Biostatistics (2013), **14**, 2, *pp.* 244–258
doi:10.1093/biostatistics/kxs038
Advance Access publication on October 15, 2012

Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis

JUN CHEN

Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

FREDERIC D. BUSHMAN

Department of Microbiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

JAMES D. LEWIS, GARY D. WU

Division of Gastroenterology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

HONGZHE LI*

Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

hongzhe@upenn.edu

Summary

Motivated by studying the association between nutrient intake and human gut microbiome composition, we developed a method for structure-constrained sparse canonical correlation analysis (ssCCA) in a high-dimensional setting. ssCCA takes into account the phylogenetic relationships among bacteria, which provides important prior knowledge on evolutionary relationships among bacterial taxa. Our ssCCA formulation utilizes a phylogenetic structure-constrained penalty function to impose certain smoothness on the linear coefficients according to the phylogenetic relationships among the taxa. An efficient coordinate descent algorithm is developed for optimization. A human gut microbiome data set is used to illustrate this method. Both simulations and real data applications show that ssCCA performs better than the standard sparse CCA in identifying meaningful variables when there are structures in the data.

Keywords: Dimension reduction; Graph; Phylogenetic tree; Regularization; Variable selection.

1. Introduction

A microbiome is a collection of micro-organisms (mostly bacteria) in a certain environment such as the human gut. The development of next generation sequencing methods such as 454 pyrosequencing and

*To whom correspondence should be addressed.

Solexa sequencing enables researchers to study the microbiome composition by directly sequencing the environmental DNAs. A commonly used sequencing strategy is to sequence a variable region of the 16S ribosomal RNA (rRNA) gene in the bacterial genome, and this variable region can be used for taxonomic classification by comparing it with existing 16S rRNA gene databases. Such 16S data can eventually produce a taxonomic profile for each sample, that is, the abundance for all the identified taxa. However, bacterial taxa are not independent of one another and are related evolutionarily by a phylogenetic tree. Taxa that are phylogenetically close usually behave similarly or have similar biological functions. Such phylogenetic tree information has been effectively utilized in the commonly used UniFrac distance between two microbiome samples (Lozupone and Knight, 2005). In an attempt to visualize the human gut microbiomes from different samples, Purdom (2011) proposed a phylogenetic tree-based principle component analysis (PCA) on the 16S data set. This phylogenetic PCA was shown to separate the environmental samples in a biologically more sensible way than the standard PCA.

In this paper, we consider another commonly used dimension-reduction method, canonical correlation analysis (CCA), that can be used to relate the bacteria taxa with environmental covariates when the number of covariates is large. Our motivating example is a data set generated from a human gut microbiome study at the University of Pennsylvania, where we aim to associate nutrient intake to the bacterial composition in the human gut (see Section 6 for details). We have both the nutrient intake data and the bacterial abundance data measured on 99 individuals and are interested in selecting the bacterial taxa and nutrients that are most closely correlated. CCA aims to identify the linear combinations of two sets of variables that are maximally correlated with each other and provides an important tool to summarize the overall dependency structures between the two sets of variables. It has been applied to linking two sets of high-dimensional genomic data measured on the same set of samples (Parkhomenko *and others*, 2009).

The standard CCA, however, does not perform variable selection and hence usually lacks biological interpretability, especially when the dimension of variables is high. When the number of variables exceeds the number of observations, CCA cannot be applied directly due to singularity of the covariance matrix. To overcome these two major limitations, various types of sparse CCA (sCCA) have been proposed and developed and applied to genomic data analysis (Parkhomenko *and others*, 2009; Witten *and others*, 2009). In sCCA, a sparsity penalty function such as the l_1 penalty is often imposed on the linear coefficients in order to explain the correlation between two data sets using the least number of variables. The sparsity constraint in sCCA not only makes the computation feasible but also increases the biological interpretability of the selected variables.

Available approaches to sCCA do not, however, exploit the prior structure information among the variables. In many applications, there exists some structure among the set of variables in the CCA analysis. These structures can be simple group structures such as gene sets or graphical structures such as gene networks in genomic studies. By including this prior structure information of the data, one can gain better biological insight from the analysis. This has been clearly demonstrated in sparse regression analysis (Li and Li, 2008).

In this paper, we utilize the phylogenetic tree structure of the data from human microbiome studies in CCA analysis. The phylogenetic information from the bacterial taxa could guide us to select relevant taxa in the context of CCA by inducing a tendency to select closely related taxa together, since these taxa are very likely to be associated with the covariates in a similar fashion. We propose to develop a structure-constrained sCCA (ssCCA), where we impose an additional structure-constrained penalty function based on the phylogenetic tree structure. The ssCCA extends the sCCA formulation of Witten *and others* (2009) by imposing a smoothness penalty for the loading coefficients of the taxa based on their closeness on the phylogenetic tree. We also develop an efficient coordinate descent algorithm to implement ssCCA. Our simulations that mimic real microbiome data demonstrate that ssCCA can result in much better performance in selecting bacteria that are associated with other environmental variables. Our analysis of the microbiome and nutrient data has concluded that fat-related nutrients are closely related to human gut microbiome composition, a conclusion that agrees with a previous analysis of the data set (Wu *and others*, 2011).

The rest of the paper is organized as follows. The data structure from 16S microbiome and the concept of phylogenetic tree-structured data are presented in Section 2. A brief review of CCA and the formulation of ssCCA is given in Section 3. Details of the coordinate descent algorithm are presented in Section 4. Results from simulation studies to evaluate our method are given in Section 5. An application to a real human microbiome study to associate nutrient intake with bacterial abundance is presented in Section 6. Finally, a brief discussion of the methods and results is presented in Section 7.

2. 16S microbiome data processing, phylogenetic tree and Laplacian matrix

Typical gut microbiome study involves the collection of fecal samples, isolating all bacterial DNA and then sequencing it using next generation sequencing machines such as the 454 genome sequencer. Since each bacterial cell is assumed to have the same number of copies of this gene, the basic step of a 16S microbiome study is to count different versions of the sequences, and then to identify to which bacteria the versions correspond; in this way, the types and abundance of different bacteria in a sample are determined. After preprocessing of the raw sequences, the 16S sequences are either mapped to an existing phylogenetic tree in a taxonomy-dependent way (Matsen *and others*, 2010) or clustered into operational taxonomic units (OTUs) at a certain similarity level in a taxonomic-independent way (Caporaso *and others*, 2010). At the 97% similarity level, these OTUs are used to approximate the biological species.

