Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Constructing Predictive Microbial Signatures at Multiple Taxonomic Levels

Tao Wang & Hongyu Zhao

To cite this article: Tao Wang & Hongyu Zhao (2017) Constructing Predictive Microbial Signatures at Multiple Taxonomic Levels, Journal of the American Statistical Association, 112:519, 1022-1031, DOI: 10.1080/01621459.2016.1270213

To link to this article: https://doi.org/10.1080/01621459.2016.1270213

View supplementary material

Accepted author version posted online: 05 Jan 2017.
Published online: 30 Oct 2017.

Submit your article to this journal

Article views: 710

View related articles

View Crossmark data

Citing articles: 4 View citing articles

Full Terms & Conditions of access and use can be found at
https://amstat.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, 1022–1031, Applications and Case Studies
https://doi.org/10.1080/01621459.2016.1270213

Taylor & Francis
Taylor & Francis Group

Check for updates

Constructing Predictive Microbial Signatures at Multiple Taxonomic Levels

Tao Wang[a] and Hongyu Zhao[b]

[a]Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China; [b]Department of Biostatistics, Yale University, New Haven, CT

ABSTRACT

Recent advances in DNA sequencing technology have enabled rapid advances in our understanding of the contribution of the human microbiome to many aspects of normal human physiology and disease. A major goal of human microbiome studies is the identification of important groups of microbes that are predictive of host phenotypes. However, the large number of bacterial taxa and the compositional nature of the data make this goal difficult to achieve using traditional approaches. Furthermore, the microbiome data are structured in the sense that bacterial taxa are not independent of one another and are related evolutionarily by a phylogenetic tree. To deal with these challenges, we introduce the concept of variable fusion for high-dimensional compositional data and propose a novel tree-guided variable fusion method. Our method is based on the linear regression model with tree-guided penalty functions. It incorporates the tree information node-by-node and is capable of building predictive models comprised of bacterial taxa at different taxonomic levels. A gut microbiome data analysis and simulations are presented to illustrate the good performance of the proposed method. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2015
Accepted November 2016

KEYWORDS
Compositional data;
High-dimensional data;
Microbiome signatures;
Tree-structured regression

1. Introduction

The human microbiota is the aggregate of microorganisms that live inside and on the human body. They include bacteria, archea, fungi, and viruses, with bacteria alone estimated to outnumber human cells within an individual by an order of magnitude (Savage 1977). The totality of these microbes and the genes they encode is called the human microbiome.

High-throughput DNA sequencing technologies, coupled with bioinformatics developments, have enabled rapid advances in our understanding of the human microbiome (Kuczynski et al. 2011). For example, several studies have revealed that host-associated microbial communities are likely to play key roles in normal physiology (Dethlefsen and Relman 2011), development (Koenig et al. 2011), and diseases such as inflammatory bowel disease (Sartor 2008). To characterize the microbial community structure and function, marker gene sequencing and whole-genome shotgun or metagenomic sequencing are two widely used approaches (Petrosino et al. 2009). After sequences have been acquired, they are usually clustered into Operational Taxonomic Units (OTUs): groups of sequences that are intended to correspond to taxonomic clades or monophyletic groups. QIIME (Caporaso et al. 2010) and AMPHORA (Wu and Eisen 2008) are two commonly used tools to infer OTUs. The abundance of an OTU is defined as the number of sequences in that OTU. The microbial community is often described by a list of OTUs and their frequencies, and a phylogenetic tree.

One of the goals of human microbiome studies is to understand how microbial communities affect health and correlate with disease (Peterson et al. 2009). In particular, it is important to identify groups of OTUs (species or taxa, which are used interchangeably) that are consistently predictive of host phenotypes (Knights et al. 2011). To develop new statistical methods to determine the association between microbial profiles and phenotypes such as obesity, however, it is crucial to take into consideration the challenges inherent in microbiome sequence data.

First, the OTUs are related to one another by a phylogenetic tree. Appropriate use of this hierarchical structure can lead to more informative analyses. For example, to estimate the taxonomic diversity between communities, distance measures that use the tree information, such as UniFrac (Lozupone and Knight 2005), are more effective at revealing ecological patterns than distance measures that use only sets of isolated OTUs. To develop predictive models, on the other hand, using the structure of the tree allows us to borrow statistical strength across phylogenetically close OTUs by encouraging them to have similar effects on the trait of interest. Chen et al. (2013) proposed to use the phylogeny in canonical correlation analysis by treating the tree as a special case of undirected graphs and then imposing a tree-constrained smoothness penalty. More recently, Tanaseichuk, Borneman, and Jiang (2014) adopted a tree-guided group-lasso penalty that incorporates the hierarchy in the microbiome space, and developed a new method for classifying microbial community samples.

Second, a host phenotype may be characterized by the microbial community through several taxa of varying phylogenetic depth (Knights et al. 2011). In this case, the goal of microbiome analysis is to determine, for the given phenotype, which taxa matter and at what taxonomic level. Intuitively, treating

CONTACT Hongyu Zhao hongyu.zhao@yale.edu Department of Biostatistics, Yale University, PO Box 208034, 60 College Street, New Haven, CT 06520-8034.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.
© 2017 American Statistical Association

intermediate nodes of the phylogenetic tree as candidate features will lead to more accurate and more interpretable results. To our knowledge, however, few methods are available for constructing predictive microbial signatures at multiple taxonomic levels (Zhang et al. 2015).

Third, the number of sequencing reads varies greatly across samples. In other words, the abundance of species is measured in relative terms. Consequently, these count data are compositional (Aitchison 1986) and should be treated appropriately. To select the taxa that are associated with a phenotype, Lin et al. (2014) recently proposed an l_1 regularization methodology under a linear log-contrast model that accounts for the natural constraint of compositional data. An alternative approach is to model the counts explicitly. For example, Xia et al. (2013) proposed to use a logistic normal multinomial model to characterize the dependency of the community composition.

The rest of the article is organized as follows. In Section 2.1, we introduce the linear model, the biological assumption underlying variable fusion, and two methods for variable fusion. We then propose two tree-guided fused lasso penalties that use the tree topology as well as branch lengths in Section 2.2. In Section 2.3, we consider the compositional aspect of the microbiome data and the non-identifiability of regression coefficients caused by it, and adopt fusion methods we proposed to deal with the identifiability issue. Computation and tuning are covered in Section 2.4. Application of the proposed methods to a gut microbiome dataset and simulations are presented in Sections 3 and 4, respectively. We conclude with discussion in Section 5.

2. Methodology

Suppose we have an outcome vector $\boldsymbol{y} = (y_i)$ and a community-by-OTU matrix $\mathbf{X} = (x_{ij})$, where y_i is the phenotype for the ith community sample and x_{ij} denotes the observed abundance of OTU j in sample i, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We also have a binary tree T that describes the phylogenetic relationships among the OTUs. In this tree, leaves correspond to OTUs and internal nodes represent taxa at different levels. We assume that the tree information is known a priori, although in many applications it can be learned from data using a variety of methods (Nielsen 2005). There may be uncertainties in this tree, and we will discuss the robustness of our method in the presence of tree uncertainties in the following.

