The Annals of Applied Statistics
2017, Vol. 11, No. 2, 771–791
DOI: 10.1214/16-AOAS1017

STRUCTURED SUBCOMPOSITION SELECTION IN REGRESSION AND ITS APPLICATION TO MICROBIOME DATA ANALYSIS[1]

BY TAO WANG* AND HONGYU ZHAO*,†

*Shanghai Jiao Tong University** *and Yale University*†

Compositional data arise naturally in many practical problems and the analysis of such data presents many statistical challenges, especially in high dimensions. In this article, we consider the problem of subcomposition selection in regression with compositional covariates, where the relationships among the covariates can be represented by a tree with leaf nodes corresponding to covariates. Assuming that the tree structure is available as prior knowledge, we adopt a symmetric version of the linear log contrast model, and propose a tree-guided regularization method for this structured subcomposition selection. Our method is based on a novel penalty function that incorporates the tree structure information node-by-node, encouraging the selection of subcompositions at subtree levels. We show that this optimization problem can be formulated as a generalized lasso problem, the solution of which can be computed efficiently using existing algorithms. An application to a human gut microbiome study and simulations are presented to compare the performance of the proposed method with an l_1 regularization method where the tree structure information is not utilized.

1. Introduction. Compositional data arise in many disciplines such as geology (geochemical elements), economy (income or expenditure distribution), and ecology (abundances of different species), just to name a few. A compositional data point, or composition for short, can be represented by a positive real vector with as many components as considered. The special and intrinsic feature of compositional data is that the components of a composition are naturally subject to a unit-sum constraint [Aitchison (1986)]. This paper is concerned with regression problems where the covariates are compositional data. Our work is motivated from a data set generated from a human gut microbiome study [Wu et al. (2011)], where the authors explored the relationship between Body Mass Index (BMI) of an individual and the composition of the gut microbiome this person carries (see Section 3 for details).

The human microbiome is the collection of microorganisms that live inside and on the human body. It plays a major role in health and disease in humans, and is

Received November 2015; revised December 2016.

[1]Supported by National Natural Science Foundation of China (11601326) and NIH Grant R01 GM59507.

Key words and phrases. Compositional data analysis, feature selection, homogeneity, log ratio transformations, penalized regression, phylogenetic tree, the lasso.

sometimes referred to as our forgotten organ. To determine the types and abundances of bacteria, the 16S ribosomal RNA (rRNA) gene targeted sequencing is commonly used in microbiome studies. After sequencing, the raw sequences are often clustered into operational taxonomic units (OTUs) at the species level. Each OTU can then be assigned to a taxonomic identity by comparing it to a reference 16S rRNA database. In addition, a phylogenetic tree can be constructed based on the distances between the OTUs. See Navas-Molina et al. (2013) for more details. Because the sampling depths can vary across different samples, a normalization step is often adopted in practice to convert abundances into proportions. Such normalized data are (1) compositional, since for each sample the relative abundances must sum to one, and (2) high-dimensional, since the number of OTUs is large compared to the number of subjects in the study.

Let $Y \in \mathbb{R}$ denote the response variable (e.g., BMI) and $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ the composition of p components (e.g., OTUs). Note that observations of $\boldsymbol{X}$ lie in the $(p-1)$-dimensional simplex:

$$\mathbb{S}^{p-1} = \left\{ (x_1, \ldots, x_p)^\top \in \mathbb{R}^p : x_1 > 0, \ldots, x_p > 0, \sum_{j=1}^{p} x_j = 1 \right\}.$$

To remove the unit-sum constraint and then apply classical statistical methods, Aitchison (1986) suggested transforming the composition $\boldsymbol{X}$ in $\mathbb{S}^{p-1}$ to a vector in $\mathbb{R}^{p-1}$, by a log ratio transformation,

$$\log\left(\frac{X_j}{X_p}\right) = \log(X_j) - \log(X_p), \qquad j = 1, \ldots, p-1.$$

In this paper, we consider the linear log contrast model of Aitchison and Bacon-Shone (1984)

$$Y = \sum_{j=1}^{p-1} \alpha_j \log\left(\frac{X_j}{X_p}\right) + \varepsilon, \tag{1.1}$$

where $\alpha_1, \ldots, \alpha_{p-1}$ are regression coefficients, and the error term ε has mean zero and constant variance. Here, we assume implicitly that the composition $\boldsymbol{X}$ contains all regression information about the response Y, and thus the actual amount of the mixture of the p components is irrelevant. Let $\alpha_p = -\sum_{j=1}^{p-1} \alpha_j$. We can write (1.1) as

$$Y = \sum_{j=1}^{p} \alpha_j \log(X_j) + \varepsilon, \qquad \sum_{j=1}^{p} \alpha_j = 0, \tag{1.2}$$

that is, the mean response, $E(Y \mid \boldsymbol{X})$, is a linear contrast of $\boldsymbol{X}$ in the log scale.

Extracting useful information from high-dimensional data is an important focus of recent statistical research and practice, and for this task penalized loss function

minimization has been shown to be very effective. A popular example is the penalization of the l_2 loss in the usual linear model by the l_1 norm of the parameter vector, known as the lasso [Tibshirani (1996)]. Recently, Lin et al. (2014) proposed a lasso-type procedure for simultaneous component selection and parameter estimation in model (1.2). They assumed that the model is sparse in the sense that many of the coefficients α_j are zero. To see how their method works, we first introduce some notation and definitions.

For a nonempty subset $\mathcal{S} \subseteq \{1, \ldots, p\}$, define

$$X_j^{\mathcal{S}} = \frac{X_j}{\sum_{l \in \mathcal{S}} X_l}, \qquad j \in \mathcal{S}$$

and

$$\mathcal{L}(\mathcal{S}) = \sum_{j \in \mathcal{S}} \alpha_j \log(X_j^{\mathcal{S}}).$$

We call $\{X_j^{\mathcal{S}}, j \in \mathcal{S}\}$, or $\mathcal{S}$ for short, a subcomposition formed from the full composition $\{1, \ldots, p\}$. If $\sum_{j \in \mathcal{S}} \alpha_j = 0$, we call $\mathcal{L}(\mathcal{S})$ a linear contrast of $\mathcal{S}$ (in the log scale). Under model (1.2), a subcomposition $\mathcal{S}$ is said to be inactive if the expected response $E(Y \mid \boldsymbol{X})$ depends only on the subcomposition $\mathcal{S}^c$, the complement of $\mathcal{S}$; $\mathcal{S}$ is said to be active if the expected response $E(Y \mid \boldsymbol{X})$ depends on $\mathcal{S}$ through a linear contrast $\mathcal{L}(\mathcal{S})$. Note that when the cardinality of $\mathcal{S}$, denoted by $|\mathcal{S}|$, is 1, $\{X_j^{\mathcal{S}}, j \in \mathcal{S}\} = \{1\}$ and $\mathcal{L}(\mathcal{S}) = 0$, hence $\mathcal{S}$ is inactive. Loosely speaking, a subcomposition is a subset of components (e.g., a group of bacterial taxa).

Let $\mathcal{A} = \{j : \alpha_j \neq 0\}$. Then, under model (1.2), $\sum_{j \in \mathcal{A}} \alpha_j = 0$ and $E(Y \mid \boldsymbol{X}) = \sum_{j \in \mathcal{A}} \alpha_j \log(X_j^{\mathcal{A}})$. By definition, $\mathcal{A}$ is active as long as $|\mathcal{A}| > 1$. In other words, what the lasso (or component selection in general) actually targets is a single subcomposition composed of selected components. The reason is that the linear log contrast model is different from the standard linear model. For illustration, we consider a toy example, in which $p = 8$ and the mean function has the form $E(Y \mid \boldsymbol{X}) = \log(X_1) - \log(X_2) + \log(X_3) - \log(X_4)$. Then the set consisting of $\{5, 6, 7, 8\}$ is inactive, the set consisting of $\{1, 2, 3, 4\}$ is active, and the latter can be partitioned into two active subcompositions, $\{1, 2\}$ and $\{3, 4\}$. Clearly, the lasso can pick up $\{1, 2, 3, 4\}$, but it cannot tell whether or not the two subcompositions formed from it are active. In human microbiome studies, an important objective is to identify groups of bacterial species present in a body habitat (e.g., the gut) that are predictive of a phenotype (e.g., BMI) [Knights et al. (2011)]. The lasso only provides an approximate solution (i.e., a single group) for this purpose, a practical disadvantage.

