$, not on the original log-ratio matrix $\mathbf{Z}$. For the subcompositional model (3), it is easy to see that $\widetilde{\mathbf{Z}}$ is the matrix of the centered log-ratio (CLR) transformation of the original taxonomic composition [Aitchison (1982)], where

$$\widetilde{\mathbf{Z}}_{gs} = \log \frac{X_{gs}}{\sqrt[m_g]{\prod_{s=1}^{m_g} X_{gs}}}.$$

CLR has been shown to be effective in transforming compositional data to approximately multivariate normal in many real compositional and microbiome data [Aitchison (1982), Kurtz et al. (2015)]. Conditions 3 and 4 are therefore reasonable assumptions in our setting.

The following Lemma shows that if $\gamma = c\sqrt{\log p/n}$ in Algorithm 2 is properly chosen, then $\mathbf{\Omega}$ is in the feasible set of the optimization problem (10) with a large probability.

LEMMA 1. *Let* $\widehat{\mathbf{\Sigma}} \equiv (\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}})/n$ *be the empirical covariance. For any constant* $c > 0$, *the following holds true*:

$$\mathbb{P}\left\{ |\mathbf{\Omega}\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - \mathbf{P}_\mathbf{C})|_\infty \ge c\sqrt{\frac{\log p}{n}} \right\} \le 2p^{-c''},$$

where $c'' = (c^2 C_{\min})/(24e^2\kappa^4 C_{\max}) - 2$.

```
Algorithm 2: Constructing a debiased estimator
  Input: $Y$, $\mathbf{Z}$, $\widehat{\beta}^n$, and $\gamma$.
  Output: $\widehat{\beta}^u$
    Let $\widehat{\beta}^n$ be the regularized estimator from optimization problem (6).
    Set $\widetilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{I}_p - \mathbf{P_C})$.
    Set $\widehat{\mathbf{\Sigma}} \equiv (\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}})/n$.
    for $i = 1, 2, \dots, p$ do:
    Let $m_i$ be a solution of the convex program:
    (10)  $\text{minimize } m^\top \widehat{\mathbf{\Sigma}} m$
          $\text{subject to } \|\widehat{\mathbf{\Sigma}} m - (\mathbf{I}_p - \mathbf{P_C}) e_i\|_\infty \le \gamma.$
    end for
    Set $\mathbf{M} = (m_1, \dots, m_p)^\top$, set
    (11)  $\widetilde{\mathbf{M}} = (\mathbf{I}_p - \mathbf{P_C})\mathbf{M}(\mathbf{I}_p - \mathbf{P_C}).$
    Define the estimator $\widehat{\beta}^u$ as follows:
    (12)  $\widehat{\beta}^u = \widehat{\beta}^n + \frac{1}{n}\widetilde{\mathbf{M}}\widetilde{\mathbf{Z}}^\top (Y - \widetilde{\mathbf{Z}}\widehat{\beta}^n).$
```

4.2. *Asymptotic distribution and inference.* To obtain the asymptotic distribution of the debiased estimator $\widehat{\beta}^u$, an additional assumption on $\widetilde{\mathbf{Z}}$ is required.

CONDITION 5. The inequality $(3\tau - 1)\delta_{2s}^-(\widetilde{\mathbf{Z}}/\sqrt{n}) - (\tau + 1)\delta_{2s}^+(\widetilde{\mathbf{Z}}/\sqrt{n}) \ge 4\tau\phi_0$ holds for a constant $\phi_0 > 0$, where for any matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\delta_k^+(\mathbf{A})$ and $\delta_k^-(\mathbf{A})$ are the upper and lower restricted isometry property (RIP) constants of order k defined as

$$\delta_k^+(\mathbf{A}) = \sup\left\{\frac{\|\mathbf{A}\alpha\|_2^2}{\|\alpha\|_2^2} : \alpha \in \mathbb{R}^m \text{ is } k\text{-sparse vector}\right\},$$

$$\delta_k^-(\mathbf{A}) = \inf\left\{\frac{\|\mathbf{A}\alpha\|_2^2}{\|\alpha\|_2^2} : \alpha \in \mathbb{R}^m \text{ is } k\text{-sparse vector}\right\}.$$

Condition 5 means that $\delta_{2s}^-(\widetilde{\mathbf{Z}}/\sqrt{n})$ and $\delta_{2s}^+(\widetilde{\mathbf{Z}}/\sqrt{n})$ should be close, that is, any $2s$ columns of the CLR transformed compositional data matrix $\widetilde{\mathbf{Z}}/\sqrt{n}$ should be close to orthonormal.

The following theorem gives the asymptotic distribution of the debiased estimates of the regression coefficients.

THEOREM 1. *Consider the linear model* (5) *with β as an s-sparse vector, and let $\widehat{\beta}^u$ be defined as in equation* (12) *in Algorithm* 2. *Then*,

$$\sqrt{n}(\widehat{\beta}^u - \beta) = B + \Delta, \qquad B|\mathbf{Z} \sim N(0, \sigma^2 \widetilde{\mathbf{M}}\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{M}}^\top),$$
$$\Delta = \sqrt{n}(\widetilde{\mathbf{M}}\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - \mathbf{P_C}))(\beta - \widehat{\beta}^n).$$