2.1. Linear Regression Model and Variable Fusion

Since the goal is to understand the role of the structured microbiome space in explaining a phenotype, an effective approach is to pose the problem in a regression framework. In this article, we restrict our discussion to cases where the outcome variable is continuous or quantitative, although the general approach is applicable to cases where the trait of interest is discrete or qualitative. Specifically, we consider the linear model

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i,$$

where β_0 is an intercept, $\beta_1, \ldots, \beta_p$ are regression coefficients, and $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed (iid) errors with zero mean and finite variance.

High-throughput characterizations of microbial communities produce high-dimensional data. The scale of data may remain very large even after sequences have been assigned to OTUs. As such, dimension reduction is necessary. Our method is based on the biological assumption that phylogenetically close taxa may have similar associations with a host phenotype. Evidence supporting this can be found in several studies, see, for example, Ahn et al. (2013), Dey et al. (2013), and Garcia et al. (2014). Under this assumption, one natural way to handle the dimensionality problem is to conduct variable fusion by shrinking some $|\beta_{j_1} - \beta_{j_2}|$ to zero. For example, we can solve the pairwise fused lasso (She 2010) problem

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{1\le j_1<j_2\le p} |\beta_{j_1} - \beta_{j_2}|,$$

where $\lambda > 0$ is the tuning parameter. This method fuses variables to each other and induces a clustering of variables.

Variable fusion is crucial in high-dimensional problems for interpretability and for improved predictive performance. However, the pairwise fused lasso treats all pairs of OTUs equally and thus fails to exploit the phylogeny of the OTUs. The recently proposed structure-constrained canonical correlation analysis (Chen et al. 2013) used a ridge-type penalty $\lambda \sum_{j_1<j_2} (\beta_{j_1} - \beta_{j_2})^2/d^2_{j_1 j_2}$ to smooth the coefficients of two OTUs j_1 and j_2 based on their closeness $d_{j_1 j_2}$ on the phylogenetic tree. Similarly, we can use a tree-weighted penalty

$$\lambda \sum_{1\le j_1<j_2\le p} w_{j_1 j_2}|\beta_{j_1} - \beta_{j_2}|, \tag{1}$$

where the $w_{j_1 j_2}$ are tree-based weights. The assumption that closely related taxa are likely to share a similar effect implies that $w_{j_1 j_2}$ should be small when OTUs j_1 and j_2 are distantly related. Throughout this article, we set $w_{j_1 j_2} = d^{\alpha}_{j_1 j_2}$, with $\alpha \le 0$ being another tuning parameter to be determined. Then, we need to solve the following problem:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{1\le j_1<j_2\le p} w_{j_1 j_2}|\beta_{j_1} - \beta_{j_2}|. \tag{2}$$

The pairwise fused lasso is a generalization of the ordinary fused lasso (Tibshirani et al. 2005) defined by

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p-1} w_{j(j+1)}|\beta_j - \beta_{j+1}|. \tag{3}$$

The latter is intended for situations in which variables have an ordering along which smoothness is expected. Since any ordering of OTUs is arbitrary, the ordinary fused lasso may be misleading.

2.2. Tree-Guided Fused Lasso

In a classification context, Tanaseichuk, Borneman, and Jiang (2014) recently proposed a method that uses the tree topology. To be specific, let V denote the collection of all nodes or vertices of the tree T. For each $v \in V$, let $L_v \subseteq \{1, \ldots, p\}$ be the set of indices of OTUs that correspond to the leaves of the subtree rooted at v, and let $\beta_{L_v} = (\beta_j, j \in L_v)$. They adopted a group-lasso-type penalty $\lambda \sum_{v \in V} \|\beta_{L_v}\|_2$ based on the grouping of the OTUs along the tree. Here, $\|\cdot\|_2$ denotes the Euclidean norm.

However, their penalty is not able to harness a predictive microbial signature made of a set of multi-level taxa, and as we will see in the next section, it does not take into account the compositional aspect of the microbiome data. To address these problems, we propose a novel tree-guided variable fusion method. The basic idea is to treat the internal nodes as potential features. In the development below, we construct two tree-weighted fused lasso penalties that encode the tree topology.

Denote by N the set of internal nodes of T. For each $v \in N$, let c_{v_1} and c_{v_2} be the two child nodes of v. For any set S, we denote by $|S|$ the size of S. Let $\beta = (\beta_1, \ldots, \beta_p)^\top$. The first tree-guided fused lasso penalty is defined by

$$\lambda \sum_{v \in N} w_{c_{v_1} c_{v_2}} \left| {}_1D_v^\top \beta \right|, \tag{4}$$

where ${}_1D_v \in \mathbb{R}^p$ is an indicator vector with jth entry $1/|L_{c_{v_1}}|$ if $j \in L_{c_{v_1}}$, $-1/|L_{c_{v_2}}|$ if $j \in L_{c_{v_2}}$, and 0 otherwise. In other words, the two child nodes c_{v_1} and c_{v_2} each take a proportion of the weight $w_{c_{v_1} c_{v_2}}$ of the parent node $v \in N$, relative to the sizes of their subtrees.

Denote by L the set of leaves of T. Without loss of generality, let $L = \{1, \ldots, p\}$ and $N = \{p+1, \ldots, |V|\}$. Note that, for a binary tree with p leaves, $|N| = p - 1$ and $|V| = 2p - 1$. To design the second penalty, we use a bottom-up approach by first computing the penalty terms for all internal nodes with size 2 subtrees, then all with size 3 subtrees if any exist, and so on. Specifically, define A to be the level set such that, for each $s \in A$, there is $v \in N$ such that $s = |L_v|$. Let s_h denote the hth smallest element of A. Let $E = (e_{ij})$ be a $|V| \times p$ matrix with $e_{jj} = 1$ for $j \in L$ and 0 otherwise. We denote by E_j the jth row of E. The second tree-guided fused lasso penalty is defined by

$$\lambda \sum_{v \in N} w_{c_{v_1} c_{v_2}} \left| {}_2D_v^\top \beta \right|, \tag{5}$$

where for $h = 1, \ldots, |A|$, we set ${}_2D_v = E_{c_{v_1}} - E_{c_{v_2}}$ and update $E_v = (E_{c_{v_1}} + E_{c_{v_2}})/2$, for all $v \in N$ such that $|L_v| = s_h$.

We now consider the tree-guided fused lasso problem

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{v \in N} w_{c_{v_1} c_{v_2}} \left| D_v^\top \beta \right|, \tag{6}$$

where $D_v = {}_1D_v$ or $D_v = {}_2D_v$. Since each internal node represents the abundance of a taxonomic lineage, by incorporating the tree information *node-by-node*, the estimated microbial signature from our tree-guided fused lasso tends to be composed of a few taxonomic units at different depths. Formally, the variables indexed by L_u for $u \in N$ are fused together, defining a new variable indexed by u, if $D_v^\top \beta = 0$ for all the internal nodes v of the subtree rooted at u. Note that although in this article we focus on binary trees, the idea here can be naturally extended to K-ary trees.