Since a composition carries only relative information, subcompositions, which preserve ratio relationships, are natural objects of investigation in compositional data analysis. Indeed, subcompositional analysis has a long history, for example, in geology, and is a major theme of Aitchison (1986). It turns out that in the regression setting, the counterpart of component selection should be subcomposition

selection. In this paper, we assume that there is a partition of the full composition $\{1, \ldots, p\}$ into $K+1 \geq 2$ nonoverlapping subcompositions $\mathcal{S}_k$, $k = 1, \ldots, K+1$, such that $|\mathcal{S}_k| > 1$ for $k = 1, \ldots, K$, and

$$E(Y \mid \boldsymbol{X}) = \sum_{k=1}^{K} \sum_{j \in \mathcal{S}_k} \alpha_j(k) \log\bigl(X_j^{\mathcal{S}_k}\bigr), \qquad \sum_{j \in \mathcal{S}_k} \alpha_j(k) = 0,\ j = 1, \ldots, K, \tag{1.3}$$

that is, the subcomposition $\mathcal{S}_{K+1}$ is inactive, and the expected response depends on K subcompositions formed from a nonoverlapping partition of $\mathcal{S}_{K+1}^c$. To identify subcompositions (say, groups of bacterial species) that are predictive of an outcome (say, BMI), we need to infer K, the number of linear contrasts, and the corresponding coefficients within subcompositions.

The problem of subcomposition selection is challenging. First, the total number of all possible partitions of the full composition, which is the pth Bell number [Rota (1964)], is much larger than that of all possible subsets of components, and hence it is computationally infeasible, even for a moderate p, to enumerate over all possible least squares regressions for identifying the best partition. Second, in some situations, the partition of $\mathcal{S}_{K+1}^c$ is not unique, and hence it is impossible to identify a particular set of subcompositions. In the above toy example, the active subcomposition $\{1, 2, 3, 4\}$ can be divided into two active subcompositions in two ways: (i) $\{1, 2\}$ and $\{3, 4\}$, and (ii) $\{1, 4\}$ and $\{2, 3\}$. There seems to be little to distinguish between (i) and (ii) in terms of goodness-of-fit. For the first issue, it is desirable to develop a regularization method for subcomposition selection, analogous to the lasso for variable selection. For the second issue, if there is prior knowledge on possible subcompositions that one may wish to see selected jointly, then we can perform knowledge-based subcomposition selection. This is the case in our motivating example in which a phylogenetic tree that encodes the relationships among the species is available. Each node of the tree, which corresponds to a subcomposition of species, potentially represents a bacterial lineage of biological importance. Note that microbial community changes can occur at different levels of granularity, and finding microbial signatures at multiple granularities can both provide much insight into the underlying biology and improve prediction. For example, many studies have suggested that changes at smaller phylogenetic lineages (at the bottom of the tree) than phyla (at the top of the tree) are more relevant to obesity [Turnbaugh et al. (2008), Zhang et al. (2009, 2010), Fleissner et al. (2010)]. Generally, when the tree structure is unknown a priori, it can be learned from (sequence) data using hierarchical agglomerative clustering or other more sophisticated methods. Although next generation sequencing has allowed the rapid growth in the field of microbial ecology and the number of finished and ongoing studies, there is a serious dearth of principled statistical methods that take into account

special and inherent characteristics of microbiome data. In this paper, we look at the problem from the viewpoint of feature selection and contribute to the field by proposing a regularization method for selecting subcompositions that accounts for both the compositional nature of the data and the tree structure among the bacterial taxa. We then apply it to identify a few bacterial lineages that are predictive of BMI. Our method encourages the selection of subcompositions represented by tree nodes, and thus belongs to the class of methods that allow the identified features to reflect the structural information [see, e.g., Garcia et al. (2014), Jenatton, Audibert and Bach (2011), Jenatton et al. (2011), Kim and Xing (2009, 2012), Yuan and Lin (2006), Zhao, Rocha and Yu (2009)]. However, there are two major differences. First, we use the linear log contrast model instead of the traditional linear model, and treat subcompositions rather than individual components as features. Second, because of the distinction in the model, our methodology is more similar to the lasso than to group-lasso-type methods. Indeed, the optimization problem can be formulated as a generalized lasso problem [Tibshirani and Taylor (2011)], the solution of which can be computed efficiently using existing algorithms. In other words, the selection of subcompositions (say, groups of microbial taxa) can be achieved somewhat surprisingly through a simple lasso-type procedure. In Section 2.1, we introduce another symmetric version of the linear log contrast model. In Section 2.2, we propose a tree-structured regularization method for subcomposition selection. Computation and tuning are discussed in Section 2.3. An application to a gut microbiome data set and simulations are presented in Sections 3 and 4, respectively. Concluding remarks are made in Section 5.

2. Methodology.

2.1. *Another equivalent model.* To avoid imposing the zero-sum constraint explicitly on the coefficients, we define the centered log ratio transformation

$$Z_j = \log(X_j) - \frac{1}{p}\sum_{l=1}^{p}\log(X_l), \qquad j = 1, \ldots, p,$$

and consider the following model:

$$Y = \sum_{j=1}^{p}\beta_j Z_j + \varepsilon. \tag{2.1}$$

Define

$$Z_j^{\mathcal{S}} = \log(X_j^{\mathcal{S}}) - \frac{1}{|\mathcal{S}|}\sum_{l\in\mathcal{S}}\log(X_l^{\mathcal{S}}), \qquad j \in \mathcal{S}.$$

We assume that

$$E(Y \mid \mathbf{Z}) = \sum_{j=1}^{p}\beta_j Z_j = \sum_{k=1}^{K}\sum_{j\in\mathcal{S}_k}\beta_j(k) Z_j^{\mathcal{S}_k}, \tag{2.2}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$. Note that for each $k \in \{1, \ldots, K\}$, the coefficients $\beta_j(k)$ in (2.2) are identifiable up to a common additive constant, due to the fact that $\sum_{j \in \mathcal{S}_k} Z_j^{\mathcal{S}_k} = 0$. It is thus the relative, rather than absolute, value of $\beta_j(k)$ that matters. This in turn implies that linear contrasts of $\beta_1(k), \ldots, \beta_{|\mathcal{S}_k|}(k)$, such as $\beta_j^*(k) = \beta_j(k) - \sum_{l \in \mathcal{S}_k} \beta_l(k)/|\mathcal{S}_k|$, $j = 1, \ldots, |\mathcal{S}_k|$, are estimable.

Let $\beta_j^* = \beta_j - \sum_{l=1}^p \beta_l/p$, $j = 1, \ldots, p$. Then

$$E(Y \mid \mathbf{Z}) = \sum_{j=1}^{p} \beta_j^* \log(X_j) = \sum_{k=1}^{K} \sum_{j \in \mathcal{S}_k} \beta_j^*(k) \log\bigl(X_j^{\mathcal{S}_k}\bigr). \tag{2.3}$$

Hence, model (1.2) with (1.4) and model (2.1) with (2.2) are equivalent by setting $\alpha_j = \beta_j^*$ and $\alpha_j(k) = \beta_j^*(k)$. Both models appear to be natural with penalization, but as we will see, the latter also turns out to be computationally convenient.

It is standard to check the inactivity of a subcomposition or a subset of components, since the corresponding hypothesis is a linear hypothesis within a linear model. As a result, one way to identify $\mathcal{S}_{K+1}^c$ is to incorporate the test into a subset search procedure, for example, a stepwise backward or forward search. However, best subset selection is not only computationally intensive but may also be unsatisfactory in terms of stability [Breiman (1995)]. Alternatively, regularization methods, such as the lasso [Lin et al. (2014)], provide a direct estimate of $\mathcal{S}_{K+1}^c$, and are applicable in high dimensions. Subcomposition selection, on the other hand, seeks not only $\mathcal{S}_{K+1}^c$, but also subcompositions formed from it, removing possible redundancy in it. To understand this, we note that, by (2.3), $E(Y \mid \mathbf{Z}) = \sum_{j \in \mathcal{S}_{K+1}^c} \tilde{\beta}_j \log(X_j^{\mathcal{S}_{K+1}^c})$ for some $\tilde{\beta}_j$ such that $\sum_{j \in \mathcal{S}_{K+1}^c} \tilde{\beta}_j = 0$. Hence, for component selection, the true number of free parameters is $p - |\mathcal{S}_{K+1}| - 1$. On the other hand, the true number of free parameters for subcomposition selection is $p - |\mathcal{S}_{K+1}| - K$, due to K constraints $\sum_{j \in \mathcal{S}_k} \beta_j^*(k) = 0, k = 1, \ldots, K$. Therefore, subcomposition selection should yield better performance than component selection when $K > 1$.

2.2. *Tree-structured subcomposition selection.* As mentioned in the Introduction, the problem of subcomposition selection is computationally intensive, and moreover, its solution may not be unique. One way to deal with these issues is to restrict our attention to a smaller set of solutions. Since subcompositions represent subsets of components, we are interested in identifying subcompositions composed of closely related components. In this paper, we use a tree to represent the closeness or homogeneity among components. We elaborate on this in the following.