Further, assume Conditions (1)–(5) *hold. Then setting $\lambda = r\tilde{c}\sigma\sqrt{(\log p)/n}$ in optimization problem* (6) *and $\gamma = c\sqrt{(\log p)/n}$ in Algorithm* 2, *the following holds true*:

$$\mathbb{P}\left\{\|\Delta\|_\infty > \frac{c\tilde{c}k_0(\tau k_0 + 1)}{\phi_0} \cdot \frac{\sigma s \log p}{\sqrt{n}}\right\} \le 2p^{-c'} + 2p^{-c''},$$

where $K = \max_i \sqrt{\widehat{\mathbf{\Sigma}}_{i,i}}$ and constants c' and c'' are given by

$$c' = \frac{\tilde{c}^2}{2K^2} - 1, \qquad c'' = \frac{c^2 C_{\min}}{24e^2\kappa^4 C_{\max}} - 2.$$

Theorem 1 says that $N(0, \sigma^2\widetilde{\mathbf{M}}\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{M}}^\top)$ can be used to approximate the distribution of $\widehat{\beta}^u$ with proper choices of c and $\tilde{c}$ (or, equivalently, γ and λ). This leads to the following corollary that can be used to construct asymptotic confidence intervals and p-values for β in a high-dimensional linear model with linear constraints (4).

COROLLARY 1. *Let $\widehat{\sigma}$ be a consistent estimator of σ.*

1. *Define $\delta_i(\alpha, n) = \Phi^{-1}(1-\alpha/2)\widehat{\sigma}n^{-1/2}[\widetilde{\mathbf{M}}\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{M}}^\top]_{i,i}^{1/2}$. Then $I_i = [\widehat{\beta}_i^u - \delta_i(\alpha, n), \widehat{\beta}_i^u + \delta_i(\alpha, n)]$ is an asymptotic two-sided level $1-\alpha$ confidence interval for β_i.*
2. *For individual hypothesis $H_{0,i} : \beta_i = 0$ versus $H_{0,i} : \beta_i \neq 0$, an asymptotic p-value can be constructed as follows*:

$$P_i = 2\left[1 - \Phi\left(\frac{n^{1/2}|\widehat{\beta}_i^u|}{\widehat{\sigma}[\widetilde{\mathbf{M}}\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{M}}^\top]_{i,i}^{1/2}}\right)\right].$$

The following lemma shows that with Condition 2, the diagonal elements of $\widetilde{\mathbf{M}}\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{M}}^\top$ are nonzero with a γ that is not too large.

LEMMA 2. *Let $\widetilde{\mathbf{M}}$ be the matrix obtained by equation* (11). *Then for $\gamma < (1 - (\mathbf{P_C})_{i,i})/k_0$ and all $i = 1, \dots, p$,*

$$[\widetilde{\mathbf{M}}\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{M}}^\top]_{i,i} \ge \frac{(1 - (\mathbf{P_C})_{i,i} - k_0\gamma)^2}{\widehat{\mathbf{\Sigma}}_{i,i}}.$$

4.3. *Selection of the tuning parameters.* In real applications, the estimator $\widehat{\beta}^n$, tuning parameter λ and estimation of noise level $\widehat{\sigma}$ are obtained through scaled LASSO [Sun and Zhang (2012)]. Specifically, the following two steps are iterated until convergence:

$$\widehat{\beta}^n \leftarrow \underset{\mathbf{C}^\top \beta = 0}{\operatorname{argmin}} \{ \|Y - \widetilde{\mathbf{Z}}\beta\|_2^2 + 2n\lambda_0 \widehat{\sigma} \|\beta\|_1 \},$$

$$\widehat{\sigma}^2 \leftarrow \|Y - \widetilde{\mathbf{Z}}\widehat{\beta}\|_2^2 / n,$$

where $\lambda_0 = \sqrt{2} L_n(k/p)$, $L_n(t) = n^{-1/2}\Phi^{-1}(1-t)$, Φ^{-1} is the quantile function for standard normal and k is the solution of $k = L_1^4(k/p) + 2L_1^2(k/p)$. Then $\widehat{\lambda} = \lambda_0 \widehat{\sigma}$ and $\gamma = a\hat{\lambda}/\hat{\sigma}$ are used in Algorithm 2, where $a = 1/3$ is used in all simulations and real data analysis in this paper.

5. Association between body mass index and gut microbiome.

The gut microbiome plays an important role in food digestion and nutrition absorption. Wu et al. (2011) reported a cross-sectional study to examine the relationship between micronutrients and gut microbiome composition, where the fecal samples of 98 healthy volunteers from the University of Pennsylvania were collected together with demographic data such as body mass index, age and sex. The DNAs from the fecal samples were analyzed by 454/Roche pyrosequencing of 16S rRNA gene segments of the V1–V2 region. After the pyrosequences were denoised, a total of about 900,000 16S reads were obtained with an average of 9165 reads per sample and 3068 operational taxonomic units (OTUs) were obtained. These OTUs were combined into 87 genera that appeared in at least one sample. Out of these 87 genera, 42 genera had zero counts in more than 90% of the samples and were removed from our analysis. The remaining 45 relatively common genera belong to four phyla, *Actinobacteria, Bacteroidetes, Firmicutes* and *Proteobacteria*. Since dysbiosis of gut microbiome has been shown to be associated with obesity [Ley et al. (2005), Ley et al. (2006), Turnbaugh et al. (2006)], it is interesting to identify the bacterial genera that are associated with BMI after adjusting for total fat and caloric intakes. In the following analysis, the zero count was replaced by the maximum rounding error of 0.5 commonly used in compositional and microbiome data analysis [Aitchison (2003), Kurtz et al. (2015)]. Since the number of reads is very large, replacing zero with other very small counts does not affect our results. These read counts are then converted into compositions of the genera or subcompositions of the genera within the phylum.

5.1. *Analysis of the data at the genus-level.* The proposed method was first applied to perform regression analysis with BMI as the response and the log-transformed compositions of the 45 genera as the covariates. In addition, total fat intake and total caloric intake were also included as the covariates in the model.