It turns out that the ordinary fused lasso, pairwise fused lasso, and tree-guided fused lasso are nicely encapsulated by the generalized lasso (Tibshirani and Taylor 2011) formulation

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \|D(\alpha)\beta\|_1, \tag{7}$$

where $D(\alpha) \in \mathbb{R}^{m \times p}$ is a tree-weighted penalty matrix and $\|\cdot\|_1$ denotes the l_1 norm. For the pairwise fused lasso, $m = p \times (p-1)/2$, while for the ordinary fused lasso and tree-guided fused lasso, $m = p - 1$. Note also that each component of $D(\alpha)\beta$ denotes a linear contrast of $\beta_1, \ldots, \beta_p$, due to the nature of these fusion penalties.

For illustration, Figure 1 shows two binary trees, each with $p = 4$ leaves. For the pairwise fused lasso, the weighted penalty matrix is

$$D(\alpha) = \begin{pmatrix} d_{12}^\alpha & -d_{12}^\alpha & 0 & 0 \\ d_{13}^\alpha & 0 & -d_{13}^\alpha & 0 \\ d_{14}^\alpha & 0 & 0 & -d_{14}^\alpha \\ 0 & d_{23}^\alpha & -d_{23}^\alpha & 0 \\ 0 & d_{24}^\alpha & 0 & -d_{24}^\alpha \\ 0 & 0 & d_{34}^\alpha & -d_{34}^\alpha \end{pmatrix} \in \mathbb{R}^{6 \times 4}.$$

Here, the d_{j_1, j_2} are calculated from branch lengths $t_1, \ldots, t_6$ and are tree dependent. Consider the tree in the left panel. The penalty matrices for the two tree-guided fused lasso penalties are

$$D(\alpha) = \begin{pmatrix} d_{12}^\alpha & -d_{12}^\alpha & 0 & 0 \\ d_{37}^\alpha/2 & d_{37}^\alpha/2 & -d_{37}^\alpha & 0 \\ d_{46}^\alpha/3 & d_{46}^\alpha/3 & d_{46}^\alpha/3 & -d_{46}^\alpha \end{pmatrix} \in \mathbb{R}^{3 \times 4}$$

and

$$D(\alpha) = \begin{pmatrix} d_{12}^\alpha & -d_{12}^\alpha & 0 & 0 \\ d_{37}^\alpha/2 & d_{37}^\alpha/2 & -d_{37}^\alpha & 0 \\ d_{46}^\alpha/4 & d_{46}^\alpha/4 & d_{46}^\alpha/2 & -d_{46}^\alpha \end{pmatrix} \in \mathbb{R}^{3 \times 4},$$

respectively. For the tree in the right panel, the penalty matrices for our tree-guided fused lasso penalties are the same (in general they are not),

$$D(\alpha) = \begin{pmatrix} d_{12}^\alpha & -d_{12}^\alpha & 0 & 0 \\ 0 & 0 & d_{34}^\alpha & -d_{34}^\alpha \\ d_{67}^\alpha/2 & d_{67}^\alpha/2 & -d_{67}^\alpha/2 & -d_{67}^\alpha/2 \end{pmatrix} \in \mathbb{R}^{3 \times 4}.$$

2.3. Compositional Data and Identifiability

For data collected in practical studies, different samples usually have very different numbers of sequences. To deal with this issue of sequencing depth, a typical step is to convert raw counts into proportions. The resulting percentage data, denoted by $\mathbf{Z} = (z_{ij})$ with $z_{ij} = x_{ij} / \sum_{k=1}^{p} x_{ik}$, are compositional in that the relative abundances must sum up to one, $\sum_{j=1}^{p} z_{ij} = 1$. This

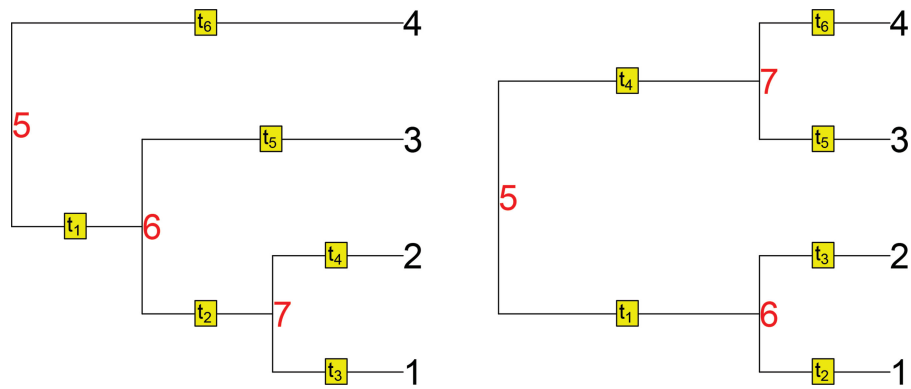

Figure 1. Two binary trees, each with four leaves and three internal nodes. Branch lengths are labeled as $t_1, \ldots, t_6$.

property is consistent with the fact that when the relative abundance of one OTU increases, the relative abundance of the rest of the community must necessarily decrease. However, it renders standard multivariate statistical methods inappropriate.

To understand this, we assume that $\{(\epsilon_i, z_{i1}, \ldots, z_{ip}), i = 1, \ldots, n\}$ is an iid random sample on $(\epsilon, Z_1, \ldots, Z_p)$ with $\sum_{j=1}^{p} Z_j = 1$, and $y_1, \ldots, y_n$ are generated from the model

$$Y = \beta_0 + \sum_{j=1}^{p} Z_j \beta_j + \epsilon, \tag{8}$$

where $E(\epsilon) = 0$ and $E(\epsilon^2) < \infty$. Here, we assume that the microbiome's effect on the outcome variable is mediated through its taxonomic composition but not the sampling depth. This working assumption is reasonable since for the microbiome data, the number of sequencing reads can vary substantially across samples, and more importantly, the actual amount of the mixture of components is usually unknown.

From (8), we have

$$\beta_0 = E(Y) - \sum_{j=1}^{p} E(Z_j)\beta_j$$

and

$$Y - E(Y) = \sum_{j=1}^{p} \{Z_j - E(Z_j)\}\beta_j + \epsilon.$$

Furthermore, it is easy to check that

$$Y - E(Y) = \sum_{j=1}^{p} \{Z_j - E(Z_j)\}(\beta_j + c) + \epsilon$$

for any constant c. That is, the coefficients β_j are identifiable only up to a common additive constant. The situation here formally resembles standard analysis of variance, in which the dummy variables that code a multi-level categorical predictor sum up to one. Although the idea of variable fusion, or the biological assumption underlying it, is independent of the unit-sum constraint, variable fusion can handle this identifiability issue, as we show below.

