Frontiers in Genetics

TYPE Methods
PUBLISHED 20 November 2024
DOI 10.3389/fgene.2024.1489694

OPEN ACCESS

EDITED BY
Weihua Guan,
University of Minnesota Twin Cities,
United States

REVIEWED BY
Qiwei Li,
The University of Texas at Dallas, United States
Liangliang Zhang,
Case Western Reserve University, United States

*CORRESPONDENCE
Xianyang Zhang,
zhangxiany@stat.tamu.edu
Jun Chen,
chen.jun2@mayo.edu

†These authors have contributed equally to this work

RECEIVED 01 September 2024
ACCEPTED 31 October 2024
PUBLISHED 20 November 2024

CITATION
Deng L, Tang Y, Zhang X and Chen J (2024) Structure-adaptive canonical correlation analysis for microbiome multi-omics data . *Front. Genet.* 15:1489694.
doi: 10.3389/fgene.2024.1489694

Structure-adaptive canonical correlation analysis for microbiome multi-omics data

Linsui Deng[1]†, Yanlin Tang[2]†, Xianyang Zhang[3]* and Jun Chen[4]*

[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen, China, [2]School of Statistics, East China Normal University, Shanghai, China, [3]Department of Statistics, Texas A&M University, College Station, TX, United States, [4]Department of Quantitative Health Sciences, Rochester, MN, United States

Sparse canonical correlation analysis (sCCA) has been a useful approach for integrating different high-dimensional datasets by finding a subset of correlated features that explain the most correlation in the data. In the context of microbiome studies, investigators are always interested in knowing how the microbiome interacts with the host at different molecular levels such as genome, methylol, transcriptome, metabolome and proteome. sCCA provides a simple approach for exploiting the correlation structure among multiple omics data and finding a set of correlated omics features, which could contribute to understanding the host-microbiome interaction. However, existing sCCA methods do not address compositionality, and its application to microbiome data is thus not optimal. This paper proposes a new sCCA framework for integrating microbiome data with other high-dimensional omics data, accounting for the compositional nature of microbiome sequencing data. It also allows integrating prior structure information such as the grouping structure among bacterial taxa by imposing a "soft" constraint on the coefficients through varying penalization strength. As a result, the method provides significant improvement when the structure is informative while maintaining robustness against a misspecified structure. Through extensive simulation studies and real data analysis, we demonstrate the superiority of the proposed framework over the state-of-the-art approaches.

KEYWORDS

canonical correlation analysis, compositional effect, phylogenetic tree, structural information, dimension reduction, variable selection

1 Introduction

The human microbiome is the collection of microorganisms and their genetic makeup associated with the human body. It plays a critical role in human health and disease ranging from gastrointestinal diseases to various cancers (Sepich-Poore et al., 2021). To gain more mechanistic insights, multi-omics approaches have been increasingly employed in microbiome studies to elucidate the intricate interplay between the environment, the human microbiome and the host at different molecular levels (Hasin et al., 2017; Lloyd-Price et al., 2019). Although many multi-omics datasets have been generated in the past few years, it is unclear how to integrate them efficiently. One useful tool for multi-omics data integration is to perform canonical correlation analysis (CCA). CCA, due to Hotelling (1936), connects two sets of variables by finding a linear combination of variables that maximally correlate. However, the standard CCA fails when the sample size is strictly less than the number of variables as one can find meaningless solutions with correlations equal to one. Also, it does not perform variable selection and hence lacks interpretability. To circumvent these problems, sparse CCA (sCCA) has been proposed, aiming to find pairs of sparse canonical directions by imposing sparsity penalty. The first sCCA algorithm was presented by Parkhomenko et al. (2007), which, however, lacks exact criterion and biconvexity. Witten et al. (2009) applied the penalized matrix decomposition to cross-product matrix and yielded a straightforward formulation for sCCA. Some closely related methods include Parkhomenko et al., 2009; Lê Cao et al., 2009. Hardoon and Shawe-Taylor (2011) expressed the sCCA model as a primal-dual Rayleigh quotient, which takes the primal representation and kernel representation as the first view and second view, respectively. Chu et al. (2013) reformed CCA into a trace maximization problem and computed the sparse solution by the linearized Bregman method. To exploit the potential structural information among features, various forms of structure-adaptive sCCA have been proposed (Lin et al., 2013; Chen et al., 2012; Mohammadi-Nejad et al., 2017). In particular, Chen et al. (2013) proposed the structure-constrained sCCA (ssCCA) to exploit the phylogenetic structure in microbiome data.