The model was fit with the constraint that the sum of the coefficients corresponding to the 45 genera is zero, assuming

$$E(\mathrm{BMI}) = \sum_{g=1}^{45} \beta_g \log(X_g) + \gamma_1 \textsc{Fat} + \gamma_2 \textsc{Calorie},$$

where $\sum_{g=1}^{45} \beta_g = 0$, and $\log(X_g)$ is the logarithm of the relative abundance of the gth genus. The goal of this analysis is to identify the bacteria genera that are associated with BMI.

Figure 1 shows the estimated regression coefficients from LASSO with one constraint and their debiased estimates together with the 95% confidence intervals of the regression coefficients. Four genera were statistically significant with p-value of 0.0251 for *Alistipes*, 0.0031 for *Clostridium*, 0.0031 for *Acidaminococcus* and 0.0042 for *Allisonella*, respectively. These four genera were exactly the same genera identified using stability selection by Lin et al. (2014). They belong to two bacterial phyla, *Bacteroidetes* and *Firmicutes*. The results indicate that Alistipes in the Bacteroidetes phylum is negatively associated with BMI, which is consistent with previous findings that the gut microbiota in obese mice and humans tend to have a lower proportion of *Bacteroidetes* [Ley et al. (2005), Ley et al. (2006), Turnbaugh et al. (2006)]. However, for the *Firmucutes* phylum, both the positively

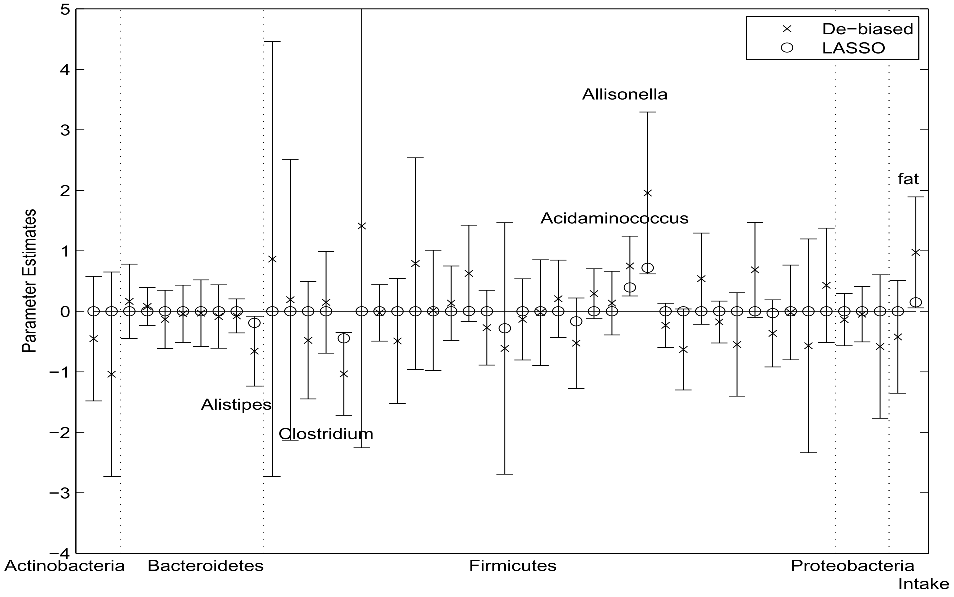

FIG. 1. *Analysis of gut microbiome data. Lasso estimates, debiased estimates and* 95% *confidence intervals of the regression coefficients in the model treating the composition of* 45 *genera as covariates together with total fat and caloric intakes. Dashed vertical lines separate bacterial genus into different phyla.*

associated (*Acidaminococcus* and *Allisonella*) and negatively associated (*Clostridium*) genera were observed to be associated with BMI, suggesting that obesity may be associated with changes in gut microbiome composition at a lower taxonomic level than previously thought.

5.2. *Subcomposition analysis.* The proposed method was then applied to subcomposition analysis where the number of sequencing reads were converted into compositions of genera within each phylum. This creates four subcompositions of the genera within four phyla. This analysis aims to answer the question of whether the composition of genera within a given phylum is associated with BMI, where the log-transformed genera subcompositions are treated as predictors together with total fat and caloric intakes as covariates in the following model:

$$E(\text{BMI}) = \sum_{g=1}^{4}\sum_{s=1}^{m_g} \beta_{gs}\log(X_{gs}) + \gamma_1\text{FAT} + \gamma_2\text{CALORIE},$$

where $\sum_{s=1}^{m_g}\beta_{gs} = 0$ for $g = 1, \dots, 4$, and $\log(X_{gs})$ is the logarithm of the relative abundance of the sth genus of the gth phylum.

Figure 2 shows the LASSO estimates, debiased estimates and 95% confidence interval of the coefficients of the 45 genera. Four genera were statistically significant