The method proposed in this paper is mainly applied to OTU-based 16S data where each of the N 16S sequences belongs to one of p OTUs. Each OTU is characterized by a representative DNA sequence and can be assigned a taxonomic lineage by comparison to a known bacterial 16S rRNA database (Wang *and others*, 2007). Most species-level OTUs are in extremely low abundance with a large proportion of OTUs being simply singletons, possibly due to a sequencing error. We can further aggregate the OTUs from the same genus to form genus-level OTUs and perform analysis at the genus level, which is more robust to sequencing error and can reduce the number of variables significantly. A distance between any two OTUs can be computed using the OTU representative sequences based on an evolution model such as the Jukes–Cantor, Kimura, and Felsenstein model, and a phylogenetic tree for the OTUs can be built based on these distances (Felsenstein, 2003).

Let $\mathbf{x} = (x_1, x_2, \dots, x_p)^{\mathrm{T}}$ represent the vector of the relative abundance of p OTUs obtained from the 16S sequencing, where each OTU is a leaf node of a phylogenetic tree of all the OTUs. We first construct an adjacency matrix using a pairwise distance matrix between any two OTUs. With the given phylogenetic tree, we can use the patristic distance, which is the sum of the branch lengths linking the two OTUs. The distance is usually normalized to the scale of [0,1], with 0 for identity and 1 for complete difference. Denote by d_{jk} the distance between OTU j and OTU k. We then form a $p \times p$ adjacency matrix $\mathbf{A}$ with the diagonal elements of 1 and the jkth element between OTUs j and k defined as $a_{jk} = \phi(d_{jk})$, where ϕ is some decreasing function. Several possible functions are $a_{jk} = (1 - d_{jk}^m)$, $a_{jk} = \exp(-d_{jk}^m)$, or $a_{jk} = 1/d_{jk}^m$, where the power $m > 0$ determines how much weight one puts on closely related OTUs. In this paper, we define the adjacent matrix as

$$a_{jk} = 1/d_{jk}^2 \quad \text{for } j \neq k. \tag{2.1}$$

By taking the square of d_{jk}, large edge weight is given to closely related OTUs and meanwhile, the edge weights for distantly related OTUs are made small. As shown later, this adjacent matrix is only used in the definition of the smoothness penalty. The choice of the adjacent matrix definition should not greatly affect the variable selection results.

The phylogenetic tree is a special case of general undirected graphs, and the adjacency matrix is related to the Laplacian matrix associated with the graph. For a given adjacency matrix $\mathbf{A}$, define

$\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_p)$, where $d_j = \sum_{k=1}^{p} a_{jk}$. The associated Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$ (Chung, 1997). The Laplacian matrix $\mathbf{L}$ is associated with a labeled weighted graph $\mathcal{G} = (V, \mathcal{E}, w)$ with vertex set $V = 1, \ldots, p$ and edge set $\mathcal{E} = \{(j,k) : (j,k) \in V \times V\}$. Here a_{jk} is the weight of edge (j,k) and d_j is the degree of vertex j. For a given vector $\mathbf{u}$, it is easy to show that

$$\mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u} = \sum_{1 \leqslant j < k \leqslant p} a_{jk}(u_j - u_k)^2, \tag{2.2}$$

which measures the smoothness of the vector $\mathbf{u}$ with respect to the labeled weighted graph $\mathcal{G}$. Based on this interpretation, Li and Li (2008) proposed a smoothness penalty of the form $\mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u}$ in high-dimensional regression settings. The structure constraint displays a local smoothing effect by encouraging the variables that are linked on the prior graphical structure to have similar coefficients. In the next section, we extend sCCA to include this smoothness penalty to further encourage some smoothness of the coefficients in linear projections.

3. Structure-constrained sparse canonical correlation analysis

We consider the CCA between two random vectors $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)^{\mathrm{T}}$ and $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q)^{\mathrm{T}}$, where vector $\mathbf{x}$ contains an abundance of p OTUs on a given phylogenetic tree and $\mathbf{y}$ is the q-dimensional vector of the environmental covariates. Suppose that we have collected n i.i.d. samples of $\mathbf{x}$ and $\mathbf{y}$, denoted by $\mathbf{X}$ and $\mathbf{Y}$, respectively. Assume both are column-standardized to have mean 0 and variance 1. Let $\mathbf{A}$ be the adjacency matrix defined in the previous section based on the phylogenetic tree structure and $\mathbf{L}$ be the corresponding Laplacian matrix.

CCA aims to find two projection directions $\mathbf{u}_1 \in \mathbb{R}^p$ and $\mathbf{v}_1 \in \mathbb{R}^q$ so that

$$(\mathbf{u}_1, \mathbf{v}_1) = \arg\max_{\mathbf{u},\mathbf{v}} \text{Corr}(\mathbf{u}^{\mathrm{T}}\mathbf{x}, \mathbf{v}^{\mathrm{T}}\mathbf{y}) = \arg\max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{v}}{\sqrt{(\mathbf{u}^{\mathrm{T}}\Sigma_{\mathbf{xx}}\mathbf{u})(\mathbf{v}^{\mathrm{T}}\Sigma_{\mathbf{yy}}\mathbf{v})}},$$

where $\Sigma_{\mathbf{xx}}$, $\Sigma_{\mathbf{yy}}$, and $\Sigma_{\mathbf{xy}}$ are covariance and cross-covariance matrices. This maximization is equivalent to

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{v} \quad \text{subject to } \mathbf{u}^{\mathrm{T}}\Sigma_{\mathbf{xx}}\mathbf{u} = 1 \quad \text{and} \quad \mathbf{v}^{\mathrm{T}}\Sigma_{\mathbf{yy}}\mathbf{v} = 1. \tag{3.1}$$

Here $\mathbf{u}_1, \mathbf{v}_1$ are called the first pair of canonical vectors, while the new variables $\eta_1 = \mathbf{u}_1{}^{\mathrm{T}}\mathbf{x}$, $\xi_1 = \mathbf{v}_1{}^{\mathrm{T}}\mathbf{y}$ are called the first pair of canonical variables or latent variables and $\rho_1 = \text{Corr}(\eta_1, \xi_1)$ is referred to as the first canonical correlation. When data are available, one estimates $\mathbf{u}_1$ and $\mathbf{v}_1$ by replacing $\Sigma_{\mathbf{xy}}$, $\Sigma_{\mathbf{xx}}$, and $\Sigma_{\mathbf{yy}}$ by the observed sample cross-covariance and covariance matrices $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, and $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$, respectively.

When the dimensions p and q are high, regularization is required in order to obtain a unique solution to the optimization problem (3.1). Given the tuning parameters $c_1 > 0$, $c_2 > 0$, $c_3 > 0$, we propose the following ssCCA criterion that extends the sCCA of Witten *and others* (2009):

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \quad \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v} \\ &\text{subject to} \quad \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} \leqslant 1, \ \mathbf{v}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{v} \leqslant 1, \ \|\mathbf{u}\|_1 \leqslant c_1, \ \|\mathbf{v}\|_1 \leqslant c_2, \ \mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u} \leqslant c_3, \end{aligned} \tag{3.2}$$

where $\|\mathbf{u}\|_1 = \sum_{i=1}^{p} |u_i|$ and $\|\mathbf{v}\|_1 = \sum_{i=1}^{p} |v_i|$ are sparsity l_1 penalty functions. Different from the sCCA formulation, we impose another structure constraint on the coefficient vector $\mathbf{u}$ through the quadratic Laplacian quantity defined in (2.2), $\mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u} \leqslant c_3$. This constraint encourages smoothness of the estimated coefficients of the OTUs that are closely related on the phylogenetic tree. A smaller value of the tuning parameter c_3 results in a smoother estimate of the coefficient vector $\mathbf{u}$ over the phylogenetic tree.