Advances in next-generation sequencing technologies have enabled the direct sequencing of microbial DNA to determine microbiome composition, using either targeted or shotgun approaches (Wensel et al., 2022). The resulting microbiome data is typically in the form of a count table that records the frequencies of detected taxa in specific samples. However, due to the complexities inherent in the sequencing process, the total count for a sample reflects the sequencing effort rather than the actual microbial load at the sampling site. Consequently, microbiome data are inherently compositional, meaning that we only have information about the relative abundances of taxa. This compositionality presents significant challenges in the statistical analysis of microbiome data. A change in the (absolute) abundance of one taxon can lead to apparent changes in the relative abundances of all other taxa, complicating the identification of the actual causal taxa (Yang and Chen, 2022). The compositional nature also renders many standard multivariate statistical models inappropriate or inapplicable (Aitchison, 1982). Many efforts have been made to address the compositionality in different contexts of microbiome data analysis. For example, Friedman and Alm (2012) developed an iterative procedure named SparCC that allows inference of correlations for compositional data by assuming that the number of taxa is large and the true correlation network is sparse. Lin et al. (2014) dealt with the variable selection in regression with compositional covariates. Jiang et al. (2019) addressed zero inflation and detected pairs of associated compositional and non-compositional covariates using a Bayesian zero-inflated negative binomial regression model. However, existing CCA methods including ssCCA could not address the compositional effects, potentially reducing its precision in recovering relevant taxa.

We propose a new sCCA framework for integrating microbiome data with other high-dimensional omics data. The framework specifically addresses the compositional nature of the microbiome data. It also allows integrating prior structure information by imposing a "soft" constraint on the coefficients through varying penalization strength. As a result, the method provides significant improvement when the structure is informative while maintaining robustness against a misspecified structure. The developed tool aims to be an important resource for investigators to understand the interplay between the microbiome and host, decipher the molecular mechanisms underlying microbiome-disease association, and identify potential microbial targets for intervention.

This paper is organized as follows. Section 2 introduces the new sCCA framework for integrating microbiome compositional data with (non-)compositional high-dimensional data. Section 3 extends the new framework to incorporate additional prior structural information. In Section 4, we conduct numerical simulations to demonstrate the effectiveness of our proposed methods. Section 5 applies the proposed methods in a real microbiome study to investigate the association between gut bacteria and its metabolic output. We conclude with a discussion in Section 6.

2 Compositional sCCA

2.1 Formulation

Let us consider two random vectors $\mathbf{X} = (X_1, \dots, X_p)^\top$ and $\mathbf{Y} = (Y_1, \dots, Y_q)^\top$, where $\mathbf{X}$ contains the composition of p taxa and $\mathbf{Y}$ is a q-dimensional vector of non-compositional covariates. The nature of the composition makes $\mathbf{X}$ lie in a $(p-1)$-dimensional positive simplex. To address the compositionality, Aitchison and Bacon-Shone (1984) proposed applying the log-ratio transformation to compositional covariates resulting in $\mathbf{Z}_{/p} = (\log(X_1/X_p), \dots, \log(X_{p-1}/X_p))$, where X_p is chosen as the reference component. CCA for compositional data can be formulated to find canonical coefficients $\mathbf{a} = (a_1, \dots, a_q)^\top$ and $\mathbf{b}_{-p} = (b_1, \dots, b_{p-1})^\top$ so that the correlation between $\mathbf{a}^\top \mathbf{Y}$ and $\mathbf{b}_{-p}^\top \mathbf{Z}_{/p}$ is maximized. Note that $\mathbf{b}_{-p}^\top \mathbf{Z}_{/p} = \sum_{j=1}^{p-1} b_j \log(X_j/X_p) = \sum_{j=1}^{p} b_j \log(X_j)$ with $b_p = -\sum_{j=1}^{p-1} b_j$. To avoid the choice of a reference component, we can write the term $\mathbf{b}_{-p}^\top \mathbf{Z}_{/p}$ in a symmetric form by noticing that $\mathbf{b}_{-p}^\top \mathbf{Z}_{/p} = \mathbf{b}^\top \mathbf{Z}$, where $\mathbf{Z} = (\log(X_1), \dots, \log(X_p))$ and $\mathbf{b} = (b_1, \dots, b_p)$ with

$$\mathbf{b} \in \mathcal{B}_p := \left\{ \mathbf{c} = \left(c_1, \dots, c_p\right)^\top : \sum_{j=1}^{p} c_j = 0 \right\}.$$

Therefore, the compositional CCA aims to find $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ such that

$$\left(\tilde{\mathbf{a}}, \tilde{\mathbf{b}}\right) = \underset{\mathbf{a} \in \mathbb{R}^q, \mathbf{b} \in \mathcal{B}_p}{\operatorname{argmax}} \operatorname{Corr}\left(\mathbf{a}^\top \mathbf{Y}, \mathbf{b}^\top \mathbf{Z}\right) = \underset{\mathbf{a} \in \mathbb{R}^q, \mathbf{b} \in \mathcal{B}_p}{\operatorname{argmax}} \frac{\operatorname{Cov}\left(\mathbf{a}^\top \mathbf{Y}, \mathbf{b}^\top \mathbf{Z}\right)}{\sqrt{\operatorname{Var}\left(\mathbf{a}^\top \mathbf{Y}\right) \operatorname{var}\left(\mathbf{b}^\top \mathbf{Z}\right)}}.$$