Specifically, to make the p parameters identifiable, we can impose a constraint on them. Note that, for each $k \in \{1, \ldots, p\}$,

$$Y - E(Y) = \sum_{j=1}^{p} \{Z_j - E(Z_j)\}\beta_j(k) + \epsilon,$$

where $\beta_j(k) = \beta_j - \beta_k$ reflects the difference between β_j and β_k. The constraint is $\beta_k(k) = 0$. Without loss of generality, assume that both Y and $(Z_1, \ldots, Z_p)^\top$ are centered, so the intercept is not included in the regression. To overcome the dimensionality problem, an alternative to dimension reduction is variable selection. For example, to get a sparse solution, we can consider the lasso (Tibshirani 1996) problem

$$\underset{\beta_j(k),\, j \neq k}{\text{minimize}} \sum_{i=1}^{n} \left\{ y_i - \sum_{j \neq k} z_{ij}\beta_j(k) \right\}^2 + \sum_{j \neq k} |\beta_j(k)|. \tag{9}$$

If $\beta_j(k)$ is estimated to be zero, then OTUs j and k are fused together. However, the lasso has two drawbacks. First, the solution to (9) depends on the choice of the reference index k. Second, the corresponding fusion pattern, which is the same as the sparsity pattern, is very rough (only those components with zero coefficients are fused together).

To overcome these problems, we note that

$$\beta_{j_1}(k) - \beta_{j_2}(k) = \beta_{j_1} - \beta_{j_2}$$

for all $1 \leq j_1, j_2 \leq p$. This in turn means that it is the relative, rather than absolute, value of β_j that matters, and that the choice of k has no impact on the differences $\beta_{j_1} - \beta_{j_2}$. Thus, to handle the dimensionality problem for compositional data, instead of conducting variable selection by penalizing $|\beta_j(k)|$ directly, we should conduct variable fusion by shrinking some $|\beta_{j_1}(k) - \beta_{j_2}(k)|$ to zero, as we did in Sections 2.1 and 2.2. Taking the compositional aspect of the microbiome data into account, we now arrive at the following generalized lasso problem:

$$\underset{\beta_j(k),\, j \neq k}{\text{minimize}} \sum_{i=1}^{n} \left\{ y_i - \sum_{j \neq k} z_{ij}\beta_j(k) \right\}^2 + \lambda \|D(\alpha)\beta(k)\|_1. \tag{10}$$

Special instances of $D(\alpha)$ give rise to the ordinary fused lasso, pairwise fused lasso, and tree-guided fused lasso. Since each row of $D(\alpha)$ is a contrast vector, the solution to (10) is independent of k.

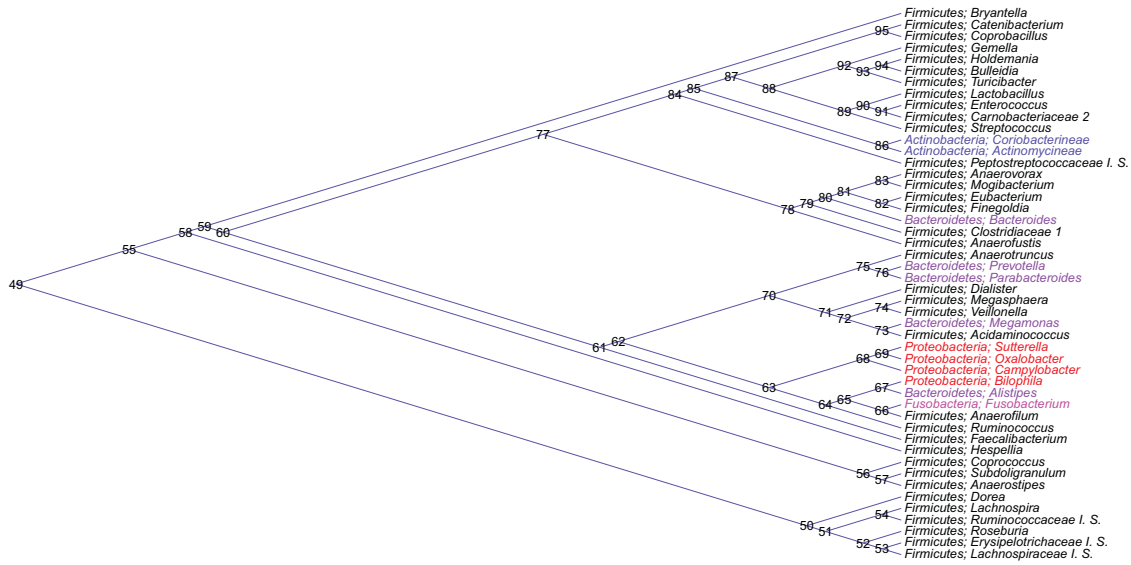

Figure 2. The phylogenetic tree of 48 selected genera for the gut microbiome data. Internal nodes are labeled as 49, . . . , 95.

2.4. Computation and Tuning

We use the dual path algorithm of Tibshirani and Taylor (2011) to optimize (10). The algorithm is efficiently implemented in the `genlasso` R package. Given $D(\alpha)$, the main function in `genlasso` computes the solution path for all values of the tuning parameter λ. In addition, it provides an unbiased estimate of the degrees of freedom of the fit (Tibshirani and Taylor 2011, 2012). This is very important, since we can then use different model selection criteria, which employ degrees of freedom to assess risk, to select the tuning parameter.

Let us denote the solution of problem (10) by $\{\hat{\beta}_j(\lambda, \alpha; k), j \neq k\}$. Define

$$\mathrm{IC}(\lambda, \alpha; k) = \log\{\mathrm{RSS}(\lambda, \alpha; k)\} + \kappa \times df(\lambda, \alpha; k), \qquad (11)$$

where $\mathrm{RSS}(\lambda, \alpha; k) = \sum_{i=1}^{n}\{y_i - \sum_{j\neq k} z_{ij}\hat{\beta}_j(\lambda, \alpha; k)\}^2$, κ is a penalty factor, and $df(\lambda, \alpha; k)$ is an unbiased estimate of the degrees of freedom of $\{\hat{\beta}_j(\lambda, \alpha; k), j \neq k\}$. Two popular criteria are AIC, which uses $\kappa = 2$, and BIC, which uses $\kappa = \log(n)$. For each α, we select λ by minimizing $\mathrm{IC}(\lambda, \alpha; k)$ along the path. Throughout the numerical studies of this article, we set $k = 1$ and explore three values of α: -2, -1, and 0. The chosen α is the one giving the smallest IC value.

3. Gut Microbiome Data Analysis

The human gut harbors diverse microbes that play fundamental roles in the well-being of their host (Clemente et al. 2012). For example, studies have suggested that the gut microbiome is implicated in the etiopathogenesis of obesity (Turnbaugh et al. 2008; Ley 2010), although the casual nature of this association is somewhat controversial (Finucane et al. 2014; Rosenbaum, Knight, and Leibel 2015).