It has been shown that in other high-dimensional problems, treating the covariance matrix as diagonal can yield good results (Tibshirani *and others*, 2003; Witten *and others*, 2009). For this reason, rather than using (3.2) as our ssCCA criterion, following the same strategy adopted by many of the existing sCCA algorithms (Parkhomenko *and others*, 2009; Witten *and others*, 2009), we substitute in the identity matrix I for $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ in the ssCCA formulation (3.2), which gives the ssCCA formulation that we use in this paper:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v} \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1,\ \|\mathbf{v}\|_2^2 \leqslant 1,\ \|\mathbf{u}\|_1 \leqslant c_1,\ \|\mathbf{v}\|_1 \leqslant c_2,\ \mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u} \leqslant c_3. \tag{3.3}$$

4. Coordinate descent algorithm for the ssCCA

4.1 *Algorithm to obtain the first ssCCA factor*

To facilitate computation, we write constraints on $\mathbf{u}$ in Lagrangian form and the ssCCA criterion (3.3) becomes:

$$\min_{\mathbf{u},\mathbf{v}} \left(-\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v} + \frac{1}{2}\|\mathbf{u}\|_2^2 + \lambda_1\|\mathbf{u}\|_1 + \frac{\lambda_2}{2}\mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u}\right)$$
$$\text{subject to} \quad \|\mathbf{v}\|_2^2 \leqslant 1, \|\mathbf{v}\|_1 \leqslant c_2, \tag{4.1}$$

where $\lambda_1 \geqslant 0$, $\lambda_2 \geqslant 0$, and $c_2 > 0$ are tuning parameters. Note that when $\lambda_2 = 0$, ssCCA is reduced to sCCA. Since the Laplacian penalty function $(\lambda_2/2)\mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u}$ is convex in $\mathbf{u}$, the criterion (4.1) remains biconvex in $\mathbf{u}$ and $\mathbf{v}$, such that we can still use an iterative method to solve this optimization problem.

Algorithm to obtain the first ssCCA factor

(1) Initialize $\mathbf{v}$ as the first right singular vector with unity l_2 norm from the singular value decomposition of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

(2) Iterate until convergence:

(a)

$$\mathbf{u} \leftarrow \arg\min_{\mathbf{u}} \left(-\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v} + \frac{1}{2}\|\mathbf{u}\|_2^2 + \lambda_1\|\mathbf{u}\|_1 + \frac{\lambda_2}{2}\mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u}\right),$$

which can be solved by a graph-constrained regression problem (Li and Li, 2008):

$$\mathbf{u} \leftarrow \arg\min_{\mathbf{u}} \left(\frac{1}{2}\|\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v} - \mathbf{u}\|_2^2 + \lambda_1\|\mathbf{u}\|_1 + \frac{\lambda_2}{2}\mathbf{u}^{\mathrm{T}}\mathbf{L}\mathbf{u}\right).$$

(b) $\mathbf{v} \leftarrow \arg\min_{\mathbf{v}} -\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v}$ subject to $\|\mathbf{v}\|_2^2 \leqslant 1$, $\|\mathbf{v}\|_1 \leqslant c_2$, which is given by

$$\mathbf{v} \leftarrow \frac{S\{(\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y})^{\mathrm{T}}, \delta\}}{\|S\{(\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y})^{\mathrm{T}}, \delta\}\|_2},$$

where $S(.,.)$ is the soft-thresholding function, i.e.

$$S(a,b) = \begin{cases} \operatorname{sgn}(a)(|a| - b) & \text{if } |a| > b, \\ 0 & \text{Otherwise,} \end{cases}$$

and $\delta = 0$ if this results in $\|\mathbf{v}\|_1 \leqslant c_2$; otherwise, δ is chosen so that $\|\mathbf{v}\|_1 = c_2$. The choice of δ can be determined using a binary search (Witten *and others*, 2009).

Let $\mathbf{L} = \mathbf{U}\Gamma\mathbf{U}^{\mathrm{T}}$ and $\mathbf{S} = \mathbf{U}\Gamma^{1/2}$. Then step 2(a) can be converted into a simple Lasso problem as in Li and Li (2008):

$$\mathbf{u} \leftarrow \arg\min_{\mathbf{u}} \left(\frac{1}{2} \|\mathbf{A}^{*}\mathbf{u} - \mathbf{b}^{*}\|_2^2 + \lambda_1 \|\mathbf{u}\|_1 \right),$$

where

$$\mathbf{A}^{*}_{2p\times p} = \begin{pmatrix} \mathbf{I}_{p\times p} \\ \sqrt{\lambda_2}\mathbf{S}^{\mathrm{T}} \end{pmatrix}, \quad \mathbf{b}^{*}_{2p} = \begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v} \\ \mathbf{0}_p \end{pmatrix},$$

$\mathbf{I}_{p\times p}$ is a $p \times p$ identity matrix, and $\mathbf{0}_p$ is a p-dimensional vector of 0's. Note that no intercept is included in this Lasso problem and a coordinate descent algorithm can be implemented to obtain the solution at a given λ_1 (Friedman *and others*, 2007).

Though the objective function is biconvex, i.e. is convex in either $\mathbf{u}$ or $\mathbf{v}$, it is not convex in $(\mathbf{u}^{\mathrm{T}}, \mathbf{v}^{\mathrm{T}})^{\mathrm{T}}$, so the coordinate descent algorithm does not necessarily converge to the global optimum; however, by using the first right singular vector of the covariance matrix as the initial starting point, it does converge to a stationary point (Tseng and Yun, 2009) and interpretable solutions.