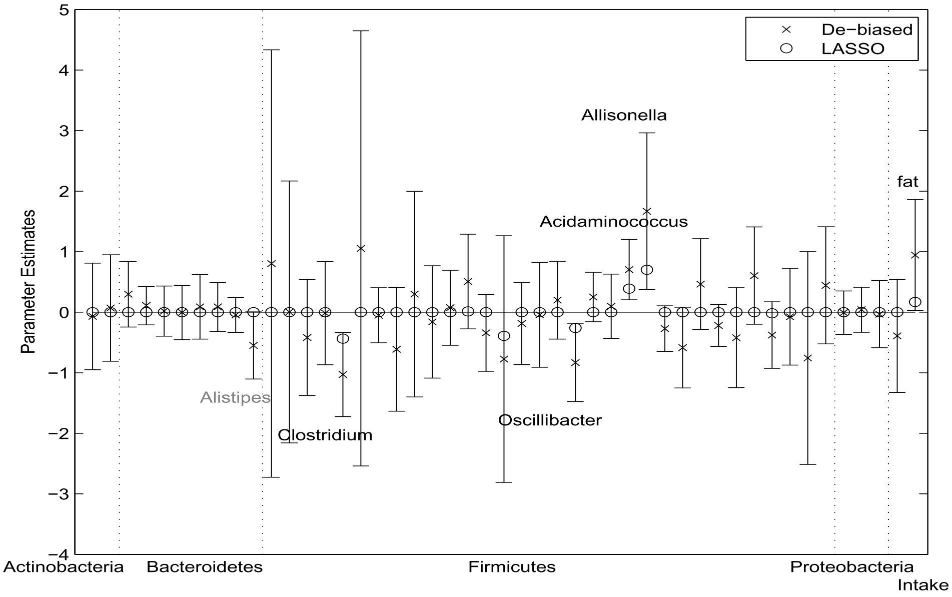

FIG. 2. *Analysis of gut microbiome data. Lasso estimates, debiased estimates and* 95% *confidence intervals of the regression coefficients in the model treating the subcompositions of the genera in each phylum as covariates together with total fat and caloric intakes. Dashed vertical lines separate bacterial genus into different phyla.*

with a p-value of 0.0036 for *Clostridium*, 0.0056 for *Acidaminococcus*, 0.0116 for *Allisonella* and 0.0111 for *Oscillibactor*. All four genera belong to phylum *Firmicutes*, indicating that the subcomposition of the bacterial genera within *Firmicutes* is associated with BMI. The genus *Alistipes* has a p-value of 0.0523 in this analysis, which is marginally significant. It is interesting that the bacterial genus *Oscillibactor* was identified as one of the two bacterial genera that are negatively associated with BMI. *Oscillibacter* was observed to be increased on the resistent starch and reduced carbohydrate weight loss diets [Walker et al. (2011)] in a strictly diet-controlled experiment in obese men, which may explain its negative association with BMI. A recent study also identified *Oscillibacter*-like organisms as a potentially important gut microbe that mediates high fat diet-induced gut dysfunction [Lam et al. (2012)]. It is possible that *Oscillibacter* directly regulates components involved in the maintenance of gut barrier integrity.

Figure 3 shows the predicted BMI using leave-one-out cross-validation (LOOCV). In each round of LOOCV, the variables were selected based on the estimated 95% confidence intervals and the prediction was performed using re-fitted coefficients of the selected bacterial genera together with calorie and fat intakes. An $R^2 = 0.1576$ was obtained between the observed and predicted values. As a comparison, fitting the model with one linear constraint at the genus level resulted in $R^2 = 0.1361$ based on LOOCV, indicating some gain in prediction by the subcompositional analysis.

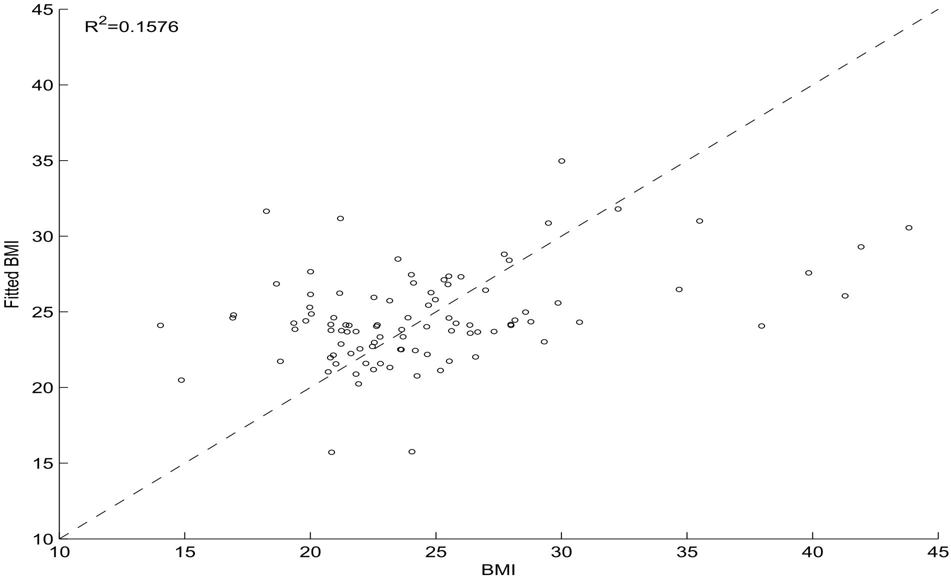

FIG. 3. *Analysis of gut microbiome data. Observed and predicted* BMI *using LOOCV and variables selected based on* 95% *confidence intervals together with total fat and caloric intakes.*

6. Simulation evaluation and comparisons.

In order to simulate the compositional covariates, an $n \times p$ matrix $\mathbf{W}$ of taxon counts is first generated with each row of $\mathbf{W}$ being generated from a log-normal distribution $lnN(\nu, \mathbf{\Sigma})$, where $\mathbf{\Sigma}_{ij} = \zeta^{|i-j|}$ with $\zeta = 0.2$ or 0.5 is the covariance matrix to reflect different levels of correlation between the taxa counts. Parameters $\nu_j = p/2$ for $j = 1, \dots, 5$ and $\nu_j = 1$ for $j = 6, \dots, p$ are set to allow some taxa to be much more abundant than others, as often observed in real microbiome compositional data. The compositional covariate matrix $\mathbf{Z}$ is obtained by normalizing the simulated taxa counts as

$$z_{ij} = \log\left(\frac{w_{ij}}{\sum_{k=1}^{p} w_{ik}}\right), \qquad i = 1, \dots, n, j = 1, \dots, p.$$