Suppose that the relationships among the p components can be represented as a tree T with the set of nodes $\mathcal{V}$. In this tree, there are p leaf nodes, one for each component, and there are $|\mathcal{V}| - p$ internal nodes indicating groups of components

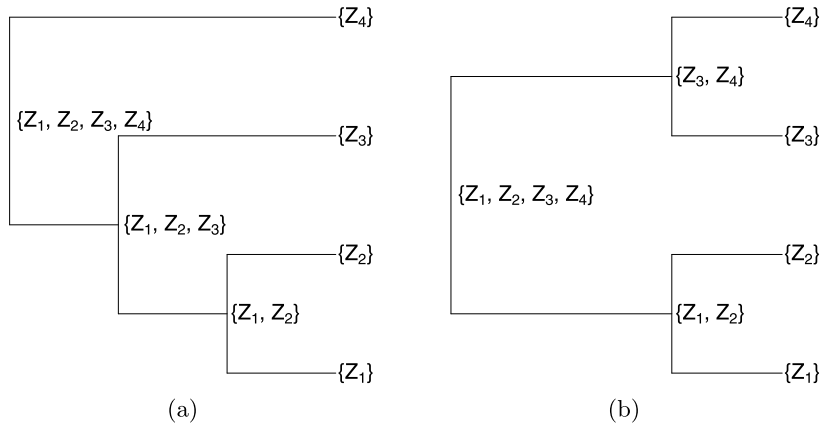

FIG. 1. *Two binary trees, each with four leaf nodes and three internal nodes. In either tree, each node is associated with a subset of components.*

at different levels. In the context of microbiome data, the evolutionary relationships among OTUs is encoded by a phylogenetic tree, where leaf nodes correspond to OTUs and internal nodes represent bacterial taxa at multiple taxonomic levels. Figure 1 shows two binary trees, each with four leaf nodes and three internal nodes. While an unbalanced tree, like the one shown in Figure 1(a), has some long paths and some short ones, a balanced tree, like the one shown in Figure 1(b), has all leaf nodes at the same depth. Throughout, we assume that the tree structure is available as prior knowledge.

Let v_{root} denote the root node of T. For each node $v \in \mathcal{V}$, let T_v denote the subtree rooted at v. Clearly, for an internal node v near the bottom of T, the components that correspond to the leaf nodes of T_v are highly homogenous, whereas for v near v_{root}, the components associated with T_v are relatively more heterogenous. Consider now an arbitrary subcomposition. Because balanced trees are extremely rare, it is very likely that the components of this subcomposition are heterogenous. To encourage the selection of homogenous subcompositions (e.g., bacterial lineages near the bottom of the phylogenetic tree), we develop a novel tree-structured regularization method as follows.

Let $\boldsymbol{e}_j$ be the p-dimensional vector whose jth element is 1 and other elements are 0, for $j = 1, \ldots, p$. For each node $v \in \mathcal{V}$, we define

$$\boldsymbol{f}_v = \sum_{j \in \mathcal{L}_v} \boldsymbol{e}_j,$$

where $\mathcal{L}_v \subseteq \{1, \ldots, p\}$ is the set of indices of the components that correspond to the leaf nodes of T_v. Denote by $\mathcal{L}$ the set of leaf nodes and $\mathcal{I}$ the set of internal nodes of T. For each $v \in \mathcal{I}$, $\boldsymbol{f}_v$ represents a tree-node-based group of components.

Let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\top$. Our tree-guided regularization penalty is defined as

$$
\begin{aligned}
J^*(\boldsymbol{\beta}^*; \lambda_1, \lambda_2) &= \lambda_1 \sum_{v \in \mathcal{L}} |\boldsymbol{f}_v^\top \boldsymbol{\beta}^*| + \lambda_2 \sum_{v \in \mathcal{I}} |\boldsymbol{f}_v^\top \boldsymbol{\beta}^*| \\
&= \lambda_1 \sum_{j=1}^{p} |\beta_j^*| + \lambda_2 \sum_{v \in \mathcal{I}} |\boldsymbol{f}_v^\top \boldsymbol{\beta}^*|.
\end{aligned} \tag{2.4}
$$

If $\boldsymbol{f}_v^\top \boldsymbol{\beta}^* = 0$ for some leaf node $v \in \mathcal{L}$, then the corresponding component is removed from the model. On the other hand, if $\boldsymbol{f}_v^\top \boldsymbol{\beta}^* = 0$ for some internal node $v \in \mathcal{I}$, then a partition occurs at v. As we move from leaf nodes to the root, the first time $\boldsymbol{f}_v^\top \boldsymbol{\beta}^* = 0$ happens at an internal node v, this defines a subcomposition. Therefore, the first term in (2.4) is for component selection, while the second term is for subcomposition selection that induces homogeneity at the subtree level. Note that it is the singularity of the l_1 norm that results in the selection of components and subcompositions.

We now return to the toy example. For illustration, we assume that each component corresponds to an OTU, and a phylogenetic tree over the eight OTUs is available. We also assume that the first four OTUs are biologically important, and the subcomposition $\{1, 2, 3, 4\}$ is associated with a subtree shown in Figure 1(b). If $\beta_j^* = 0$ for all $j > 4$, then the subcomposition $\{5, 6, 7, 8\}$ is correctly set to be inactive. We then concentrate on the subcomposition $\{1, 2, 3, 4\}$ and the corresponding subtree. If, in addition, $\boldsymbol{f}_v^\top \boldsymbol{\beta}^* = 0$ for all three internal nodes v of this subtree, then $\{1, 2, 3, 4\}$ is partitioned into $\{1, 2\}$ and $\{3, 4\}$. As we mentioned before, an alternative partition of $\{1, 2, 3, 4\}$ is given by $\{1, 4\}$ and $\{2, 3\}$. However, this partition is unlikely, especially when λ_2 is large, because the penalty on it is much larger than that for the former one. Also, the first partition yields two bacterial taxa, and hence is biologically more interpretable than the second one.

Suppose we have a sample of n observations $\{(x_{i1}, \ldots, x_{ip})^\top, y_i\}$ for $i = 1, \ldots, n$. To select subcompositions and estimate parameters simultaneously, we consider the convex optimization problem:

$$
\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2n} \|\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2 + J^*(\boldsymbol{\beta}^*; \lambda_1, \lambda_2), \tag{2.5}
$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, $\mathbf{Z} = (z_{ij}) \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, and $\|\cdot\|_2$ denotes the usual l_2 norm. We call our method Tree-guided Automatic Subcomposition Selection Operator (TASSO). When $\lambda_2 = 0$, TASSO reduces to the lasso method for component selection [Lin et al. (2014)]. Throughout the paper, TASSO is reserved for $\lambda_2 > 0$ to remind ourselves that it incorporates the tree information.

2.3. *Computation and tuning.* Denote by $\mathbf{P}_1$ the centering matrix of size p. Since $\boldsymbol{\beta}^* = \mathbf{P}_1 \boldsymbol{\beta}$ and $\boldsymbol{f}_{v_{\text{root}}}^\top \boldsymbol{\beta}^* = 0$, we can write

$$
J^*(\boldsymbol{\beta}^*; \lambda_1, \lambda_2) = J(\boldsymbol{\beta}; \lambda_1, \lambda_2) = \lambda_1 \sum_{j=1}^{p} |\tilde{\boldsymbol{e}}_j^\top \boldsymbol{\beta}| + \lambda_2 \sum_{v \in \mathcal{I}_1} |\tilde{\boldsymbol{f}}_v^\top \boldsymbol{\beta}|, \tag{2.6}
$$

where $\tilde{\boldsymbol{e}}_j = \mathbf{P}_1 \boldsymbol{e}_j$, $\tilde{\boldsymbol{f}}_v = \mathbf{P}_1 \boldsymbol{f}_v$, and $\mathcal{I}_1 = \mathcal{I} \setminus \{v_{\text{root}}\}$. Let $\mathbf{P}_2 = (\mathbf{P}_1 \boldsymbol{f}_v, v \in \mathcal{I}_1)^\top \in \mathbb{R}^{|\mathcal{I}_1| \times p}$. In matrix notation,

$$J(\boldsymbol{\beta}; \lambda_1, \lambda_2) = \lambda_1 \|\mathbf{P}_1 \boldsymbol{\beta}\|_1 + \lambda_2 \|\mathbf{P}_2 \boldsymbol{\beta}\|_1, \tag{2.7}$$

where $\|\cdot\|_1$ denotes the l_1 norm. The penalty matrix $\mathbf{P}_1$ is independent of the tree structure. For example, when $p = 4$,

$$\mathbf{P}_1 = \begin{pmatrix} 1 - 1/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 1 - 1/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1 - 1/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 1 - 1/4 \end{pmatrix} \in \mathbb{R}^{4 \times 4}.$$