To demonstrate the use of our method, we applied it to a dataset from a human gut microbiome study carried out at the University of Pennsylvania (Wu et al. 2011). Specifically, for this study, fecal samples were obtained from 98 subjects and bacterial DNA was extracted using a standard protocol (Wu et al. 2010). After multiplexed 454/Roche pyrosequencing, about 900,000 high quality, partial (∼370 bp) 16S rRNA gene sequences were generated. These sequences were analyzed by the QIIME pipeline with default parameter settings (Navas-Molina et al. 2013): more than 17,000 OTUs were identified and their counts quantified at 97% of sequence similarity, the taxonomy was assigned using the RDP classifier (Wang et al. 2007), and the phylogenetic tree was inferred by FastTree 2 (Price, Dehal, and Arkin 2010). To reduce the number of OTUs, we merged them into genera (genus-level OTUs). We further dropped genera that were observed in less than 5 of the samples. Since the sampling depths were very different for different samples, we normalized counts into proportions. In our analysis, the microbiome data were summarized as a matrix $\mathbf{Z} = (z_{ij}) \in \mathbb{R}^{98\times 48}$ of relative abundances of 48 genus-level OTUs, and a phylogenetic tree T as displayed in Figure 2. The tree was rooted using an outgroup. More information about the tree can be found in Supplementary Materials, Section A. On the other hand, clinical measurements, including Body Mass Index (BMI), were also available for these 98 subjects. We are interested in constructing a predictive microbial signature for BMI that accounts for the phylogeny relating OTUs.

Note that our method requires a phylogenetic tree, but not a taxonomic table. While a phylogenetic tree can always be learned from molecular sequences, 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 addition, the taxonomy may not be consistent with the tree structure (see Figure 2).

We used Monte-Carlo cross-validation (also called sample splitting) to evaluate the prediction performance of various methods, including the lasso defined in (9). More specifically, we sampled 80% of the observations without replacement, fitted the model using that subsample, and used the remaining

Table 1. The averages of the model size (MS) and prediction error (PERR) with standard deviations in parentheses, based on 200 random splits of 98 samples, are reported for the lasso (LASSO), ordinary fused lasso (OFL), pairwise fused lasso (PFL), and two versions of tree-guided fused lasso (TFL-1 and TFL-2).

		MS	PERR
LASSO	AIC	7.385 (4.272)	35.282 (29.892)
	BIC	2.985 (1.816)	31.039 (14.208)
OFL	AIC	8.895 (4.417)	36.574 (22.765)
	BIC	2.245 (1.508)	30.518 (14.298)
PFL	AIC	3.220 (1.635)	32.280 (15.335)
	BIC	2.185 (1.349)	30.730 (13.844)
TFL-1	AIC	12.950 (6.377)	36.860 (23.746)
	BIC	2.840 (2.396)	30.382 (14.280)
TFL-2	AIC	13.830 (6.046)	36.053 (21.420)
	BIC	3.380 (2.916)	30.372 (14.120)

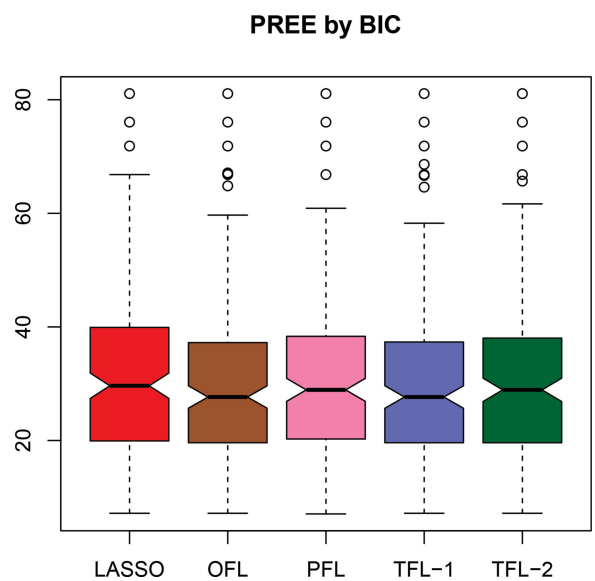

Figure 3. Boxplots of the prediction error (PREE) using BIC, over 200 random splits of 98 samples, for the lasso (LASSO), ordinary fused lasso (OFL), pairwise fused lasso (PFL), and two versions of tree-guided fused lasso (TFL-1 and TFL-2).

observations to assess the predictive accuracy of each method. The results of applying this procedure 200 times are summarized in Table 1 and Figure 3. As we can see, the AIC criterion was inferior to BIC in terms of prediction accuracy. Furthermore, when BIC was used, tree-guided variable fusion helped improve prediction slightly.

For the tree-guided fused lasso with BIC, we checked the fusion pattern along the tree. Table 2 reports the subtrees (indexed by the internal nodes of T) that appeared in the model at least 160 out of 200 times. Here, a subtree is said to be in the model if all its leaves are fused together. We see that there was considerable overlap in the sets of subtrees from two versions of the tree-guided fused lasso.

To assist in the interpretation of results, Table 3 shows the taxonomic memberships for some of the genus-level OTUs that were fused along the tree. In this table, the first column "Internal node" lists the internal nodes whose leaves were all fused together by our method. For example, for the internal node 50, its six leaves (i.e., Lachnospiraceae I. S., Erysipelotrichaceae I. S., Roseburia, Ruminococcaceae I. S., Lachnospira, and Dorea) were fused. We can see that these OTUs were grouped into key phyla in the human gut microbiota: Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria. Furthermore, some OTUs were also mapped together at lower taxonomic levels. It has been experimentally shown that in humans on weight-reduction diets, the decrease in Bacteroidetes was accompanied by an increase in Firmicutes (Ley 2010). In addition, a study with human twins manifested the interplay between Bacteroidetes and Actinobacteria in obese individuals, with a reduction of the former and a corresponding increase in the latter (Turnbaugh et al. 2008). Generally, the observed shift in the relative abundances of these phyla is closely related to energy harvest (Kinross, Darzi, and Nicholson 2011).

Table 2. Subtrees appearing in the model at least 160 times, based on 200 random splits of 98 samples, are reported for two versions of the tree-guided fused lasso (TFL-1 and TFL-2).

	Labels of subtrees
TFL-1	50–57, 69–76, 83, 86, 89–91, 95
TFL-2	50–57, 69–76, 83, 86, 91, 95

4. Simulations

In this section, we present a simulation study to evaluate the performance of the proposed methods. To mimic the kind of data that we might see in metagenomic applications, we use the compositional data matrix $\mathbf{Z}$ and the phylogenetic tree T from the gut microbiome data described in the previous section, where $n = 98$ and $p = 48$. The continuous outcome is generated from the model

$$y_i = \sum_{j=1}^{p} z_{ij}\beta_j + \epsilon_i$$

for $i = 1, \ldots, n$, where $\beta_1, \ldots, \beta_p$ are specified in Table 4, and $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$. Three values of σ are considered: 1.1, 1.8, and 2.7. The corresponding signal-to-noise ratios are 4.10, 1.53, and 0.68, respectively. We note that the β_j here were learned from the real data by the second version of the tree-guided fused lasso. Figure 4 shows a graphical summary of the fusion pattern among the OTUs. As we can see, three OTUs are fused but not along the tree (two OTUs are fused together if they have the same coefficients), hence the fusion pattern is not fully consistent with the tree topology.