4.2 *Choosing tuning parameters*

The tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, c_2)$ control the model complexity and have to be tuned. We use an M-fold two-stage cross-validation (CV) method to choose $\boldsymbol{\lambda}$. First, we divide all the samples into M disjoint subgroups, also known as folds, and denote the index of samples in the mth fold by I_m for $m = 1, \ldots, M$. The M-fold cross-validated function is defined as

$$\mathrm{CV}(\boldsymbol{\lambda}) = \frac{1}{M} \sum_{m=1}^{M} \mathrm{Corr}\{\mathbf{X}_m^{\mathrm{T}}\hat{\mathbf{u}}_{-m}(\boldsymbol{\lambda}), \mathbf{Y}_m^{\mathrm{T}}\hat{\mathbf{v}}_{-m}(\boldsymbol{\lambda})\}, \tag{4.2}$$

where Corr(., .) is the correlation function and $\hat{\mathbf{u}}_{-m}(\boldsymbol{\lambda}), \hat{\mathbf{v}}_{-m}(\boldsymbol{\lambda})$ is the estimate of $\mathbf{u}, \mathbf{v}$ based on the samples $(\bigcup_{m=1}^{M} I_m)\backslash I_m$ with $\boldsymbol{\lambda}$ as the tuning parameter. It is well known that CV can perform poorly in tuning parameter selection for problems involving l_1 penalties due to biases in parameter estimates (Meinshausen and Bühlmann, 2006). To reduce the shrinkage problem, we reestimate the non-zero coefficients without penalization by performing singular value decomposition on the training data set excluding the variables with zero coefficients in the penalized procedure. Specifically, for a given tuning parameter $\boldsymbol{\lambda}$, we recalculate the loading coefficients using the variables that are selected by ssCCA and use these coefficients in the CV score (4.2). This avoids bias of the estimates due to penalization. We then choose $\boldsymbol{\lambda}^{*} = \mathrm{argmax}_{\boldsymbol{\lambda}}\mathrm{CV}(\boldsymbol{\lambda})$ as the best tuning parameters. From our simulations, we observe that the two-stage CV procedure almost always performs better than standard CV without reestimating the parameters.

5. Simulation studies

We present Monte Carlo simulations to evaluate ssCCA in identifying the relevant variables that explain the correlation between two multivariate vectors. The solution of sCCA is obtained by setting $\lambda_2 = 0$ in ssCCA. The simulations are carried out to mimic an association study between nutrient intake and genus-level OTU abundance that is presented in Section 6. Since the phylogenetic tree implies distances between the OTUs, we simulate the distance matrix directly. Specifically, since OTUs are often clustered on the phylogenetic tree, we generate random OTU clusters of size 1–15, where the OTU cluster members are sequentially indexed. If two OTUs are from the same cluster (e.g. from the same taxonomic rank *family*), their distance is drawn from a uniform distribution on (0.1, 0.2); if two OTUs are from different clusters, then their distance is drawn from a uniform distribution on (0.2, 1). We then construct the adjacency matrix $\mathbf{A}$ using the method (2.1) based on the distances.

5.1 *Simulation based on a latent variable model*

We use a latent variable model to generate the data matrices $\mathbf{X}$ and $\mathbf{Y}$ where the dependency between these two sets of variables is induced by a latent random variable ζ and the variances in $\mathbf{x}$, $\mathbf{y}$ can be explained in part by ζ. We assume $\mathbf{x} = \zeta \mathbf{w_x} + \boldsymbol{\epsilon}_\mathbf{x}$ and $\mathbf{y} = \zeta \mathbf{w_y} + \boldsymbol{\epsilon}_\mathbf{y}$, where $\zeta \sim N(0, \sigma_\zeta^2)$, $\boldsymbol{\epsilon}_\mathbf{x}$, $\boldsymbol{\epsilon}_\mathbf{y}$ are random noise vectors that follow $\boldsymbol{\epsilon}_\mathbf{x}, \boldsymbol{\epsilon}_\mathbf{y} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$, and $\mathbf{w_x} \in \mathbb{R}^p$, $\mathbf{w_y} \in \mathbb{R}^q$ are column vectors of preset weights. The $\sigma_\epsilon/\sigma_\zeta$ ratio controls the overall association strength between $\mathbf{x}$ and $\mathbf{y}$, with a small value indicating strong association. The coefficients $\mathbf{w_x}$ and $\mathbf{w_y}$ control the relative contributions of individual variables to the overall association. We assume that only the first $p_\mathbf{x} = 10$ elements of $\mathbf{w_x}$ and the first $q_\mathbf{y} = 10$ elements of $\mathbf{w_y}$ are non-zero and take the values of $(0.1, \ldots, 0.1)$ and $(0.08, 0.084, 0.089, \ldots, 0.12)$, respectively. In addition, we let w_i and w_j be identical or similar if $\mathbf{x}_i$ and $\mathbf{x}_j$ are from the same cluster of the phylogenetic leaf nodes. We consider the scenarios where we have one relevant cluster of size 10, two relevant clusters of size 5 and 5, and three relevant clusters of size 3, 4, and 4. The highest correlation between linear combinations of $\mathbf{x}$ and $\mathbf{y}$ is given by Parkhomenko *and others* (2009):

$$\rho_{\max} = \frac{\sigma_\zeta^2}{\sqrt{(\sigma_\zeta^2 + p_\mathbf{x}\sigma_\epsilon^2)(\sigma_\zeta^2 + p_\mathbf{y}\sigma_\epsilon^2)}}. \tag{5.1}$$

We fix $\sigma_\epsilon^2 = 1$ and vary σ_ζ^2 to control the strength of the canonical correlation. When $\sigma_\zeta = 5$, $\rho_{\max} = 0.7$.

5.2 *Evaluation of the selection performance*

We evaluate the performance of our methods in terms of selecting the relevant variables that lead to correlation between random vectors $\mathbf{x}$ and $\mathbf{y}$ by considering models with various combinations of the parameters. For each simulated data set, we use 5-fold two-stage CV to select the tuning parameter values and then compute the true positive rate (TPR), false positive rate (FPR), and Matthew's correlation coefficient (MCC) to measure the selection performance for both $\mathbf{x}$ and $\mathbf{y}$. These three measures are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \quad \mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})},$$

where TP, FP, TN, and FN are true positives, false positives, true negatives, and false negatives, respectively. For each model, we generate the observed data set $\mathbf{X}$ and $\mathbf{Y}$ 100 times and summarize the TPR, FPR, and MCC as averages over 100 runs. Results from 10-fold two-stage CV are very similar and are omitted here.

We also compare the performance of different methods using the receiver operating characteristic (ROC) curve (FPR against TPR) for identifying the relevant taxa OTUs by varying the tuning parameters. Specifically, the three tuning parameters are searched over a 10 × 10 × 10 grid for a total of 1000 tuning parameter combinations. For each combination, we obtain the FPR and the TPR, which represents one point in the ROC plot. The ROC curve is then obtained by joining these points for each run. We then average the ROC curves over 100 runs to produce an average ROC curve.