By construction, $\mathbf{P}_2$ depends on the tree topology. For the tree in Figure 1(a),

$$\mathbf{P}_2 = \begin{pmatrix} 1 - 2/4 & 1 - 2/4 & -2/4 & -2/4 \\ 1 - 3/4 & 1 - 3/4 & 1 - 3/4 & -3/4 \end{pmatrix} \in \mathbb{R}^{2 \times 4},$$

and for the tree in Figure 1(b),

$$\mathbf{P}_2 = \begin{pmatrix} 1 - 2/4 & 1 - 2/4 & -2/4 & -2/4 \\ -2/4 & -2/4 & 1 - 2/4 & 1 - 2/4 \end{pmatrix} \in \mathbb{R}^{2 \times 4}.$$

Let $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$. We can further write

$$\lambda_1 \|\mathbf{P}_1 \boldsymbol{\beta}\|_1 + \lambda_2 \|\mathbf{P}_2 \boldsymbol{\beta}\|_1 = \lambda \{(1 - \alpha) \|\mathbf{P}_1 \boldsymbol{\beta}\|_1 + \alpha \|\mathbf{P}_2 \boldsymbol{\beta}\|_1\} = \lambda \|\tilde{\mathbf{D}}(\alpha) \boldsymbol{\beta}\|_1,$$

where $\tilde{\mathbf{D}}(\alpha) = \{(1 - \alpha)\mathbf{P}_1, \alpha \mathbf{P}_2^\top\}^\top \in \mathbb{R}^{(|\mathcal{V}| - 1) \times p}$. We then arrive at the optimization problem:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2n} \|\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2 + \lambda \|\tilde{\mathbf{D}}(\alpha) \boldsymbol{\beta}\|_1. \tag{2.8}$$

Thus, for each α, we need to solve a generalized lasso problem [Tibshirani and Taylor (2011)]. Note, however, that the zero-sum constraint on rows of $\mathbf{Z}$ makes the elements of $\boldsymbol{\beta}$ not identifiable. Since the penalty term depends only on the relative values of the coefficients, one can choose the jth component as the reference, that is, set $\beta_j = 0$, and solve the resulting problem. Nevertheless, this strategy has an undesirable aspect: the numerical solution depends on the choice of j. Following Lin et al. (2014), an alternative approach is to impose a zero-sum constraint on the coefficients, and then solve a constrained optimization problem. However, because $\tilde{\mathbf{D}}(\alpha)$ depends on the tree structure, it is difficult to develop an efficient algorithm, unless $\alpha = 0$.

Since the zero-sum constraint implies the exact collinearity among the columns of $\mathbf{Z}$, a natural way of coping with perfectly correlated covariates is to use an l_2 or ridge penalty [Hoerl and Kennard (1981)]. We thus consider a modified criterion:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2n} \|\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2 + \lambda \|\tilde{\mathbf{D}}(\alpha) \boldsymbol{\beta}\|_1 + \frac{\gamma}{2n} \|\boldsymbol{\beta}\|_2^2, \tag{2.9}$$

where $\gamma > 0$ is the ridge parameter.

Let $\tilde{\mathbf{y}} = (\mathbf{y}, 0, \ldots, 0) \in \mathbb{R}^{n+p}$ and $\tilde{\mathbf{Z}}(\gamma) = (\mathbf{Z}^\top, \sqrt{\gamma}\mathbf{I}_p)^\top \in \mathbb{R}^{(n+p)\times p}$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. We now write (2.9) as

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \frac{1}{2n}\|\tilde{\mathbf{y}} - \tilde{\mathbf{Z}}(\gamma)\boldsymbol{\beta}\|_2^2 + \lambda\|\tilde{\mathbf{D}}(\alpha)\boldsymbol{\beta}\|_1. \tag{2.10}$$

For each pair of α and γ, this is a standard generalized lasso problem on augmented data. To this end, we can use the dual path algorithm that is efficiently implemented in the **genlasso** R package. We denote the solution for $\boldsymbol{\beta}$ by $\boldsymbol{\beta}(\alpha, \lambda, \gamma)$.

To select the tuning parameters, we propose to use an information criterion. Define

$$\mathrm{IC}(\alpha, \lambda, \gamma) = \log\{\mathrm{RSS}(\alpha, \lambda, \gamma)\} + \kappa \times \mathrm{DF}(\alpha, \lambda, \gamma), \tag{2.11}$$

where $\mathrm{RSS}(\alpha, \lambda, \gamma) = \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}(\alpha, \lambda, \gamma)\|_2^2$, κ is a complexity factor, and $\mathrm{DF}(\alpha, \lambda, \gamma)$ denotes the effective number of parameters in $\boldsymbol{\beta}(\alpha, \lambda, \gamma)$, which is an estimate of the degrees of freedom of the fit. Three popular choices of κ are $\kappa = 2$, $\kappa = \log(n)$, and $\kappa = \log\{\log(n)\} \times \log(n)$. The corresponding criteria are known as Akaike information criterion [AIC, Akaike (1998)], Bayesian information criterion [BIC, Schwarz (1978)], and generalized information criterion [GIC, Fan and Tang (2013)], respectively. For each (α, γ), we select λ by minimizing $\mathrm{IC}(\alpha, \lambda, \gamma)$ along the path. In this paper, we concentrate on two instances of our method, namely, $\alpha = 0$ (the lasso) and $\alpha = 0.5$ (TASSO). Furthermore, for each method, we explore three values of γ: 10^{-6}, 10^{-4}, and 10^{-2}, and choose γ to be the one that gives the smallest criterion value.

We note that the results generally depend on the specific criterion used. An alternative to information criteria is the cross-validation approach. However, cross validation is computationally more intensive. Furthermore, its performance is often similar to that of AIC.

3. Gut microbiome data. The human gut carries a vast and diverse microbial ecosystem that is essential for human health [Gill et al. (2006)]. For example, studies have shown that the gut microbiome is likely to be implicated in the etiopathogenesis of obesity [Turnbaugh et al. (2006), Ley (2010)]. In this section, we apply our method to a human gut microbiome study conducted at the University of Pennsylvania [Wu et al. (2011)], in which both gut microbiome data and clinical measurements were available. Our goal is to predict obesity based on the gut microbiome composition. Specifically, we are interested in identifying a few bacterial lineages that are predictive of BMI, a widely accepted index of obesity and overweight.

Stool samples of 98 healthy volunteers were collected in this study, and bacterial DNA was extracted and then analyzed by the 454/Roche pyrosequencing of 16S rRNA gene segments of the V1–V2 region. The pyrosequences were processed by the QIIME pipeline [Caporaso et al. (2010)] with the default parameters. The counts for more than 17,000 species-level OTUs were obtained. One way of

reducing the number of OTUs is to combine them at the genus level. However, for this data set more than 25% of the total OTU counts would be discarded, which is likely to result in a biased analysis. Since we have available a phylogenetic tree of all the OTUs, an alternative is to create a nontrivial set of OTUs by agglomerating closely related OTUs using, for example, the single-linkage clustering: all leaf nodes of the tree separated by a cophenetic distance smaller than some threshold will be agglomerated into one OTU [McMurdie and Holmes (2013)]. In our analysis, we used the threshold 0.5. For each merged group of OTUs, we chose the OTU with the highest abundance to represent it. Moreover, we excluded the uncommon OTUs that occurred in less than 5 of the samples, leaving 62 OTUs. The new phylogenetic tree T is shown in Figure 2.

Since the number of sequencing reads varied drastically across samples and should not play a key role in predicting BMI [Lin et al. (2014)], we applied the centered log ratio transformation after replacing zero counts by the maximum rounding error 0.5 [Aitchison (1986), Section 11.5]. The final data set was composed of a matrix $\mathbf{Z} \in \mathbb{R}^{98\times 62}$ of log contrasts, a phylogenetic tree T, and a vector $\boldsymbol{y} \in \mathbb{R}^{98}$ of BMI values. Table 1 shows a summary of the selected models for the lasso ($\alpha = 0$) and TASSO ($\alpha = 0.5$). We can see that for the lasso, AIC selected a subcomposition of 13 components, and was the only criterion that worked, while for TASSO, AIC, and BIC picked the same model with five two-component subcompositions, and GIC selected only one subcomposition of size two.