Based on these compositional covariates, the response Y is generated through Model (2) with

$$\beta = (1, -0.8, 0.4, 0, 0, -0.6, 0, 0, 0, 0, -1.5, 0, 1.2, 0, 0, 0.3, 0, \dots, 0)$$

and $\sigma = 0.5$. Different dimension/sample size combinations $(p, n) = (50, 100)$, $(50, 200)$, $(50, 500)$, $(100, 100)$, $(100, 200)$, $(100, 500)$ are considered and the simulations are repeated 100 times for each setting. The tuning parameters are chosen using the method described in Section 4.3. The regression coefficient β used in the simulation satisfies the following 8 linear constraints:

$$\begin{array}{llll} \sum_{j=1}^{10} \beta_j = 0, & \sum_{j=11}^{16} \beta_j = 0, & \sum_{j=17}^{20} \beta_j = 0, & \sum_{j=21}^{23} \beta_j = 0, \\ \sum_{j=24}^{30} \beta_j = 0, & \sum_{j=31}^{32} \beta_j = 0, & \sum_{j=33}^{40} \beta_j = 0, & \sum_{i=41}^{p} \beta_j = 0. \end{array} \tag{13}$$

6.1. *Estimation of confidence intervals.* The model is first fitted under the correct constraints specified in (13) and the corresponding confidence intervals are obtained based on our asymptotic results. Figure 4 shows the coverage probability for various models and samples sizes, indicating that the coverage probabilities of the confidence intervals are close to the nominal level of 0.95 when the sample size is large. For small sample sizes, the empirical coverage probability is slightly greater than the nominal level of 0.95, indicating some conservativeness. Figure 5 shows the lengths of confidence intervals. As expected, larger sample sizes result in shorter lengths and larger correlations among the variables lead to increased length of the confidence intervals.

As comparisons, the model is also fitted under no constraint, one single constraint, $\sum_{j=1}^{p} \beta_j = 0$, and misspecified constraints,

$$\sum_{j=1}^{5} \beta_j = 0, \qquad \sum_{j=6}^{12} \beta_j = 0, \qquad \sum_{j=13}^{23} \beta_j = 0, \qquad \sum_{j=24}^{30} \beta_j = 0, \qquad \sum_{j=31}^{p} \beta_j = 0.$$

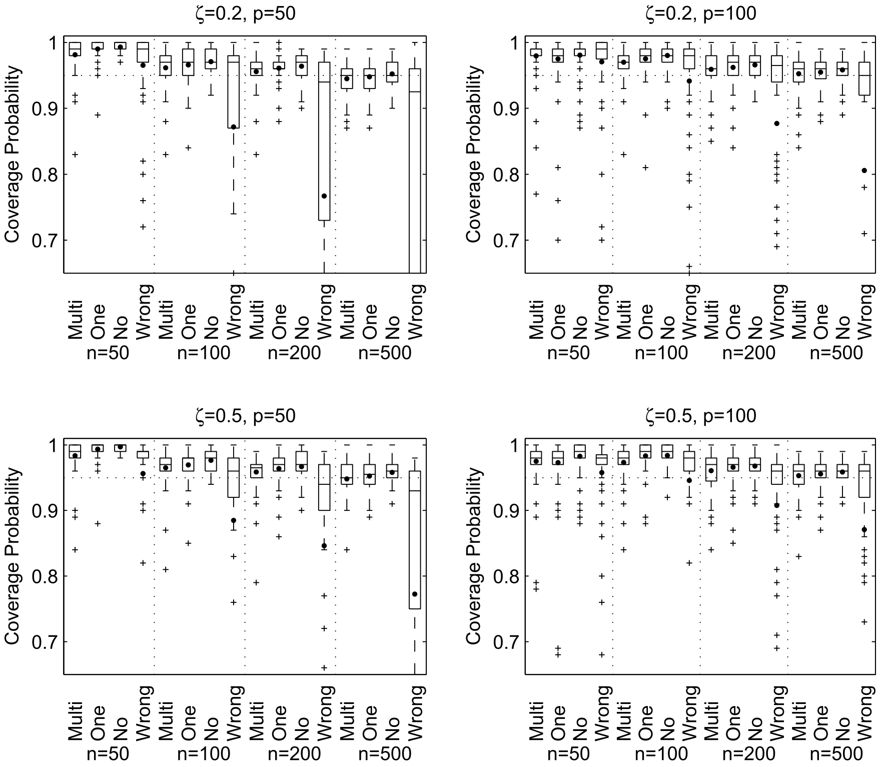

FIG. 4. *Coverage probabilities of confidence intervals based on* 100 *replications. For each model, minimum, median* (*in black line*), *mean* (*in black dot*) *and maximum of the coverage probabilities over all compositional covariates are shown. The confidence intervals are constructed using multiple, one, no and wrong linear constraints, labeled by "Multi," "One," "No" and "Wrong," respectively.*

The coverage probabilities and the lengths of the confidence intervals are given in Figure 4 and Figure 5, respectively. While the coverage probabilities are relatively less sensitive to such misspecification, the intervals estimated under the correct linear constraints are much shorter than those obtained with one or none of the linear constraints, especially when sample size is small. Using the wrong constraints results in much longer intervals with less accurate coverage.