5.3 *Comparison of ssCCA and sCCA under one latent variable model*

We consider models with various combinations of the parameters (labeled A1–D2), including the number of relevant OTU clusters, the signal strength as measured by σ_ζ^2 and the dimensions p and q and present the results in Table 1 and Figure 1. We observe that the advantage of ssCCA over sCCA is more obvious under weak association (Model A1). As the signal becomes stronger, the performance of sCCA becomes closer to ssCCA (Model A2). This agrees with our intuition: the advantage of ssCCA lies in borrowing information from closely related OTUs and, when the association is weak, pooling information across closely related OTUs can improve the OTU selection. Another interesting observation is that better selection of OTUs can lead to better selection of nutrients, which is best shown in the weak association case by obtaining a higher MCC. We also observe that as the dimension increases, both ssCCA and sCCA become less efficient in selecting relevant OTUs and nutrients (Models B1 and B2). However, ssCCA performs consistently better than sCCA in all dimensions considered. Finally, as the cluster size decreases, we do not see a significant deterioration of the selection performance of ssCCA (Models C1 and C2). ssCCA still performs better than sCCA. As long as the cluster contains more than one OTU, using structure information always improves variable selection.

Since the smoothness penalty encourages the variables that are close on the phylogenetic tree to have similar linear projection coefficients, we evaluate the sensitivity of ssCCA when this assumption does not hold. We investigate the performance of ssCCA when data contradict with our smoothness assumption. We consider the model where the first 10 elements of $\mathbf{w_x}$ have different coefficients but with the same signs, and take values that are equally spaced on $[0.08, 0.12]$ (Model D1). The performance of ssCCA is still much better than sCCA. Model D2 considers the scenario when the first five and the second five elements of $\mathbf{w_x}$ are 0.1s and -0.1s, respectively, where the coefficients are different and have different signs. This scenario violates our model assumption that closely linked OTUs have similar coefficients. The structure-constrained penalty now has an adverse effect. This is clearly seen in the ROC plot (Figure 1(D2)). However, when the CV procedure is applied to select the tuning parameters and the corresponding OTUs and nutrients, the performance of ssCCA and sCCA is very similar (Table 1). This is because if the prior structure information is not useful, CV procedure tends to select $\lambda_2 = 0$, which reduces ssCCA to sCCA. Therefore, the selection performance of ssCCA should be at least as good as sCCA, but ssCCA performs better when the prior assumption holds.

5.4 *Comparison of ssCCA and sCCA under complex models*

We compare the performance of ssCCA and sCCA under several complex models and also present the results in Table 1 and Figure 1. Under Model E, we consider the scenario when the noises are correlated with correlation $0.4^{|i-j|}$ for $\epsilon_i$and ϵ_j for both $\mathbf{x}$ and $\mathbf{y}$, where the OTU cluster members have sequential index numbers. The performances of ssCCA and sCCA are both slightly worse when compared with Model A2 when the noises are independent; ssCCA still outperforms sCCA.

We then consider Model F where we simulate count data with zeros. Specifically, we first generate the data matrix $\mathbf{X}$ as previously. We then convert it into a proportion matrix $\mathbf{P}$ and generate the counts based on $\mathbf{P}$. For the jth column X_j, we first map the column values into the range of $[0, p_j^{\max}]$ by a linear transformation $p_{ij} = ((x_{ij} - \min_i(x_{ij}))/(\max_i(x_{ij}) - \min_i(x_{ij})))p_j^{\max}$, where $p_j^{\max}$ is sampled from $[0.01, 0.1]$, so the maximum OTU abundance can vary by 10-fold. Rows of $\mathbf{P}$ are further scaled to sum up to 1. Given the OTU proportions for each sample, we generate counts using a Dirichlet-multinomial model with a total count of 1000 and an overdispersion of 0.01. Since we introduce extra variation by simulating counts, we increase the first 10 components of $\mathbf{w_x}$ to 0.4 to achieve a moderate association ($\rho_1 \sim 0.7$). Under this parameter setting, the data matrix contains about 20% 0's. To apply ssCCA and sCCA, we convert the simulated count matrix into a proportion matrix. Table 1(F) and Figure 1(F) again show that ssCCA outperforms sCCA in selecting the relevant variables.

Table 1. *Simulation results to evaluate ssCCA under models of different association signals, dimension sizes, cluster sizes, model misspecification, and complexity*

	Selection of **x** variables			Selection of **y** variables		
Method	TPR-**x**	FPR-**x**	MCC-**x**	TPR-**y**	FPR-**y**	MCC-**y**
A1: one cluster, $\sigma_\zeta = 4$, $p, q = 100$						
ssCCA	0.91 (0.20)	0.07 (0.10)	0.76 (0.22)	0.78 (0.22)	0.12 (0.12)	0.58 (0.18)
sCCA	0.70 (0.31)	0.09 (0.12)	0.56 (0.22)	0.75 (0.24)	0.12 (0.12)	0.54 (0.21)
A2: one cluster, $\sigma_\zeta = 5$, $p, q = 100$						
ssCCA	0.96 (0.10)	0.03 (0.08)	0.89 (0.16)	0.87 (0.17)	0.05 (0.09)	0.78 (0.16)
sCCA	0.89 (0.17)	0.05 (0.08)	0.79 (0.17)	0.87 (0.16)	0.05 (0.09)	0.77 (0.17)
B1: one cluster, $\sigma_\zeta = 5$, $p, q = 200$						
ssCCA	0.98 (0.08)	0.05 (0.11)	0.87 (0.19)	0.87 (0.16)	0.07 (0.11)	0.75 (0.18)
sCCA	0.89 (0.17)	0.09 (0.15)	0.74 (0.22)	0.87 (0.16)	0.08 (0.11)	0.72 (0.20)
B2: one cluster, $\sigma_\zeta = 5$, $p, q = 400$						
ssCCA	0.89 (0.30)	0.06 (0.11)	0.74 (0.33)	0.81 (0.28)	0.12 (0.13)	0.60 (0.30)
sCCA	0.77 (0.32)	0.09 (0.23)	0.66 (0.32)	0.78 (0.31)	0.11 (0.12)	0.57 (0.32)
C1: two clusters, $\sigma_\zeta = 5$, $p, q = 100$						
ssCCA	0.93 (0.14)	0.03 (0.07)	0.88 (0.15)	0.83 (0.16)	0.05 (0.09)	0.76 (0.16)
sCCA	0.87 (0.16)	0.05 (0.08)	0.78 (0.16)	0.85 (0.16)	0.06 (0.10)	0.76 (0.17)
C2: three clusters, $\sigma_\zeta = 5$, $p, q = 100$						
ssCCA	0.94 (0.11)	0.03 (0.07)	0.88 (0.15)	0.88 (0.15)	0.07 (0.11)	0.75 (0.18)
sCCA	0.89 (0.16)	0.05 (0.10)	0.80 (0.18)	0.88 (0.16)	0.07 (0.10)	0.76 (0.18)
D1: one cluster, $\sigma_\zeta = 5$, $p, q = 100$, variable coefficients of the same signs						
ssCCA	0.95 (0.11)	0.02 (0.05)	0.90 (0.13)	0.86 (0.19)	0.06 (0.10)	0.76 (0.17)
sCCA	0.87 (0.15)	0.04 (0.08)	0.79 (0.15)	0.88 (0.16)	0.07 (0.10)	0.75 (0.18)
D2: one cluster, $\sigma_\zeta = 5$, $p, q = 100$, variable coefficient of opposite signs						
ssCCA	0.89 (0.14)	0.05 (0.09)	0.81 (0.17)	0.89 (0.15)	0.08 (0.11)	0.75 (0.20)
sCCA	0.90 (0.15)	0.04 (0.09)	0.82 (0.17)	0.90 (0.14)	0.07 (0.11)	0.76 (0.19)
E: correlated noise, one cluster, $\sigma_\zeta = 5$, $p, q = 100$						
ssCCA	0.92 (0.18)	0.04 (0.07)	0.84 (0.19)	0.78 (0.21)	0.05 (0.08)	0.72 (0.17)
sCCA	0.85 (0.20)	0.05 (0.10)	0.77 (0.18)	0.82 (0.21)	0.06 (0.10)	0.73 (0.18)
F: count data, one cluster, $\sigma_\zeta = 5$, $p, q = 100$						
ssCCA	0.92 (0.16)	0.04 (0.11)	0.84 (0.17)	0.72 (0.26)	0.06 (0.14)	0.71 (0.20)
sCCA	0.72 (0.22)	0.09 (0.15)	0.62 (0.18)	0.80 (0.24)	0.08 (0.16)	0.75 (0.23)
G: two directions, one cluster, $\sigma_\zeta = 5$, $p, q = 100$						
ssCCA	0.95 (0.13)	0.03 (0.08)	0.87 (0.17)	0.85 (0.17)	0.05 (0.09)	0.76 (0.16)
sCCA	0.85 (0.19)	0.07 (0.10)	0.73 (0.17)	0.82 (0.19)	0.06 (0.09)	0.72 (0.16)
H: two directions, two clusters, model misspecification, $\sigma_\zeta = 5$, $p, q = 100$						
ssCCA	0.83 (0.20)	0.05 (0.09)	0.74 (0.19)	0.88 (0.19)	0.11 (0.13)	0.67 (0.20)
sCCA	0.87 (0.18)	0.06 (0.10)	0.76 (0.18)	0.89 (0.17)	0.10 (0.13)	0.69 (0.20)