To obtain stable selection results, we generated 1000 random subsamples containing 80% of the data, and used the AIC criterion to select the model. As illustrated in the Introduction, the lasso selects only a single subcomposition of selected components. To make feasible the comparison between our method and the lasso, we concentrated on subcompositions of size two, that is, component pairs, that appeared in at least 900 of the models. In other words, if a model included a subcomposition of size 3, then it automatically included three subcompositions of size 2. The results are reported in Table 2. We can see that 4 out of $\binom{13}{2}$ and 4 out of 5 component pairs, respectively, for the lasso and TASSO, had selection probabilities greater than 0.9. However, in terms of log contrasts there was a redundant pair for the lasso, namely, $Z_{36} - Z_{60} = (Z_{27} - Z_{36}) - (Z_{27} - Z_{60})$. For each method, we then refit a linear log contrast model based on the top four component pairs. The results are summarized in Table 3. We see that the use of the tree structure doubled the (adjusted) R^2 value. Note that, although TASSO identified only pairs of OTUs on the full data (see Table 1), it selected subcompositions of different size when applied to random subsamples.

For the three nonredundant log contrasts selected by the lasso, none of their regression coefficients was significant at the 0.05 level. However, two of the four log contrasts selected by TASSO, $Z_{26} - Z_{27}$ and $Z_{59} - Z_{60}$, had their coefficients highly significant at the 0.01 level. The four components, Z_{26}, Z_{27}, Z_{59} and Z_{60}, belonged to the phylum Firmicutes. Specifically, Z_{26} and Z_{27} were members of

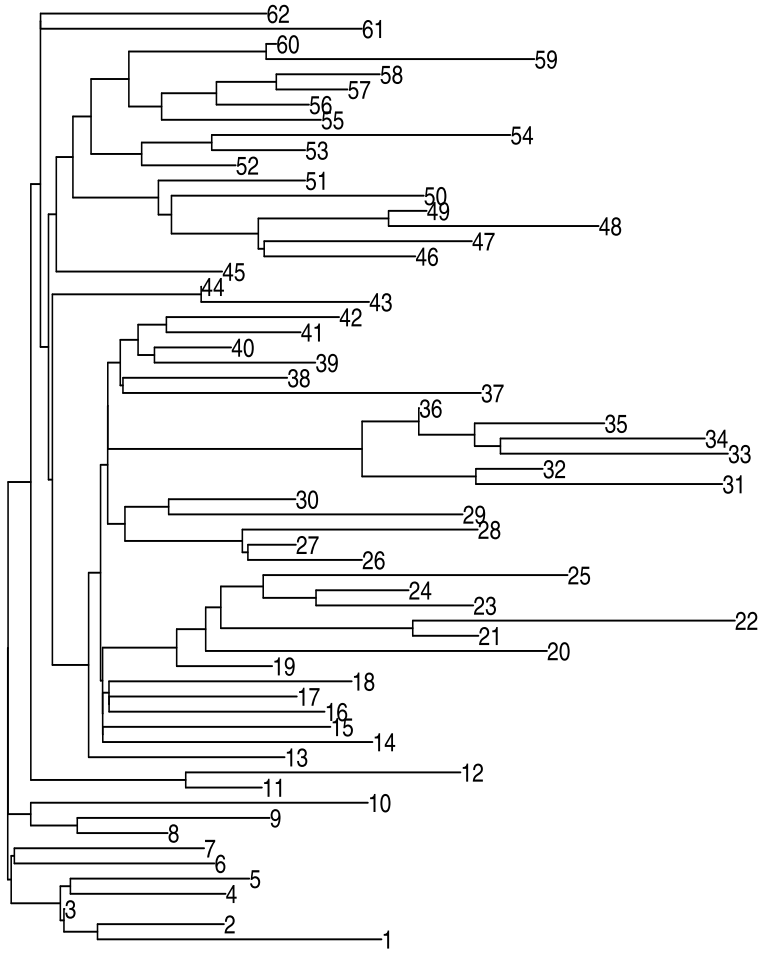

FIG. 2. *Gut microbiome data. The phylogenetic tree T. In this tree, there are* 62 *leaf nodes, labeled as* $1, \ldots, 62$*, one for each OTU, and there are* 61 *internal nodes* (*labels not shown*) *representing microbial taxa at different levels.*

TABLE 1
Gut microbiome data. A summary of the selected models from different criteria

Criterion	**Method**	**Subcompositions**
AIC	lasso	$\{Z_6, Z_{10}, Z_{11}, Z_{18}, Z_{19}, Z_{20}, Z_{27}, Z_{32}, Z_{36}, Z_{46}, Z_{51}, Z_{59}, Z_{60}\}$
	TASSO	$\{Z_{11}, Z_{12}\}, \{Z_{19}, Z_{20}\}, \{Z_{26}, Z_{27}\}, \{Z_{32}, Z_{36}\}, \{Z_{59}, Z_{60}\}$
BIC	lasso	$\varnothing$
	TASSO	$\{Z_{11}, Z_{12}\}, \{Z_{19}, Z_{20}\}, \{Z_{26}, Z_{27}\}, \{Z_{32}, Z_{36}\}, \{Z_{59}, Z_{60}\}$
GIC	lasso	$\varnothing$
	TASSO	$\{Z_{59}, Z_{60}\}$

TABLE 2
Gut microbiome data. Selection frequencies by AIC based on 1000 *random draws of* 80% *of the data*

	Component pair	Count
lasso	$\{Z_{11}, Z_{27}\}$	901
	$\{Z_{27}, Z_{36}\}$	903
	$\{Z_{27}, Z_{60}\}$	951
	$\{Z_{36}, Z_{60}\}$	903
TASSO	$\{Z_{11}, Z_{12}\}$	907
	$\{Z_{19}, Z_{20}\}$	948
	$\{Z_{26}, Z_{27}\}$	946
	$\{Z_{59}, Z_{60}\}$	973

the Veillonellaceae family, and were unclassified at the genus level, while Z_{59} and Z_{60} were members of the Erysipelotrichaceae family, with Z_{60} being unclassified at the genus level. Wu et al. (2011) showed that the Veillonellaceae was positively correlated to BMI, and Zhang et al. (2009) reported that higher proportions of the Erysipelotrichaceae were identified in morbidly obese individuals. The Erysipelotrichaceae has also been shown by several independent studies to alter in abundance in response to changes in the amount of dietary fat. For example, a bloom occurred for an uncultured member of this family, after inducing obesity in mice by feeding them a "Western" diet [Turnbaugh et al. (2008), Fleissner et al. (2010)], and four clades of this family were reported to react differently to high-fat and low-fat diets [Zhang et al. (2010)]. Although the exact role of these two bacterial families in host energy metabolism is still obscure, and deserves further research, these studies suggested that changes at much smaller phylogenetic lineages than phyla are more relevant to obesity, and our method is potentially useful in this regard. In the literature, several methods have been proposed to analyze the above data set, and they fall roughly into two categories. The first category addresses the compositional nature of microbiome data; see for example, Lin et al. (2014). The second category accounts for the phylogenetic tree over bacterial taxa; see, for example, Chen et al. (2013). As pointed out by Li (2015), methods for analyzing microbiome data must take into account the compositional nature of the

TABLE 3
Gut microbiome data. Model fits based on the top four component pairs

	Log contrasts	R^2	Adjusted R^2
lasso	$Z_{11} - Z_{27}, Z_{27} - Z_{36}, Z_{27} - Z_{60}$	0.14	0.11
TASSO	$Z_{11} - Z_{12}, Z_{19} - Z_{20}, Z_{26} - Z_{27}, Z_{59} - Z_{60}$	0.26	0.23

data. Also, how to incorporate the phylogenetic tree information is an interesting research question. Here, we provide a somewhat unified solution in the regression setting: our method satisfies the subcompositional coherence principle, and is able to select subcompositions automatically at subtree levels.