6.2. *Variable selection based on the confidence intervals.* The confidence intervals of the regression coefficients can also be applied to choose the variables of interest. For example, a variable can be selected if the nominal 95% confidence interval of the corresponding regression coefficient includes zero. Table 1 shows the true positive rate and false positive rate of the variables identified based on 95% confidence intervals under multiple constraints, one single constraint and no constraint. When the sample size is small, imposing the correct linear constraints can lead to more true discoveries while the false positive rates are still controlled

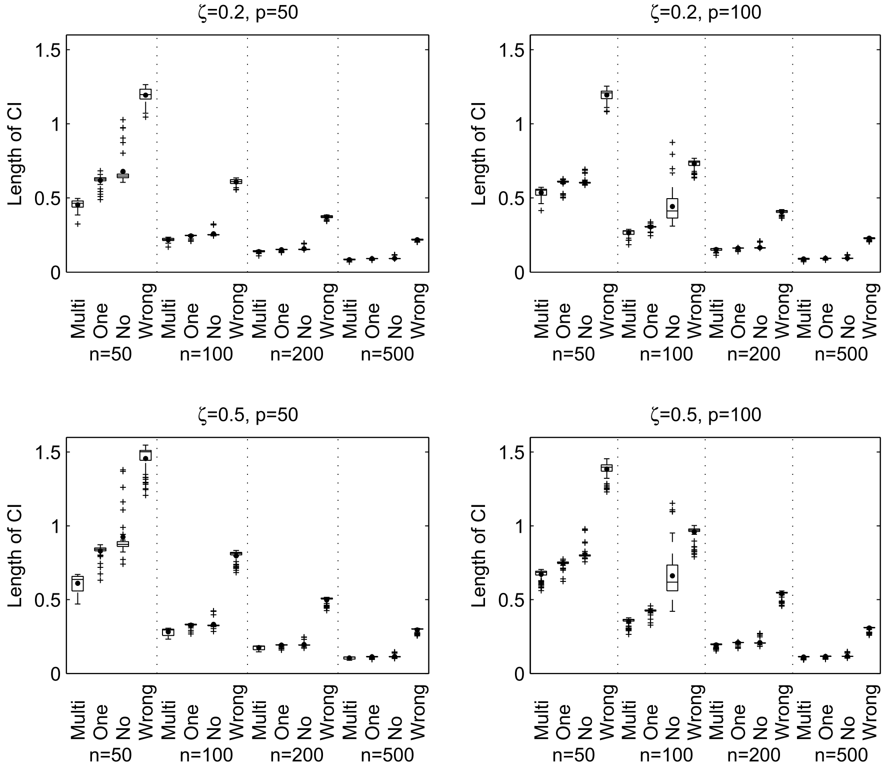

FIG. 5. *Average lengths of confidence intervals based on* 100 *replications. For each model*, *minimum*, *median* (*in black line*), *mean* (*in black dot*) *and maximum of the lengths of the intervals over all compositional covariates are shown. The confidence intervals are constructed using multiple*, *one*, *no and wrong linear constraints*, *labeled by "Multi," "One," "No" and "Wrong," respectively.*

under 5%. In contrast, the models with only one or no constraint lead to much lower true positive rates and the standard LASSO without any constraint gives the worst variable selection results.

6.3. *Prediction evaluation.* Prediction performances are also evaluated and compared for models with or without linear constraints. The prediction error $\|Y - \mathbf{Z}\hat{\beta}\|_2^2/n$ is computed from an independent test sample of size n. Table 2 shows the prediction errors of the LASSO estimator, refitted estimator with variables selected by LASSO, and refitted estimator with variables selected by the 95% confidence intervals. For each of these three estimators, model fitting and coefficient refitting and prediction are performed with multiple, one and no linear constraints. Overall, fitting the models with correct multiple constraints substantially decreases the prediction error. The LASSO estimator has the worst prediction performance, while the two refitted estimators have comparable prediction errors.

TABLE 1

True/False positive rates of the significant variables selected based on 95% *confidence intervals constructed using multiple, one and no linear constraints, labeled by "Multi," "One" and "No," respectively. Variable correlations* ζ, *numbers of variables* p *and sample sizes* (n) *are considered*

			Constraints					
Configuration			**True positive rate**			**False positive rate**		
ζ	p	n	**Multi**	**One**	**No**	**Multi**	**One**	**No**
0.2	50	50	0.9329	0.8514	0.7586	0.0121	0.0056	0.0051
		100	1.0000	1.0000	0.9957	0.0330	0.0286	0.0267
		200	1.0000	1.0000	1.0000	0.0386	0.0333	0.0328
		500	1.0000	1.0000	1.0000	0.0498	0.0477	0.0470
0.2	100	50	0.8571	0.8071	0.7700	0.0131	0.0166	0.0139
		100	1.0000	0.9857	0.9400	0.0265	0.0218	0.0173
		200	1.0000	1.0000	1.0000	0.0374	0.0353	0.0333
		500	1.0000	1.0000	1.0000	0.0441	0.0428	0.0406
0.5	50	50	0.8500	0.7486	0.6543	0.0095	0.0030	0.0019
		100	0.9971	0.9900	0.9871	0.0281	0.0240	0.0223
		200	1.0000	1.0000	1.0000	0.0351	0.0309	0.0305
		500	1.0000	1.0000	1.0000	0.0474	0.0437	0.0412
0.5	100	50	0.7643	0.7157	0.6443	0.0168	0.0173	0.0118
		100	0.9814	0.9300	0.8500	0.0227	0.0137	0.0145
		200	1.0000	1.0000	1.0000	0.0359	0.0320	0.0319
		500	1.0000	1.0000	1.0000	0.0444	0.0417	0.0409

6.4. *Simulation based on real microbiome compositional data.* Another set of simulations are conducted where the gut microbiome composition data analyzed in Section 5 are used to generate the covariates with $p = 45$ through resampling. The many zeros in the compositional data matrix are replaced with a pseudo-count of 0.05 and are renormalized to have a unit sum. For each simulation, we resample with a replacement from the rows of the compositional data matrix to achieve the required sample size. The coefficients β and noise level σ are the same as in the previous section. The sample size is chosen to be $n = 50, 100, 200$ and 500. Each setting is repeated 500 times. The coverage probability and length of confidence intervals are shown in Figure 6 for the model with multiple, one and no constraints on the coefficients. Similar conclusions are observed. The coverage probabilities are relatively less sensitive to misspecification of linear constraints, however, the intervals estimated under the correct linear constraints are shorter than those obtained with one or none of the linear constraints, especially when sample size is small. Using the wrong constraints results in much longer intervals with a less accurate coverage.