For each of 200 simulated datasets, we applied the lasso, ordinary fused lasso, pairwise fused lasso, and two versions of tree-guided fused lasso. To evaluate the performance, we computed (a) the true positive rate with respect to leaves,

$$\mathrm{TPR}_L = \frac{\text{the number of pairs of leaves correctly fused}}{\text{the number of pairs of leaves that are truly fused}},$$

(b) the false positive rate with respect to leaves,

$$\mathrm{FPR}_L = \frac{\text{the number of pairs of leaves falsely fused}}{\binom{p}{2} - \text{the number of pairs of leaves that are truly fused}},$$

(c) the true positive rate with respect to internal nodes,

$$\mathrm{TPR}_N$$

Table 3. Gut microbiome data. Taxonomic memberships for some of the genus-level OTUs fused along the tree by the tree-guided fused lasso.

Internal node	Phylum	Class	Order	Family	Genus (Leave)
50	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae I. S.
50	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae I. S.
50	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
50	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae I. S.
50	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospira
50	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Dorea
56	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerostipes
56	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
56	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus
69	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Oxalobacter
69	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Sutterella
74	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Veillonella
74	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Megasphaera
76	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae	Parabacteroides
76	Bacteroidetes	Bacteroidetes	Bacteroidales	Prevotellaceae	Prevotella
83	Firmicutes	Clostridia	Clostridiales	Incertae Sedis XIII	Mogibacterium
83	Firmicutes	Clostridia	Clostridiales	Incertae Sedis XIII	Anaerovorax
86	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Actinomycineae
86	Actinobacteria	Actinobacteria	Coriobacteridae	Coriobacteriales	Coriobacterineae
89	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
89	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Carnobacteriaceae 2
89	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
89	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
95	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	Coprobacillus
95	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	Catenibacterium

$$= \frac{\text{the number of internal nodes whose child nodes are correctly fused}}{\text{the number of internal nodes whose child nodes are truly fused}},$$

(d) the false positive rate with respect to internal nodes,

$$\mathrm{FPR}_N = \frac{\text{the number of internal nodes whose child nodes are falsely fused}}{p - 1 - \text{the number of internal nodes whose child nodes are truly fused}},$$

(e) the model size after fusion (the true value is 15), and (f) the mean squared error $\|\mathbf{Z}\beta(k) - \mathbf{Z}\hat{\beta}(k)\|_2^2/(n\sigma^2)$.

The simulation results are summarized in Table 5 for $\sigma = 1.1$, and Tables B1 and B2 in Supplementary Materials, Section B for $\sigma = 1.8$ and $\sigma = 2.7$. As we can see, the pairwise fused lasso was very aggressive, with a high false positive rate with respect to internal nodes and a small model size after fusion. We can also see that for the lasso and ordinary fused lasso, the two types of true positive rates were relatively low. On the other hand, the two versions of the tree-guided fused lasso behaved similarly and performed reasonably well: they had lower false positive rates than the pairwise fused lasso, and higher true positive rates than the lasso and ordinary fused lasso. Generally, BIC outperformed AIC, and the performance of all methods became worse as the signal-to-noise ratio decreased.

Table 4. The regression coefficients used for the simulation study.

β_1	−23.526	β_{17}	−0.168	β_{33}	−11.057
β_2	−23.526	β_{18}	−0.168	β_{34}	−11.057
β_3	−30.688	β_{19}	−0.168	β_{35}	10.720
β_4	−27.107	β_{20}	40.712	β_{36}	10.720
β_5	−27.107	β_{21}	40.712	β_{37}	10.720
β_6	13.475	β_{22}	60.721	β_{38}	10.720
β_7	−13.462	β_{23}	60.721	β_{39}	10.720
β_8	−13.462	β_{24}	−24.817	β_{40}	10.720
β_9	−13.462	β_{25}	−13.286	β_{41}	10.720
β_{10}	−0.168	β_{26}	−13.286	β_{42}	10.720
β_{11}	−0.168	β_{27}	−13.286	β_{43}	10.720
β_{12}	12.975	β_{28}	−11.057	β_{44}	10.720
β_{13}	−8.277	β_{29}	−11.057	β_{45}	10.720
β_{14}	−8.277	β_{30}	−11.057	β_{46}	10.720
β_{15}	−18.346	β_{31}	−11.057	β_{47}	10.720
β_{16}	−18.346	β_{32}	−11.057	β_{48}	−0.168

In Supplementary Materials, Section B, we further examined the impact of the reference index k on the performance. The results shown in Table B3 confirmed that the solution of the lasso varied with k. Furthermore, the ordinary fused lasso and tree-guided fused lasso were not affected much by the choice of k, while the pairwise fused lasso was numerically very unstable. This is because the penalty matrix for the pairwise fused lasso has a large number of rows, making the resulting optimization much more challenging. In Supplementary Materials, Section C, we ran simulations when the z_{ij} values were generated from a logistic normal distribution rather than fixed, and when $p > n$. The results reported in Tables C1 and C2 are qualitatively similar, except that the performance of all methods deteriorated as the sample size decreased.

Table 6 shows the average CPU seconds for computing the entire path of solutions. All timings were carried out on the Yale Louise high performance cluster (Dell m620 system, 8-core processor, 48G of memory) using R version 3.1.0. As expected, the pairwise fused lasso was much slower than its competitors. We note that the absolute numbers here might be different on different machines, and that it is the relative values that matter.

So far we have assumed that closely-related OTUs in a tree have a similar function because of shared evolutionary ancestry. In reality, this assumption could be violated (see the Discussion section). To further examine the performance of the tree-guided fused lasso, in Supplementary Materials, Section D, we conducted a simulation study in which the tree structure encoded by our penalties was partly or fully inconsistent with the fusion pattern among the true coefficients. The results reported in