4. Simulations. In this section, we conducted a simulation study to examine the performance of the proposed method. To mimic the type of tree structure that we might see in real problems, we used the phylogenetic tree T from the gut microbiome data described in the previous section, which has $p = 62$ leaf nodes. We then generated the compositional data matrix $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times p}$ and the response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ with $n = 100$ as follows. First, we simulated a data matrix $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n\times p}$ from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{ij}) \in \mathbb{R}^{p\times p}$. We then transformed $\mathbf{W}$ to $\mathbf{X}$ by setting $x_{ij} = \exp(w_{ij})/\sum_{l=1}^{p} \exp(w_{il})$. Each row of $\mathbf{X}$ has a logistic normal distribution. To allow component proportions to differ by orders of magnitude, as was the case for the gut microbiome data, we set $\mu_j = \log(p/2)$ for $j = 1, \ldots, 5$ and $\mu_j = 0$ otherwise. To describe different levels of correlations among the components, we took $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.2$ or 0.5. Finally, we applied the centered log ratio transformation $z_{ij} = \log(x_{ij}) - \sum_{l=1}^{p} \log(x_{il})/p$, and simulated the responses y_i from the model

$$y_i = \sum_{j=1}^{p} z_{ij}\beta_j^* + \varepsilon_i,$$

where ε_i are independent and normally distributed with mean zero and variance σ^2, with $\sigma = 0.5$ or 1. To specify the coefficient vector $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\top \in \mathbb{R}^p$, we considered three cases:

$$\text{(I)} \qquad \boldsymbol{\beta}^* = (1, -1, 0, 0.8, -0.8, 0, 0, -1.5, -0.5, 2, 0, \ldots, 0)^\top,$$

$$\text{(II)} \qquad \boldsymbol{\beta}^* = (1, -1, 0.8, -0.8, 0, 0, 0, -1.5, -0.5, 2, 0, \ldots, 0)^\top,$$

and

$$\text{(III)} \qquad \boldsymbol{\beta}^* = (1, -0.8, 0.6, 0, 0, 0, 0, -1.5, -0.5, 1.2, 0, \ldots, 0)^\top.$$

Case (I) implies three subcompositions, $\{1, 2\}$, $\{4, 5\}$, and $\{8, 9, 10\}$, all consistent with the tree structure. Case (II) also implies three subcompositions $\{1, 2\}$, $\{3, 4\}$, and $\{8, 9, 10\}$. However, $\{3, 4\}$ is not associated with an internal node. Case (III) determines one single subcomposition $\{1, 2, 3, 8, 9, 10\}$, and so requires only component selection. In summary, the structure of the phylogenetic tree T perfectly matches the subcompositions in case (I), partly matches the subcompositions in case (II), and is completely misleading in case (III).

To evaluate the performance, we used the following measures: (1) the l_2 loss $\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\|_2$, where $\hat{\boldsymbol{\beta}}^* = \mathbf{P}_1\hat{\boldsymbol{\beta}}$, (2) the relative model error $\|\mathbf{Z}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*)\|_2^2/(n\sigma^2)$,

(3) the true positive rate, (4) the false positive rate, (5) the effective number of parameters, that is, the estimated degrees of freedom, and (6) the number of subcompositions. For each configuration in each case, we simulated 200 data sets. The results are summarized in Tables 4 and 5. Several conclusions can be made as follows.

First, the BIC criterion generally performed better than AIC and GIC. For the lasso ($\alpha = 0$), AIC over-selected components, and for TASSO ($\alpha = 0.5$), AIC over-selected subcompositions. While GIC competed well with BIC for TASSO, it performed poorly for the lasso when $\rho = 0.5$ and $\sigma = 1$, with a very low true positive rate. Second, TASSO performed consistently better than the lasso in terms of the l_2 loss and the relative model error. As expected, the use of tree information in case (I) substantially enhanced the performance. The improvement in case (II) when the tree structure partly matches the subcompositions, indicates that TASSO enjoyed a certain degree of robustness relative to possibly misspecified subcompositions. One explanation for the success of TASSO in case (III) was that the cost of falsely selecting subcompositions was compensated by the parsimonious use of degrees of freedom. As we can see, TASSO appeared to have a smaller effective number of parameters than the lasso did, as was especially the case for BIC. A possible reason is that, like the elastic net method of coupling l_1 and l_2 regularizations [Zou and Hastie (2005)], TASSO has two penalty terms, a lasso penalty plus a tree-guided penalty, and this second penalty alleviates the instability of the lasso. Finally, as we increased σ from 0.5 to 1, the performance of both methods deteriorated.

5. Discussion. Next generation sequencing (e.g., 454 pyrosequencing and Illumina shotgun sequencing), which is becoming cheaper and faster, has allowed much larger surveys of microbial communities, with more reads in total. However, our statistical methods for extracting useful information from microbiome studies have not been developed as quickly as experimental techniques. In particular, there is a serious dearth of principled tools that can account for special and inherent features of microbiome data. In this paper, we have considered the problem of subcomposition selection in regression problems with high-dimensional and compositional covariates. Rather than searching through all possible solutions, we have considered a setting where the relationships between the components can be represented as a tree, and proposed a structured regularization method to select subcompositions at subtree levels. We have demonstrated the superiority of our method over the lasso (which actually selects a single subcomposition of selected components) through an application to a human gut microbiome study and simulations. In particular, unlike with the lasso, our method identified subcompositions (composed of members of the families Veillonellaceae and Erysipelotrichaceae) that are likely to play important roles in obesity.

It is known that data from experiments with mixtures are compositional, due to the constraint on components composing the mixture. To reduce the effect of multicollinearity in the analysis of mixture experiments, St. John (1984) suggested the

TABLE 4

Means and standard deviations (*in parentheses*) *of the* l_2 *loss*, *relative model error* (*RME*), *true positive rate* (*TPR*), *false positive rate* (*FPR*), *effective number of parameters* (*ENP*), *and number of subcompositions* (*NSC*), *based on* 200 *replications*, *are reported when* $\rho = 0.2$

			l_2 loss	RME	TPR	FPR	ENP	NSC
					Case (I)			
$\sigma = 0.5$	AIC	lasso	0.39 (0.14)	0.38 (0.26)	1.00 (0.02)	0.57 (0.22)	36.60 (12.66)	1.00 (0.00)
		TASSO	0.31 (0.13)	0.26 (0.17)	1.00 (0.00)	0.45 (0.29)	29.56 (16.44)	6.49 (3.44)
	BIC	lasso	0.33 (0.11)	0.34 (0.39)	1.00 (0.02)	0.16 (0.07)	13.04 (3.90)	1.00 (0.00)
		TASSO	0.24 (0.06)	0.18 (0.09)	1.00 (0.00)	0.05 (0.04)	6.76 (2.54)	4.89 (1.84)
	GIC	lasso	0.38 (0.12)	0.43 (0.41)	1.00 (0.02)	0.11 (0.04)	10.28 (2.36)	1.00 (0.00)
		TASSO	0.26 (0.07)	0.22 (0.11)	1.00 (0.00)	0.02 (0.02)	5.23 (1.21)	3.85 (1.00)
$\sigma = 1$	AIC	lasso	0.77 (0.25)	0.36 (0.14)	1.00 (0.02)	0.57 (0.22)	36.53 (12.75)	1.00 (0.00)
		TASSO	0.61 (0.25)	0.25 (0.16)	1.00 (0.00)	0.44 (0.29)	29.04 (16.26)	6.29 (3.27)
	BIC	lasso	0.66 (0.15)	0.32 (0.13)	0.99 (0.04)	0.16 (0.07)	12.94 (3.93)	1.00 (0.00)
		TASSO	0.48 (0.12)	0.18 (0.09)	1.00 (0.00)	0.05 (0.04)	6.69 (2.54)	4.83 (1.73)
	GIC	lasso	0.75 (0.19)	0.41 (0.19)	0.99 (0.06)	0.12 (0.04)	10.13 (2.39)	1.00 (0.00)
		TASSO	0.53 (0.14)	0.22 (0.11)	1.00 (0.02)	0.02 (0.02)	5.22 (1.32)	3.86 (1.04)
					Case (II)			
$\sigma = 0.5$	AIC	lasso	0.39 (0.12)	0.36 (0.14)	1 (0)	0.57 (0.22)	36.63 (12.49)	1.00 (0.00)
		TASSO	0.34 (0.13)	0.30 (0.16)	1 (0)	0.53 (0.26)	34.03 (14.99)	5.80 (3.34)
	BIC	lasso	0.34 (0.08)	0.33 (0.13)	1 (0)	0.16 (0.06)	12.84 (3.66)	1.00 (0.00)
		TASSO	0.29 (0.07)	0.24 (0.10)	1 (0)	0.09 (0.06)	9.12 (3.47)	5.42 (2.17)
	GIC	lasso	0.38 (0.09)	0.40 (0.16)	1 (0)	0.11 (0.04)	10.54 (2.44)	1.00 (0.00)
		TASSO	0.32 (0.08)	0.30 (0.13)	1 (0)	0.06 (0.03)	7.19 (1.96)	4.38 (1.58)
$\sigma = 1$	AIC	lasso	0.77 (0.24)	0.36 (0.14)	1.00 (0.02)	0.57 (0.22)	36.63 (12.50)	1.00 (0.00)
		TASSO	0.69 (0.25)	0.30 (0.16)	1.00 (0.00)	0.53 (0.27)	34.28 (15.16)	5.69 (3.33)
	BIC	lasso	0.69 (0.16)	0.33 (0.13)	1.00 (0.03)	0.16 (0.06)	12.82 (3.67)	1.00 (0.00)
		TASSO	0.57 (0.14)	0.24 (0.10)	1.00 (0.00)	0.09 (0.06)	9.11 (3.52)	5.45 (2.08)
	GIC	lasso	0.78 (0.20)	0.42 (0.19)	0.99 (0.06)	0.11 (0.04)	10.33 (2.41)	1.00 (0.00)
		TASSO	0.63 (0.17)	0.30 (0.15)	1.00 (0.04)	0.06 (0.03)	7.12 (1.98)	4.35 (1.55)
					Case (III)			
$\sigma = 0.5$	AIC	lasso	0.37 (0.15)	0.35 (0.27)	1.00 (0.01)	0.52 (0.24)	33.98 (13.58)	1.00 (0.00)
		TASSO	0.35 (0.13)	0.31 (0.15)	1.00 (0.00)	0.52 (0.26)	34.26 (14.74)	5.75 (2.90)
	BIC	lasso	0.31 (0.11)	0.31 (0.45)	1.00 (0.03)	0.11 (0.06)	11.08 (3.37)	1.00 (0.00)
		TASSO	0.25 (0.06)	0.23 (0.10)	1.00 (0.00)	0.10 (0.06)	10.55 (3.35)	3.78 (2.11)
	GIC	lasso	0.35 (0.12)	0.40 (0.49)	1.00 (0.03)	0.07 (0.04)	8.79 (2.34)	1.00 (0.00)
		TASSO	0.27 (0.07)	0.27 (0.11)	1.00 (0.00)	0.07 (0.03)	8.84 (1.87)	2.65 (1.55)
$\sigma = 1$	AIC	lasso	0.74 (0.26)	0.34 (0.16)	1.00 (0.02)	0.52 (0.24)	34.02 (13.61)	1.00 (0.00)
		TASSO	0.68 (0.27)	0.30 (0.16)	0.99 (0.04)	0.51 (0.26)	33.28 (14.53)	6.06 (2.91)
	BIC	lasso	0.62 (0.16)	0.29 (0.16)	0.99 (0.05)	0.11 (0.06)	10.93 (3.40)	1.00 (0.00)
		TASSO	0.51 (0.13)	0.23 (0.11)	0.97 (0.08)	0.10 (0.06)	10.55 (3.25)	3.99 (2.08)
	GIC	lasso	0.72 (0.21)	0.40 (0.24)	0.97 (0.08)	0.06 (0.04)	8.43 (2.37)	1.00 (0.00)
		TASSO	0.55 (0.14)	0.28 (0.13)	0.97 (0.09)	0.07 (0.03)	8.64 (1.93)	2.67 (1.48)