Five-fold two-stage CV is used to select the tuning parameters. As a comparison, results from sCCA are also presented. Each column represents a measure of selection performance for OTU (**x**) or nutrient (**y**). TPR, true positive rate; FPR, false positive rate; MCC, Matthew's correlation coefficient. The results are averaged over 100 replications with SD indicated in the parenthesis.

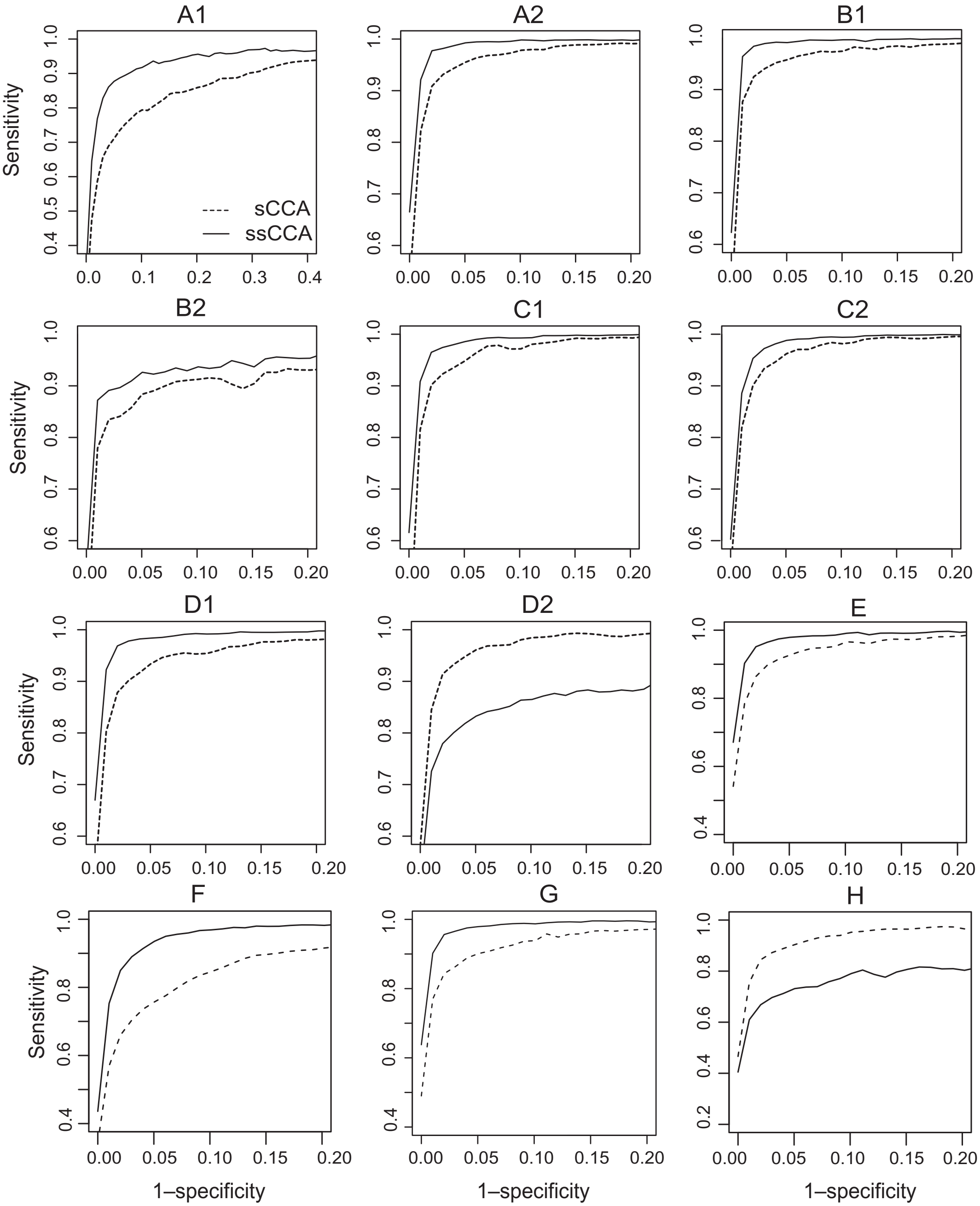

Fig. 1. ROC curves for selecting the OTUs using the ssCCA and sCCA for Models A1–H. The corresponding model parameters are given in Table 1.