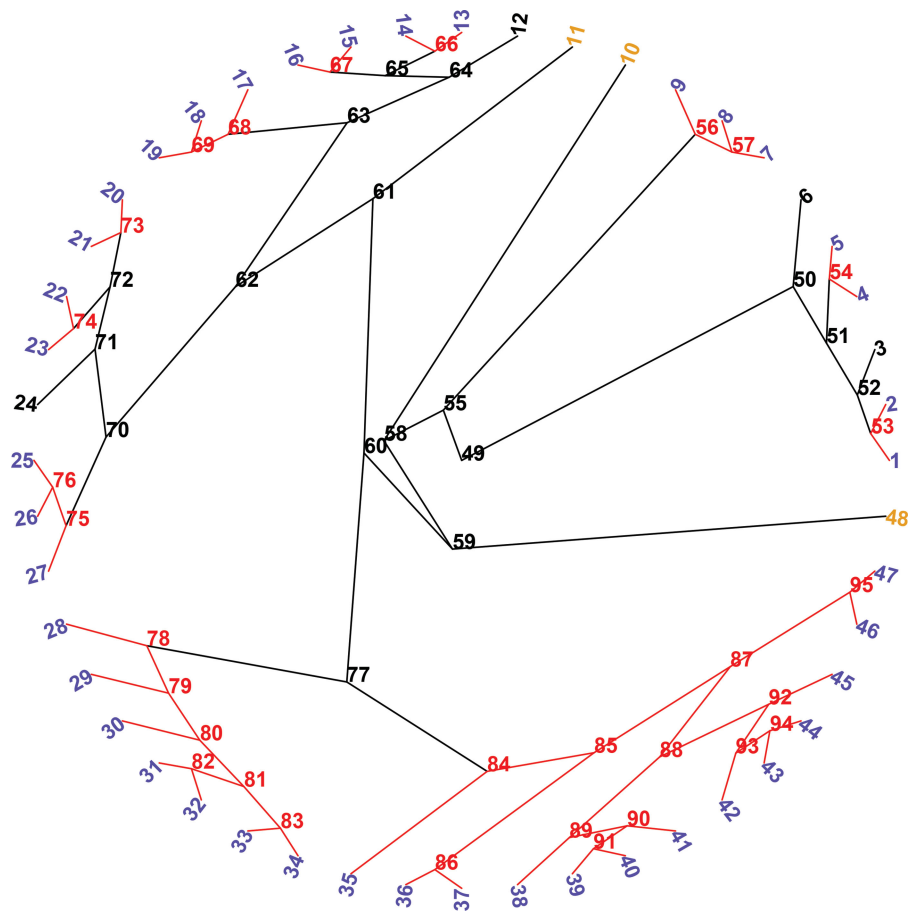

Figure 4. The fusion pattern for the simulation study. Black leaves stand for OTUs that are not fused, blue leaves represent OTUs that are fused along the true, orange leaves show OTUs that are fused but not along the tree, and red internal nodes indicate subtrees whose leaves are fused.

Tables D1 and D2 indicate that the tree-guided fused lasso achieved a level of robustness when the tree structure was partly inconsistent with the fusion pattern, and had inferior performance when the tree was fully inconsistent.

5. Discussion

In this article, we have introduced the concept of variable fusion for dimension reduction in linear regression with compositional covariates (i.e., bacterial species or OTUs) related in a phylogenetic tree. To encourage hierarchically close species to have similar effects on the phenotype, we have proposed the tree-weighted pairwise fused lasso using branch lengths as weights. To construct a predictive microbial signature composed of taxa at different levels, we have designed two tree-guided fused lasso penalties that use the tree topology as well as branch lengths. Both real data analysis and simulations have shown that the tree-guided fused lasso performs better than other methods, such as the lasso and ordinary fused lasso.

There are three possible extensions. First, it is easy to extend tree-guided fused lasso penalties from binary trees to general K-ary trees. Second, we can extend the tree-guided fused lasso from linear models to generalized linear models. Finally, tree-guided variable fusion can be extended to other multivariate analysis methods, such as principal component analysis. Below we discuss some issues related to our methodology.

The idea of fusion and the method of fusion along the tree are very natural under the biological assumption that

Table 5. The averages of the true positive rate (TPR_L) and false positive rate (FPR_L) with respect to leaves, true positive rate (TPR_N) and false positive rate (FPR_N) with respect to internal nodes, model size after fusion (MS, with standard deviation in parentheses), and mean squared error (MSE, with standard deviation in parentheses), based on 200 data replications, are reported for the lasso (LASSO), ordinary fused lasso (OFL), pairwise fused lasso (PFL), and two versions of tree-guided fused lasso (TFL-1 and TFL-2) when $\sigma = 1.1$.

		TPR_L	FPR_L	TPR_N	FPR_N	MS	MSE
LASSO	AIC	0.198	0.010	0.303	0.221	27.455 (8.792)	0.315 (0.125)
	BIC	0.435	0.047	0.606	0.446	16.255 (2.967)	0.306 (0.118)
OFL	AIC	0.467	0.018	0.389	0.021	23.385 (9.472)	0.268 (0.137)
	BIC	0.683	0.062	0.619	0.047	13.875 (2.283)	0.233 (0.113)
PFL	AIC	0.828	0.317	0.896	0.859	4.250 (1.451)	2.011 (1.011)
	BIC	0.847	0.357	0.906	0.873	3.870 (1.524)	2.058 (1.066)
TFL-1	AIC	0.559	0.016	0.438	0.002	29.450 (10.203)	0.267 (0.137)
	BIC	0.872	0.064	0.775	0.008	18.575 (3.038)	0.228 (0.097)
TFL-2	AIC	0.636	0.030	0.526	0.004	26.775 (10.454)	0.239 (0.138)
	BIC	0.903	0.071	0.838	0.007	17.660 (3.006)	0.176 (0.072)

Table 6. The average CPU timings (seconds) for computing the solution path, based on 200 data replications, are reported for the lasso (LASSO), ordinary fused lasso (OFL), pairwise fused lasso (PFL), and two versions of tree-guided fused lasso (TFL-1 and TFL-2) when $\sigma = 1.1$.

LASSO	0.43
OFL	0.72
PFL	140.76
TFL-1	0.77
TFL-2	0.66

closely-related species in a tree may function in a similar manner. However, there are situations in which this assumption could be violated. Clostridia are a good example, where some species convert dietary fiber into anti-inflammatory short chain fatty acids, while others cause severe colitis (Bartlett et al. 1978; Lopetuso et al. 2013). How to fuse bacterial taxa based on their biological function is an interesting direction for future study.

Our method assumes that the tree information is known a priori. Typically, the tree is inferred from molecular sequences, as was the case in the real-data analysis. Since our tree-guided fused lasso penalties use the discrete topology, not just tree-associated distances, accurate methods, especially likelihood-based methods (Price, Dehal, and Arkin 2010; Stamatakis 2014), are recommended for learning trees from data. Another important issue is tree rooting. We note that rooting is not part of tree inference and rooting error is in addition to tree-estimation error. We conducted a small simulation (Supplementary Materials, Section E) and found that there was a loss in accuracy if the root was wrongly inferred, suggesting that rooting error could have a large impact on tree-based methods.

To use the tree topology, we consider the number of subtree nodes as a measure of phylogenetic diversity. A potentially better measure is the branch length. For example, suppose we have a "bushy" subtree of many closely-related nodes (i.e., short branches) and a subtree with long branches. Using the subtree size would treat them equally, but fusing on a bushy tree might be more justified. Clearly, our way of simply using branch lengths as weights is inefficient. As suggested by one referee, our tree-guided fused lasso penalties are related to the so-called "phylogenetic independent contrasts" (Felsenstein 1985), which have the potential advantage of using branch lengths in a more complex but more biologically relevant way. More generally, it is an interesting topic to investigate possible improvements based on the models used in phylogenetic comparative methods (Garamszegi 2014).

To handle the dimensionality problem, variable selection is an alternative to variable fusion. The reason why we do not consider variable selection in the current framework is that, unlike variable fusion, it does not provide a natural way to deal with the compositional aspect of the microbiome data (i.e., the unit-sum constraint). Furthermore, the sparsity pattern is a very rough fusion pattern and is less biologically meaningful. Nevertheless, identifying differential features is one of the central goals for microbiome studies (Albanese et al. 2015). We are working along this line, under a different framework explored by Lin et al. (2014).

Supplementary Materials

The online supplementary materials include the following:
Section A: Additional tree information
Section B: Additional simulation results
Section C: Additional simulations
Section D: Impact of the tree structure
Section E: Issue of tree rooting

Acknowledgments

The authors gratefully acknowledge Hongzhe Li and Jun Chen for providing the data.