TABLE 5

Means and standard deviations (in parentheses) of the l_2 loss, relative model error (RME), true positive rate (TPR), false positive rate (FPR), effective number of parameters (ENP), and number of subcompositions (NSC), based on 200 replications, are reported when $\rho = 0.5$

			l_2 loss	RME	TPR	FPR	ENP	NSC
					Case (I)			
$\sigma = 0.5$	AIC	lasso	0.47 (0.16)	0.37 (0.14)	1 (0)	0.59 (0.21)	37.35 (11.81)	1.00 (0.00)
		TASSO	0.39 (0.16)	0.27 (0.17)	1 (0)	0.46 (0.28)	30.07 (16.15)	5.76 (2.93)
	BIC	lasso	0.43 (0.11)	0.35 (0.14)	1 (0)	0.20 (0.08)	15.39 (4.35)	1.00 (0.00)
		TASSO	0.30 (0.08)	0.19 (0.09)	1 (0)	0.05 (0.04)	6.78 (2.50)	4.55 (1.68)
	GIC	lasso	0.52 (0.13)	0.48 (0.21)	1 (0)	0.14 (0.05)	11.80 (2.90)	1.00 (0.00)
		TASSO	0.33 (0.09)	0.22 (0.11)	1 (0)	0.03 (0.02)	5.45 (1.42)	3.80 (1.14)
$\sigma = 1$	AIC	lasso	0.94 (0.32)	0.37 (0.14)	1.00 (0.03)	0.58 (0.21)	37.21 (11.94)	1.00 (0.00)
		TASSO	0.77 (0.32)	0.27 (0.16)	1.00 (0.02)	0.45 (0.28)	29.40 (16.14)	5.85 (2.87)
	BIC	lasso	0.88 (0.24)	0.36 (0.16)	0.97 (0.09)	0.19 (0.08)	14.93 (4.41)	1.00 (0.00)
		TASSO	0.60 (0.16)	0.18 (0.09)	0.99 (0.05)	0.05 (0.05)	6.85 (2.73)	4.51 (1.65)
	GIC	lasso	1.10 (0.34)	0.57 (0.35)	0.88 (0.20)	0.12 (0.05)	10.32 (3.52)	1.00 (0.00)
		TASSO	0.66 (0.19)	0.23 (0.12)	0.98 (0.07)	0.02 (0.02)	5.23 (1.42)	3.71 (1.09)
					Case (II)			
$\sigma = 0.5$	AIC	lasso	0.49 (0.16)	0.39 (0.15)	1 (0)	0.62 (0.20)	39.26 (11.58)	1.00 (0.00)
		TASSO	0.43 (0.15)	0.30 (0.15)	1 (0)	0.53 (0.25)	34.36 (14.16)	5.21 (2.84)
	BIC	lasso	0.47 (0.12)	0.36 (0.14)	1 (0)	0.21 (0.08)	16.03 (4.59)	1.00 (0.00)
		TASSO	0.37 (0.09)	0.24 (0.10)	1 (0)	0.10 (0.07)	9.70 (4.16)	5.20 (2.18)
	GIC	lasso	0.56 (0.14)	0.49 (0.21)	1 (0)	0.15 (0.05)	12.52 (2.93)	1.00 (0.00)
		TASSO	0.41 (0.11)	0.30 (0.14)	1 (0)	0.05 (0.04)	7.12 (2.05)	4.18 (1.50)
$\sigma = 1$	AIC	lasso	0.99 (0.33)	0.39 (0.15)	1.00 (0.02)	0.62 (0.20)	39.10 (11.69)	1.00 (0.00)
		TASSO	0.86 (0.31)	0.31 (0.15)	1.00 (0.00)	0.53 (0.25)	34.23 (14.42)	5.27 (2.83)
	BIC	lasso	0.99 (0.31)	0.41 (0.24)	0.92 (0.18)	0.19 (0.09)	14.68 (5.29)	1.00 (0.00)
		TASSO	0.74 (0.19)	0.25 (0.11)	0.99 (0.06)	0.10 (0.07)	9.37 (4.23)	4.96 (2.06)
	GIC	lasso	1.27 (0.39)	0.67 (0.39)	0.76 (0.27)	0.11 (0.05)	9.37 (3.70)	1.00 (0.00)
		TASSO	0.85 (0.25)	0.33 (0.17)	0.95 (0.11)	0.05 (0.03)	6.62 (2.11)	3.99 (1.39)
					Case (III)			
$\sigma = 0.5$	AIC	lasso	0.49 (0.24)	0.57 (2.01)	0.99 (0.06)	0.56 (0.23)	36.40 (13.10)	1.00 (0.00)
		TASSO	0.43 (0.16)	0.32 (0.16)	1.00 (0.00)	0.52 (0.26)	34.01 (14.57)	5.22 (2.58)
	BIC	lasso	0.44 (0.20)	0.64 (3.11)	0.99 (0.08)	0.15 (0.08)	13.27 (4.36)	1.00 (0.00)
		TASSO	0.32 (0.08)	0.22 (0.09)	1.00 (0.00)	0.09 (0.06)	10.27 (3.46)	3.36 (1.85)
	GIC	lasso	0.52 (0.23)	0.82 (3.84)	0.99 (0.10)	0.10 (0.05)	10.33 (2.81)	1.00 (0.07)
		TASSO	0.34 (0.08)	0.26 (0.11)	1.00 (0.00)	0.06 (0.04)	8.63 (1.97)	2.43 (1.39)
$\sigma = 1$	AIC	lasso	0.94 (0.35)	0.39 (0.36)	0.99 (0.05)	0.56 (0.24)	36.23 (13.21)	1.00 (0.00)
		TASSO	0.83 (0.31)	0.30 (0.15)	0.99 (0.05)	0.50 (0.25)	32.81 (14.22)	5.43 (2.47)
	BIC	lasso	0.90 (0.27)	0.41 (0.59)	0.91 (0.15)	0.12 (0.07)	11.14 (4.53)	1.00 (0.00)
		TASSO	0.63 (0.16)	0.22 (0.10)	0.97 (0.07)	0.09 (0.05)	9.86 (3.05)	3.38 (1.81)
	GIC	lasso	1.11 (0.30)	0.61 (0.76)	0.78 (0.20)	0.05 (0.04)	6.91 (3.06)	1.00 (0.07)
		TASSO	0.68 (0.17)	0.26 (0.11)	0.97 (0.08)	0.06 (0.03)	8.30 (1.91)	2.42 (1.40)

use of ordinary ridge regression estimator as a means for stabilizing the coefficient estimates in the fitted model. In this paper, we take an alternative strategy by using variable transformation. Specifically, we adopt the linear log contrast model instead of the linear model, and focus on the selection of subcompositions. Since the linear log contrast model guarantees the subcompositional coherence and our tree-guided regularization penalty preserves this property, the proposed procedure satisfies the subcompositional coherence principle.