Finally, we consider two models where two orthogonal directions induce the correlation between two sets of random vectors. We assume $\mathbf{x} = \zeta_1 \mathbf{w}_{\mathbf{x}}^1 + \zeta_2 \mathbf{w}_{\mathbf{x}}^2 + \boldsymbol{\epsilon}_{\mathbf{x}}$ and $\mathbf{y} = \zeta_1 \mathbf{w}_{\mathbf{y}}^1 + \zeta_2 \mathbf{w}_{\mathbf{y}}^2 + \boldsymbol{\epsilon}_{\mathbf{y}}$, where under Model G, the two directions are given by

$$\mathbf{w}_{\mathbf{x}}^1 = (\underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{0, \ldots, 0}_{90})^{\mathrm{T}}$$

and

$$\mathbf{w}_{\mathbf{x}}^2 = 0.5(\underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{-0.1, \ldots, -0.1}_{5} \underbrace{0, \ldots, 0}_{90})^{\mathrm{T}}.$$

We assume that $\mathbf{w}_{\mathbf{y}}^1$, $\mathbf{w}_{\mathbf{y}}^2$ are the same as $\mathbf{w}_{\mathbf{x}}^1$, $\mathbf{w}_{\mathbf{x}}^2$, and the OTUs from the same cluster have the same coefficients on the first direction. Under Model H, we consider model misspecification where the two directions are given by

$$\mathbf{w}_{\mathbf{x}}^1 = (\underbrace{0.1, 0.1, -0.1, -0.1, 0.1}_{5}, \underbrace{0.1, 0.1, -0.1, -0.1, -0.1}_{5}, \underbrace{0, \ldots, 0}_{90})^{\mathrm{T}}$$

and

$$\mathbf{w}_{\mathbf{x}}^2 = 0.5(\underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{0.1, \ldots, 0.1}_{5} \underbrace{0, \ldots, 0}_{90})^{\mathrm{T}},$$

and $\mathbf{w}_{\mathbf{y}}^1$, $\mathbf{w}_{\mathbf{y}}^2$ are the same as $\mathbf{w}_{\mathbf{x}}^1$, $\mathbf{w}_{\mathbf{x}}^2$. OTUs from the same cluster have coefficients of different signs on the first direction. Under Model H, ssCCA has higher true positive and lower false positive rates and higher area under the ROC curve (Table 1(G) and Figure 1(G)). Under the model misspecification (Model H), the performances of ssCCA and sCCA are comparable.

6. Application to gut microbiome data analysis

We apply ssCCA to a microbiome study on association between the nutrient intake and bacterial abundance in the human gut conducted at the University of Pennsylvania. The human gut is inhabited by trillions of bacterial cells, and some bacterial species have a profound influence on human health and disease. One goal of the study is to investigate the relationship between diet and microbiome composition and to identify a short list of potential nutrients and their associated bacteria in the human gut. For this study, both gut microbiome 16S data and nutrient intake data were available for 99 healthy subjects. Fecal samples were obtained from these 99 subjects and bacterial DNA was extracted using a standard protocol. After multiplexed 454 pyrosequencing, about 900 000 high quality, partial (~ 370 bp) 16S rRNA gene sequences were generated. These sequences were analyzed using the Qiime pipeline (Caporaso *and others*, 2010), where the sequences were clustered at 97% sequence identity into OTUs and assigned a taxonomic identity by comparing to the Ribosomal Database Project reference 16S rRNA database (Wang *and others*, 2007). We consolidated these species-level OTUs into 119 genera (genus-level OTUs) and used the representative sequence from the most abundant species-level OTU as the genus level representative sequence for distance calculation and for construction of the phylogenetic tree. In our analysis, we further excluded the uncommon genera that occurred in less than $\frac{1}{4}$ of the samples; so we only considered $p = 40$ relatively common genera (Figure 2). These 99 subjects also completed a carefully designed food frequency questionnaire (FFQ). Based on the FFQ, the daily intake for $q = 214$ nutrients were calculated for each subject by nutritionists. Because nutrient intake is clearly dependent on the overall energy consumption, we regressed the nutrient intake on the total energy consumption and took residuals as the normalized nutrient intake. Our final data set can be summarized as the OTU abundance matrix $\mathbf{X}_{99\times 40}$ and the nutrient intake