TABLE 2

Testing set prediction error of the LASSO *estimator, refitted estimator with variables selected by* LASSO, *and refitted estimator with variables selected based on* 95% *confidence intervals. For each estimator, the model was fit using multiple, one and no linear constraints. Variable correlations* ζ, *numbers of variables* p *and sample sizes* (n) *are considered*

			Constraints								
Configuration			LASSO **Estimator**			**Refitted with Selection by** LASSO			**Refitted with Selection by 95% CI**		
ζ	p	n	**Multi**	**One**	**No**	**Multi**	**One**	**No**	**Multi**	**One**	**No**
0.2	50	50	0.687	0.926	0.983	0.360	0.502	1.336	0.370	0.487	1.375
		100	0.360	0.391	0.412	0.300	0.309	1.153	0.284	0.296	1.155
		200	0.293	0.302	0.307	0.271	0.273	1.039	0.264	0.269	1.054
		500	0.265	0.269	0.270	0.259	0.261	1.025	0.255	0.258	1.034
0.2	100	50	1.027	1.429	1.438	0.484	0.776	1.531	0.496	0.602	1.483
		100	0.408	0.467	0.491	0.305	0.315	1.164	0.286	0.322	1.300
		200	0.303	0.318	0.322	0.273	0.276	1.066	0.268	0.277	1.076
		500	0.269	0.274	0.274	0.263	0.264	1.041	0.260	0.264	1.049
0.5	50	50	0.806	1.095	1.210	0.520	0.687	1.179	0.441	0.557	1.278
		100	0.400	0.476	0.454	0.300	0.319	0.959	0.283	0.301	0.963
		200	0.305	0.325	0.320	0.270	0.272	0.861	0.263	0.267	0.877
		500	0.269	0.276	0.274	0.258	0.260	0.847	0.255	0.257	0.862
0.5	100	50	1.069	1.494	1.731	0.668	0.993	1.416	0.606	0.690	1.361
		100	0.476	0.604	0.560	0.322	0.366	0.963	0.293	0.342	1.134
		200	0.323	0.358	0.342	0.271	0.273	0.884	0.265	0.270	0.896
		500	0.274	0.284	0.279	0.262	0.262	0.863	0.258	0.261	0.876

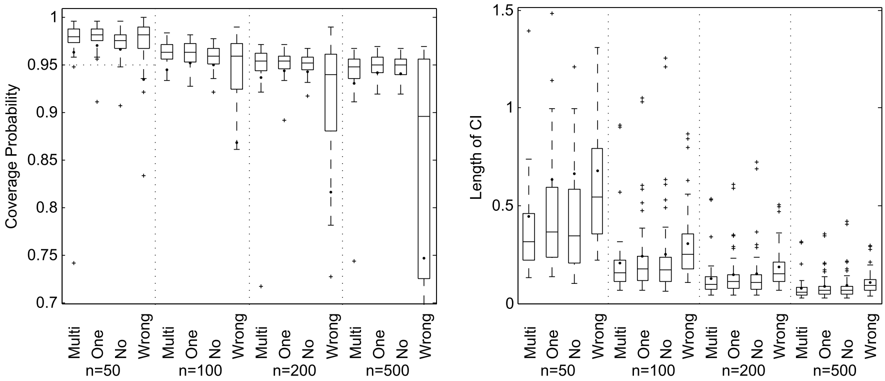

FIG. 6. *Coverage probabilities and length of confidence intervals based on* 500 *replications. Data are simulated by resampling the gut microbiome composition data in Section* 5.

7. Discussion.

This paper has considered the problem of regression analysis for microbiome compositional data obtained through 16S sequencing or metagenomic sequencing. The models and methods in this paper can be applied to identify the microbial subcompositions that are associated with a continuous response. The idea of imposing the constraints on regression coefficients was motivated by using the log-ratios as covariates. However, the method proposed does not use the log-ratios as covariates; it treats the logarithm of the relative abundances as covariates and allows the response to depend on the relative abundances of certain bacteria instead of the ratios. Imposing linear constraints on coefficients enhances the interpretability and also guarantees the subcompositional coherence. Our method allows selecting taxa in different higher rank taxa. By applying our subcompositional analysis, *Oscillibacter* genus was found to be associated with BMI, even after total fat and caloric intakes were adjusted, indicating that the gut microbiome may serve as an independent predictor for complex phenotypes such as BMI. Our simulation studies have demonstrated a clear gain in prediction performance when true linear constraints are imposed. However, the small sample size of our data did not allow us to extensively evaluate the gain in BMI prediction by incorporating the gut microbiome data.

An estimation procedure through regularization under linear constraints has been developed. In order to obtain the confidence interval of the regression coefficients, debiased estimates of the regression coefficients are obtained which are shown to be approximately normally distributed. The p optimization problems in the debiased algorithm can be solved efficiently using convex programs. For one simulated data set in Section 6, Algorithm 2 took about 36 seconds for $p = 100$ and 300 seconds for $p = 200$ on a PC with a core of Intel i7-3770 CPU 3.40 GHz. For large p, convex optimization problems can be carried out in parallel. In typical microbiome studies, p is less than 1000.