To analyze compositional data, alternative transformations to the log transformation are available, such as the square-root transformation [Scealy and Welsh (2011)] and the relative power transformation [Scealy et al. (2015)]. Nevertheless, each of these transformations has its own merits and drawbacks. We choose the log transformation mainly because the resulting linear log contrast model preserves the subcompositional coherence principle and facilitates the selection of subcompositions.

Recently, Garcia et al. (2014) proposed a regularization approach for identifying important regressor groups, subgroups, and individuals, and they applied it to a microbiome data set. Their method was mathematically and statistically sound, but it was for regression analysis with regular (yet not compositional) covariates. Specifically, in the application, the compositional nature of the microbiome data was not accounted for. As we mentioned in the Introduction, taking into account the unit-sum constraint makes our method very different from that of Garcia et al. (2014) and others in the literature.

Note that our goal here is to select subcompositions at subtree levels (i.e., groups of bacterial taxa at different taxonomic levels). Related to the present work is Shi, Zhang and Li (2016), who considered the problem of selecting subcompositions at a fixed taxonomic level (e.g., groups of species under a given genus or phylum). Assuming that a grouping of bacterial taxa is available, the work of Shi, Zhang and Li (2016) is a direct extension of Lin et al. (2014). However, our extension in methodology is original: our method is based on a novel penalty function that incorporates the topology of the phylogenetic tree node-by-node, thus encourages the selection of subcompositions at subtree levels. In practice, classifying the microbes into different taxonomies, from phylum to species level, necessitates the existence of a reference database that is often incomplete, because the vast majority of microbes have not yet been formally described. In contrast, a phylogenetic tree can always be learned from molecular sequences.

In this paper, the observed compositional covariates are assumed to lie in a strictly positive simplex. One reason is that we cannot take logarithms of zero in the linear log contrast model. In the absence of a one-to-one monotonic transformation between the real line and its nonnegative subset, the problem of zeros might not be satisfactorily resolved, and solutions generally depend on the frequency and nature of the zeros. There are two types of zeros: sampling zeros and structural zeros. In this paper, we assume implicitly that the zero does not denote the part that is completely absent but rather denotes that part that no quantifiable proportion

could be recorded to the accuracy or rounding of the measurement process. In this case, replacing zero counts by the maximum rounding error is a commonly used strategy in the literature [Aitchison (1986), Lin et al. (2014)]. Clearly, new statistical methods are needed to model "sparse" compositional data that contain many zeros and to differentiate between sampling zeros and structural zeros [Li (2015)].

The properties of the generalized lasso apply to our method, because the latter can be encapsulated by the former; see, for example, Lee, Sun and Taylor (2015). Throughout this paper, we have assumed that the tree structure is available as prior knowledge. It is a topic of our future research to explore ways to estimate tree structures completely based on data. Alternatively, it is of interest to study the theoretical behavior of our method when the prior information is not completely accurate or even erroneous. Finally, our method takes advantage of the tree topology, but not branch lengths. Since the tree-guided regularization penalty contains internal-node-based terms, one possible way of accounting for branch lengths is to first summarize such information at the internal node (i.e., subtree) level, and then incorporate the summary statistics into our penalty as weights.

Acknowledgments. We are grateful to the Editor, the Associate Editor, and three anonymous referees for their helpful comments. We thank Hongzhe Li and Jun Chen for providing the data.

REFERENCES

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London. MR0865647

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

AKAIKE, H. (1998). *Selected Papers of Hirotugu Akaike*. Springer, New York. MR1486823

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720

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

CHEN, J., BUSHMAN, F. D., LEWIS, J. D., WU, G. D. and LI, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostat.* **14** 244–258.

FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. MR3065478

FLEISSNER, C. K., HUEBEL, N., ABD EL-BARY, M. M., LOH, G., KLAUS, S. and BLAUT, M. (2010). Absence of intestinal microbiota does not protect mice from diet-induced obesity. *Br. J. Nutr.* **104** 919–929.

GARCIA, T. P., MÜLLER, S., CARROLL, R. J. and WALZEM, R. L. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: Application to gut microbiome data. *Bioinformatics* **30** 831–837.

GILL, S. R., POP, M., DEBOY, R. T., ECKBURG, P. B., TURNBAUGH, P. J., SAMUEL, B. S., GORDON, J. I., RELMAN, D. A., FRASER-LIGGETT, C. M. and NELSON, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312** 1355–1359.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

JENATTON, R., AUDIBERT, J.-Y. and BACH, F. (2011). Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.* **12** 2777–2824. MR2854347

JENATTON, R., MAIRAL, J., OBOZINSKI, G. and BACH, F. (2011). Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.* **12** 2297–2334. MR2825428

KIM, S. and XING, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet.* **5** e1000587.

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL mapping. *Ann. Appl. Stat.* **6** 1095–1117. MR3012522

KNIGHTS, D., PARFREY, L. W., ZANEVELD, J., LOZUPONE, C. and KNIGHT, R. (2011). Human-associated microbial signatures: Examining their predictive value. *Cell Host & Microbe* **10** 292–296.

LEE, J. D., SUN, Y. and TAYLOR, J. E. (2015). On model selection consistency of regularized M-estimators. *Electron. J. Stat.* **9** 608–642. MR3331852

LEY, R. E. (2010). Obesity and the human microbiome. *Curr. Opin Gastroenterol.* **26** 5–11.

LI, H. (2015). Microbiome, metagenomics and high-dimensional compositional data analysis. *Ann. Rev. Stat. Appl.* **2** 73–94.

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

MCMURDIE, P. J. and HOLMES, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8** e61217.

NAVAS-MOLINA, J. A., PERALTA-SÁNCHEZ, J. M., GONZÁLEZ, A., MCMURDIE, P. J., VÁZQUEZ-BAEZA, Y., XU, Z., URSELL, L. K., LAUBER, C., ZHOU, H., SONG, S. J., HUNTLEY, J., ACKERMANN, G. L., BERG-LYONS, D., HOLMES, S., CAPORASO, J. G. and KNIGHT, R. (2013). Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.* **531** 371–444.

ROTA, G.-C. (1964). The number of partitions of a set. *Amer. Math. Monthly* **71** 498–504. MR0161805

SCEALY, J. L. and WELSH, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 351–375. MR2815780

SCEALY, J. L., DE CARITAT, P., GRUNSKY, E. C., TSAGRIS, M. T. and WELSH, A. H. (2015). Robust principal component analysis for power transformed compositional data. *J. Amer. Statist. Assoc.* **110** 136–148. MR3338492

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHI, P., ZHANG, A. and LI, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **10** 1019–1040. MR3528370

ST. JOHN, R. C. (1984). Experiments with mixtures, ill-conditioning, and ridge regression. *J. Qual. Technol.* **16** 81–96.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

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., BÄCKHED, F., FULTON, L. and GORDON, J. I. (2008). Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host & Microbe* **3** 213–223.

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.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

ZHANG, H., DIBAISE, J. K., ZUCCOLO, A., KUDRNA, D., BRAIDOTTI, M., YU, Y., PARAMESWARAN, P., CROWELL, M. D., WING, R., RITTMANN, B. E. et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. USA* **106** 2365–2370.

ZHANG, C., ZHANG, M., WANG, S., HAN, R., CAO, Y., HUA, W., MAO, Y., ZHANG, X., PANG, X., WEI, C. et al. (2010). Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. *ISME J.* **4** 232–241.

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

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

DEPARTMENT OF BIOINFORMATICS
AND BIOSTATISTICS
SHANGHAI JIAO TONG UNIVERSITY
SHANGHAI 200240
CHINA
AND
SJTU-YALE JOINT CENTER FOR
BIOSTATISTICS
SHANGHAI JIAO TONG UNIVERSITY
SHANGHAI 200240
CHINA
E-MAIL: neowangtao@sjtu.edu.cn

DEPARTMENT OF BIOSTATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06510
USA
AND
SJTU-YALE JOINT CENTER FOR
BIOSTATISTICS
SHANGHAI JIAO TONG UNIVERSITY
SHANGHAI 200240
CHINA
E-MAIL: hongyu.zhao@yale.edu