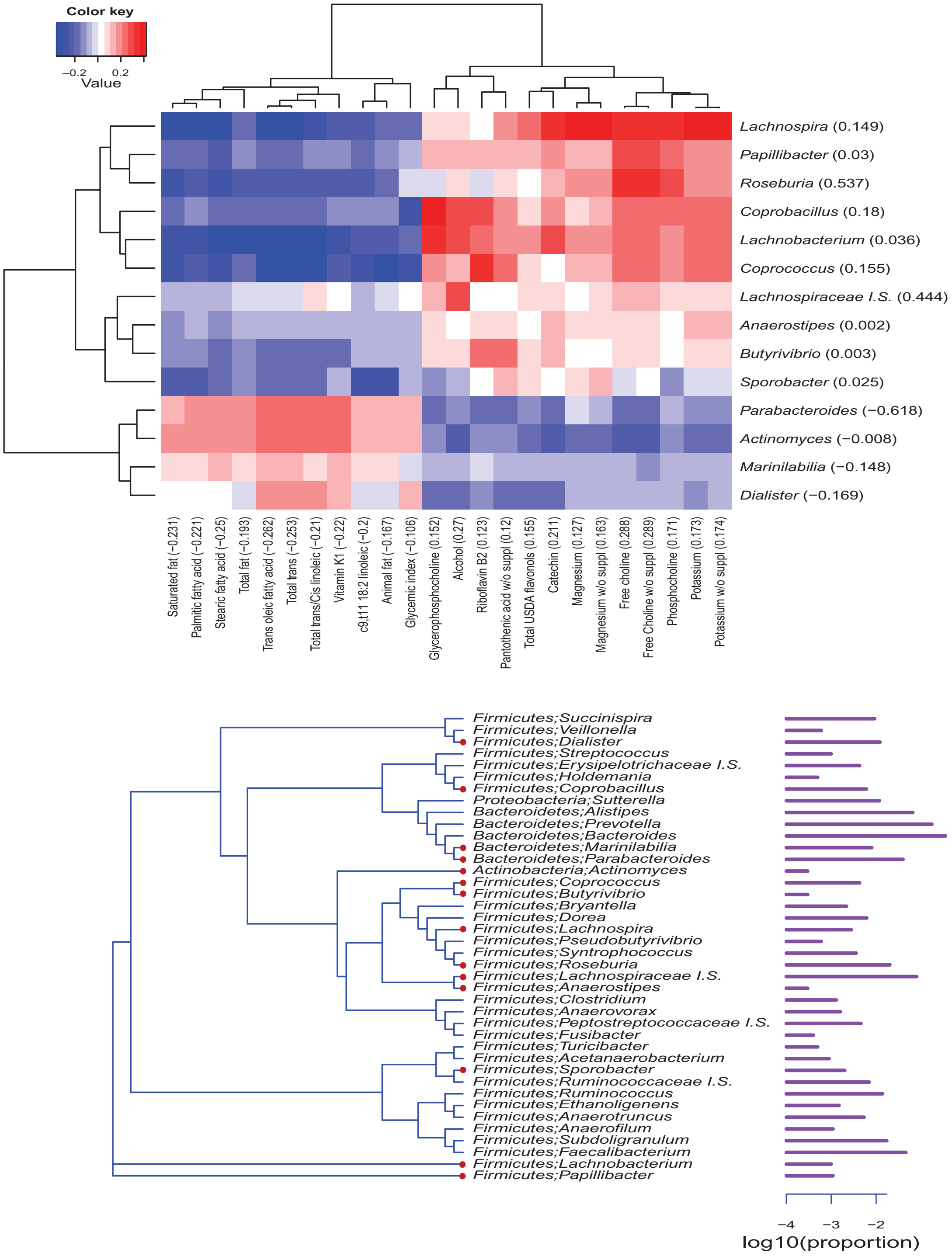

Fig. 2. Analysis of gut microbiome data. Top: heatmap that shows the correlations between the selected genera and nutrients. The number in parenthesis of each variable is the estimated loading coefficient. Red and blue colors indicate positive and negative correlations, respectively. Bottom: Phylogenetic tree of the 40 genera used in the analysis. The genera selected by ssCCA are marked with red circles. The bars on the right side indicate the average relative abundances of these genera on log 10 scale.

matrix $\mathbf{Y}_{99\times 214}$. Since the sampling depths are very different for different samples, we normalize the counts into proportions and standardize the columns to have mean 0 and variance 1.

The goal of our analysis is to investigate the overall association between gut bacteria abundance and nutrient intake. We used the method presented in (2.1) to construct the adjacency matrix $\mathbf{A}$, and the distances between any two OTUs were calculated using the "K80" model (R "ape" package, "dist.dna" function). Five-fold two-stage CV was performed to search the optimal tuning parameters on a grid of 20 × 20 × 20, and the range of the tuning parameters was set to explore all possible models: from the most dense to the most sparse model. We applied ssCCA to the data set and identified 24 nutrients and 14 genera whose linear combinations gave a cross-validated correlation of 0.42 between gut bacterial abundance and nutrients. Figure 2 shows the heatmap of pairwise correlations between these selected nutrients and OTUs, where the estimated loading coefficients are given in parentheses. The signs of the estimated loading coefficients correspond very well to the pairwise correlations. The nutrients related to fats are clustered together, while the other nutrients show association in the opposite direction.

The selected microbiome-associated nutrients are biologically interpretable. More than half of the selected nutrients are related to fat. It has been experimentally shown that fats can change the gut microbiome composition independent of obesity in a mouse study (Hildebrandt *and others*, 2009). There are also four selected nutrients related to choline, and it was found by a recent human microbiome study that the composition of the gastrointestinal microbiome changed with the choline levels of diets (Spencer *and others*, 2011). The selected nutrients are also consistent with the candidate nutrients we identified using a distance-based testing procedure (Wu *and others*, 2011). This procedure utilized the overall UniFrac distances (Lozupone and Knight, 2005) between microbiomes of any two subjects computed using both the OTU abundances and the phylogenetic relationship among them. Twenty out of 24 nutrients selected by ssCCA were in the nutrients selected by the distance-based individual testing method at the false discovery rate of 25%.

The pattern of selected OTUs is also interesting. The selected OTUs are marked with red circles in the phylogenetic tree of Figure 2. We see that the closely related OTUs tend to be selected together; for example, the genus *Parabacteroides* and *Marinilabilia*, *Butyrivibrio* and *Coprococcus*, and *Anaerostipes* and *Lachnospiraceae Incertae Sedis* are all close relatives on the tree. ssCCA tends to select closely related OTUs together by making the coefficients of neighbors similar through imposing a phylogenetic tree-constrained smoothness penalty. This feature of ssCCA can also be viewed as borrowing information from nearby OTUs; that is, if several neighbors all exhibit similar weak association, ssCCA amplifies the signal strength and selects them together. On the other hand, if some OTU exhibits low-level association but all its neighbors show the opposite evidence, ssCCA will not select that OTU.

By a comparison, an sCCA that does not account for the phylogenetic relationship among the OTUs selects only one OTU, the $FirmucuteLachnospira$, which was also selected by ssCCA, but a total of 122 nutrients. The interpretation of the result is not as clear as that from ssCCA. The resulting combinations gave a cross-validated correlation of 0.39, smaller than that obtained from ssCCA.

7. Conclusion and discussion

We have extended the sCCA to incorporate the graphical structure among the variables in CCA. When the number of variables exceeds the number of samples, using prior structure information to guide variable selection is important. The prior knowledge could lead to a solution that is biologically more interpretable. The structured sCCA utilizes the phylogenetic information to select the bacterial OTUs that are associated with covariates. The power of the ssCCA method has been demonstrated in the simulation studies, and its performance is unanimously better than sCCA in all the simulated scenarios when there are structures in the data. Even when the prior information is not completely accurate, our method still performs comparably to sCCA due to selection of the tuning parameter by CV.

One limitation of the ssCCA formulation is that it assumes a linear relationship among the variables, which may not always hold for OTU compositional/abundance data. Our analysis of the gut microbiome data did not indicate too much deviation from the linearity between OTU abundance and nutrient intake. One interesting future research direction is to develop structure-constrained non-linear measures of association and sparse non-linear CCA.

Acknowledgments

We thank two reviewers for their helpful comments. *Conflict of Interest*: None declared.

Funding

This research was supported by the National Institutes of Health (CA127334, GM097505, and DK083981).

References

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I. *and others*. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336.

Chung, F. R. K. (1997). *Spectral Graph Theory*, Volume 92. Providence, RI: American Mathematical Society.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sunderland, Massachusetts: Sinauer Associates.

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1**, 302–332.

Hildebrandt, M. A., Hoffmann, C., Sherrill-Mix, S. A., Keilbaugh, S. A., Hamady, M., Chen, Y. Y., Knight, R., Ahima, R. S., Bushman, F. and Wu, G. D. (2009). High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology* **137**, 1716–1724.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175–1182.

Lozupone, C. and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**, 8228.

Matsen, F., Kodner, R. and Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34**, 1436–1462.

Parkhomenko, E., Tritchler, D. and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology* **8**, 1.

Purdom, E. (2011). Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *Annals of Applied Statistics* **5**, 2326–2358.

Spencer, M. D., Hamp, T. J., Reid, R. W., Fischer, L. M., Zeisel, S. H. and Fodor, A. A. (2011). Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology* **140**, 976–986.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science* **18**, 104–117.

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming Series B* **117**, 387–423.

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.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W., Knights, R. *and others*. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108.

[*Received January 28, 2012; revised September 8, 2012; accepted for publication September 11, 2012*]