The general results presented in this paper can also be used for statistical inference for the log-contrast model considered in Lin et al. (2014). These types of debiased estimates were also proposed in Zhang and Zhang (2014) and van de Geer et al. (2014). Lee et al. (2016) proposed an exact inference procedure for LASSO by characterizing the distribution of a post-selection estimator conditioned on the selection event. It is interesting to extend their approach to the high-dimensional regression problems with constraints. Efron (2014) developed a bootstrap smoothing procedure for computing the standard errors and confidence intervals for predictions, which is different from what was considered in this paper. Efron's procedure can be applied directly to make inferences on predictions using the methods developed here.

Several extensions are worth considering. Model (4) can be extended to include the interaction terms of the form $\lambda_{lk}(\log x_{il} - \log x_{ik})^2$, where x_{il} and x_{ik} are the proportion of the lth and the kth component of subject i, and λ_{lk} is the coefficient that corresponds to the interaction between these two components [Aitchison and Bacon-Shone (1984)]. A similar variable selection and inference procedure can be developed. It is also interesting to develop methods for generalized linear models with high-dimensional compositional data as covariates.

Acknowledgments. We are grateful to the editor, the associate editor and three anonymous referees for their helpful comments. We thank Drs. Gary Wu, James Lews and Rick Bushman for sharing the data and Dr. Wei Lin for helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement to "Regression analysis for microbiome compositional data" (DOI: 10.1214/16-AOAS928SUPP; .pdf). The online Supplemental Materials include proofs of all lemmas and theorems [Shi, Zhang and Hongzhe (2016)].

REFERENCES

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

AITCHISON, J. (2003). *The Statistical Analysis of Compositional Data*. Blackburn Press, Cadwell, NJ.

AITCHISON, J. and BACON-SHONE, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71** 323–330.

BERTSEKAS, D. P. (1996). *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, Belmont.

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

CORNELL, J. A. (2002). *Experiments with Mixtures*: *Designs*, *Models*, *and the Analysis of Mixture Data*, 3rd ed. Wiley, New York. MR1882356

EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671

GRANT, M. and BOYD, S. (2013). CVX: Matlab software for disciplined convex programming, version 2.0 beta. Technical report. Available at http://cvxr.com/cvx.

HUSON, D. H., AUCH, A. F., QI, J. and SCHUSTER, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* **17** 377–386.

JAMES, G. M., PAULSON, C. and RUSMEVICHIENTONG, P. (2015). Penalized and constrained regression. Unpublished manuscript.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biolology* **11** e1004226.

LAM, Y. Y., HA, C. W. Y., CAMPBELL, C. R., MITCHELL, A. J., DINUDOM, A., OSCARSSON, J., COOK, D. I., HUNT, N. H., CATERSON, I. D., HOLMES, A. J. and STORLIEN, L. H. (2012). Increased gut permeability and microbiota change associate with mesenteric fat inflammation and metabolic dysfunction in diet-induced obese mice. *PLoS ONE* **7** e34233.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948

LEY, R. E., BÄCKHED, F., TURNBAUGH, P., LOZUPONE, C. A., KNIGHT, R. D. and GORDON, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. USA* **102** 11070–11075.

LEY, R. E., TURNBAUGH, P. J., KLEIN, S. and GORDON, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. *Nature* **444** 1022–1023.

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797. MR3286917

MANICHANH, C., BORRUEL, N., CASELLAS, F. and GUARNER, F. (2012). The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **9** 599–608.

QIN, J., LI, R., RAES, J., ARUMUGAM, M., BURGDORF, K. S., MANICHANH, C., NIELSEN, T., PONS, N., LEVENEZ, F., YAMADA, T. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464** 59–65.

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y., SHEN, D. et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490** 55–60.

SEGATA, N., WALDRON, L., BALLARINI, A., NARASIMHAN, V., JOUSSON, O. and HUTTENHOWER, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9** 811–814.

SHI, P., ZHANG, A. and LI (2016). Supplement to "Regression analysis for microbiome compositional data." DOI:10.1214/16-AOAS928SUPP.

SNEE, R. D. (1973). Techniques for the analysis of mixture data. *Technometrics* **15** 517–528.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

TURNBAUGH, P. J., LEY, R. E., MAHOWALD, M. A., MAGRINI, V., MARDIS, E. R. and GORDON, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444** 1027–1031.

TURNBAUGH, P. J., LEY, R. E., HAMADY, M., FRASER-LIGGETT, C. M., KNIGHT, R. and GORDON, J. I. (2007). The human microbiome project. *Nature* **449** 804–810.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

WALKER, A. W., INCE, J., DUNCAN, S. H., WEBSTER, L. M., HOLTROP, G., ZE, X., BROWN, D., STARES, M. D., SCOTT, P., BERGERAT, A., LOUIS, P., MCINTOSH, F., JOHNSTONE, A. M., LOBLEY, G. E., PARKHILL, J. and FLINT, H. J. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* **5** 220–230.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R., SINHA, R., GILROY, E., GUPTA, K., BALDASSANO, R., NESSEL, L., LI, H., BUSHMAN, F. D. and LEWIS, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334** 105–108.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

P. SHI
H. LI
DEPARTMENT OF BIOSTATISTICS AND EPIDEMIOLOGY
UNIVERSITY OF PENNSYLVANIA SCHOOL OF MEDICINE
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: pixushi@mail.med.upenn.edu
hongzhe@upenn.edu

A. ZHANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: anruzhang@stat.wisc.edu