Funding

This work was supported by Natural Science Foundation of China (11601326), and NIH grant R01 GM59507.

References

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J. J., Hayes, R. B., and Yang, L. (2013), "Human Gut Microbiome and Risk of Colorectal Cancer," *Journal of the National Cancer Institute*, 105, 1907–1911. [1023]

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, New York: Springer. [1023]

Albanese, D., De Filippo, C., Cavalieri, D., and Donati, C. (2015), "Explaining Diversity in Metagenomic Datasets by Phylogenetic-Based Feature Weighting," *PLoS Computational Biology*, 11, e1004186. [1030]

Bartlett, J. G., Chang, T. W., Gurwith, M., Gorbach, S. L., and Onderdonk, A. B. (1978), "Antibiotic-Associated Pseudomembranous Colitis due to Toxin-Producing Clostridia," *New England Journal of Medicine*, 298, 531–534. [1030]

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," *Nature Methods*, 7, 335–336. [1022]

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," *Biostatistics*, 14, 244–258. [1022,1023]

Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012), "The Impact of the Gut Microbiota on Human Health: An Integrative View," *Cell*, 148, 1258–1270. [1026]

Dethlefsen, L., and Relman, D. A. (2011), "Incomplete Recovery and Individualized Responses of the Human Distal Gut Microbiota to Repeated Antibiotic Perturbation," *Proceedings of the National Academy of Sciences*, 108, 4554–4561. [1022]

Dey, N., Soergel, D. A., Repo, S., and Brenner, S. E. (2013), "Association of Gut Microbiota with Post-Operative Clinical Course in Crohn's Disease," *BMC Gastroenterology*, 13, 131. [1023]

Felsenstein, J. (1985), "Phylogenies and the comparative method," *American Naturalist*, 125, 1–15. [1030]

Finucane, M. M., Sharpton, T. J., Laurent, T. J., and Pollard, K. S. (2014), "A Taxonomic Signature of Obesity in the Microbiome? Getting to the Guts of the Matter," *PLoS ONE*, 9, e84689. [1026]

Garamszegi, L. Z. (2014), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, New York: Springer. [1030]

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. [1023]

Kinross, J. M., Darzi, A. W., and Nicholson, J. K. (2011), "Gut Microbiome-Host Interactions in Health and Disease," *Genome Medicine*, 3, 14. [1027]

Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C., and Knight, R. (2011), "Human-Associated Microbial Signatures: Examining their Predictive alue," *Cell Host & Microbe*, 10, 292–296. [1022]

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011), "Succession of Microbial Consortia in the Developing Infant Gut Microbiome," *Proceedings of the National Academy of Sciences*, 108, 4578–4585. [1022]

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., and Knight, R. (2011), "Experimental and Analytical Tools for Studying the Human Microbiome," *Nature Reviews Genetics*, 13, 47–58. [1022]

Ley, R. E. (2010), "Obesity and the Human Microbiome," *Current Opinion in Gastroenterology*, 26, 5–11. [1026,1027]

Lin, W., Shi, P., Feng, R., and Li, H. (2014), "Variable Selection in Regression with Compositional Covariates," *Biometrika*, 101, 785–797. [1023,1030]

Lopetuso, L. R., Scaldaferri, F., Petito, V., and Gasbarrini, A. (2013), "Commensal Clostridia: Leading Players in the Maintenance of Gut Homeostasis," *Gut pathogens*, 5, 1. [1030]

Lozupone, C., and Knight, R. (2005), "UniFrac: A new Phylogenetic Method for Comparing Microbial Communities," *Applied and Environmental Microbiology*, 71, 8228–8235. [1022]

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. et al., (2013), "Advancing our Understanding of the Human Microbiome using QIIME," *Methods in Enzymology*, 531, 371–444. [1026]

Nielsen, R. (2005), *Statistical Methods in Molecular Evolution*, Vol. 6, New York: Springer. [1023]

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C. et al., (2009), "The NIH Human Microbiome Project," *Genome Research*, 19, 2317–2323. [1022]

Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A., and Versalovic, J. (2009), "Metagenomic Pyrosequencing and Microbial Identification," *Clinical Chemistry*, 55, 856–866. [1022]

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010), "FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments," *PloS ONE*, 5, e9490. [1026,1030]

Rosenbaum, M., Knight, R., and Leibel, R. L. (2015), "The Gut Microbiota in Human Energy Homeostasis and Obesity," *Trends in Endocrinology & Metabolism*, 26, 493–501. [1026]

Sartor, R. B. (2008), "Microbial Influences in Inflammatory Bowel Diseases," *Gastroenterology*, 134, 577–594. [1022]

Savage, D. C. (1977), "Microbial Ecology of the Gastrointestinal Tract," *Annual Reviews in Microbiology*, 31, 107–133. [1022]

She, Y. (2010), "Sparse Regression with Exact Clustering," *Electronic Journal of Statistics*, 4, 1055–1096. [1023]

Stamatakis, A. (2014), "RAxML version 8: a Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies," *Bioinformatics*, 30, 1312–1313. [1030]

Tanaseichuk, O., Borneman, J., and Jiang, T. (2014), "Phylogeny-Based Classification of Microbial Communities," *Bioinformatics*, 30, 449–456. [1022,1024]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, 58, 267–288. [1025]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society: Series B*, 67, 91–108. [1023]

Tibshirani, R. J., and Taylor, J. (2011), "The Solution Path of the Generalized Lasso," *The Annals of Statistics*, 39, 1335–1371. [1024,1026]

Tibshirani, R. J., and Taylor, J. (2012), "Degrees of Freedom in Lasso Problems," *The Annals of Statistics*, 40, 1198–1232. [1026]

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P. et al., (2008), "A Core Gut Microbiome in Obese and Lean Twins," *Nature*, 457, 480–484. [1026,1027]

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007), "Naive Bayesian Classifier for Rapid Assignment of rrna Sequences into the new Bacterial Taxonomy," *Applied and Environmental Microbiology*, 73, 5261–5267. [1026]

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. [1026]

Wu, G. D., Lewis, J. D., Hoffmann, C., Chen, Y.-Y., Knight, R., Bittinger, K., Hwang, J., Chen, J., Berkowsky, R., Nessel, L. et al., (2010), "Sampling and Pyrosequencing methods for Characterizing Bacterial Communities in the Human Gut using 16S Sequence Tags," *BMC Microbiology*, 10, 206. [1026]

Wu, M., and Eisen, J. A. (2008), "A Simple, Fast, and Accurate Method of Phylogenomic Inference," *Genome Biology*, 9, R151. [1022]

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013), "A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis," *Biometrics*, 69, 1053–1063. [1023]

Zhang, Q., Abel, H., Wells, A., Lenzini, P., Gomez, F., Province, M. A., Templeton, A. A., Weinstock, G. M., Salzman, N. H., and Borecki, I. B. (2015), "Selection of Models for the Analysis of Risk-Factor Trees: Leveraging Biological Knowledge to Mine Large Sets of Risk Factors with Application to Microbiome Data," *Bioinformatics*, 31, 1607–1613. [1023]