When the dimensions p and q are high (as compared to the sample size), regularization is required to encourage sparsity and to obtain a unique solution to the optimization problem. We let $\mathbf{\Sigma}_{\mathbf{YZ}} = \operatorname{Cov}(\mathbf{Y}, \mathbf{Z})$, $\mathbf{\Sigma}_{\mathbf{Y}} = \operatorname{Cov}(\mathbf{Y}, \mathbf{Y})$ and $\mathbf{\Sigma}_{\mathbf{Z}} = \operatorname{Cov}(\mathbf{Z}, \mathbf{Z})$. Define the weighted l_1 norm based on a vector of non-negative weights $\mathbf{w} = (w_1, \dots, w_p)$ for a vector $\mathbf{b}$ as $\|\mathbf{b}\|_{1,\mathbf{w}} = \sum_{j=1}^{p} w_j |b_j|$. The compositional sCCA problem can then be formulated as

$$\begin{aligned} \max_{\mathbf{a} \in \mathbb{R}^q, \mathbf{b} \in \mathbb{R}^p} \mathbf{a}^\top \mathbf{\Sigma}_{\mathbf{YZ}} \mathbf{b} \quad \text{s.t.} \quad & \mathbf{a}^\top \mathbf{\Sigma}_{\mathbf{Y}} \mathbf{a} \le 1,\ \|\mathbf{a}\|_1 \le C_{\mathbf{a}},\ \mathbf{b}^\top \mathbf{\Sigma}_{\mathbf{Z}} \mathbf{b} \le 1, \\ & \|\mathbf{b}\|_{1,\mathbf{w}} \le C_{\mathbf{b}},\ \mathbf{b} \in \mathcal{B}_p. \end{aligned} \tag{1}$$

Here $C_{\mathbf{a}}, C_{\mathbf{b}} > 0$ are some positive tuning parameters that control the global shrinkage level. The weight w_j allows different penalization strengths according to the data or prior structure information. We will elaborate it in Section 2.3.

It has been shown that in high dimensions, treating the covariance matrix as diagonal can yield good results. Following the same strategy adopted by many of the existing high-dimensional CCA algorithms (e.g., Witten et al., 2009), we substitute in the identity matrix for $\mathbf{\Sigma}_\mathbf{Y}$ and $\mathbf{\Sigma}_\mathbf{Z}$ in the CCA formulation (Equation 1). Moreover, we write the (weighted) l_1 constraints on $\mathbf{a}$ and $\mathbf{b}$ in the Lagrangian form. Given a set of n samples $\{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^{n}$, let $\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}$ be the sample cross-covariance between $\mathbf{Y}$ and $\mathbf{Z}$. We formulate the feasible CCA problem as

$$\begin{aligned}\min_{\mathbf{a}\in\mathbb{R}^q,\mathbf{b}\in\mathbb{R}^p} & -\mathbf{a}^\top\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}+\lambda_\mathbf{a}\|\mathbf{a}\|_1\\ & +\lambda_\mathbf{b}\|\mathbf{b}\|_{1,\mathbf{w}} \quad \text{s.t.} \quad \|\mathbf{a}\|_2\le 1,\ \|\mathbf{b}\|_2\le 1,\ \mathbf{b}\in\mathcal{B}_p,\end{aligned}$$

which can be solved by iteratively optimizing the objective function with respect to one parameter while fixing the other parameter. Specifically, we have the following two updating steps.

1. **Update b**: Fix $\mathbf{a}^{(t)}$ and update $\mathbf{b}$ through

$$\mathbf{b}^{(t)} \leftarrow \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p} -\left(\mathbf{a}^{(t)}\right)^\top\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}+\lambda_\mathbf{b}\|\mathbf{b}\|_{1,\mathbf{w}} \text{ s.t. } \|\mathbf{b}\|_2\le 1,\ \mathbf{b}\in\mathcal{B}_p. \quad (2)$$

2. **Update a**: Fix $\mathbf{b}^{(t)}$ and update $\mathbf{a}$ through

$$\mathbf{a}^{(t+1)} \leftarrow \operatorname*{argmin}_{\mathbf{a}\in\mathbb{R}^q} -\mathbf{a}^\top\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)}+\lambda_\mathbf{a}\|\mathbf{a}\|_1 \text{ s.t. } \|\mathbf{a}\|_2\le 1. \quad (3)$$

2.2 Algorithm

In this section, we discuss the updates in Equations 2, 3. Define the operator

$$g(\mathbf{h},\lambda,\mathbf{w}) = \operatorname*{argmin}_{\mathbf{b}\in\mathcal{B}_p} \frac{1}{2}\|\mathbf{h}-\mathbf{b}\|_2^2+\lambda\|\mathbf{b}\|_{1,\mathbf{w}}.$$

By exploring the Karush-Kuhn-Tuchker conditions, we obtain the following result.

Proposition 2.1. *Set* $\breve{\mathbf{b}} = \operatorname{argmin}_{\mathbf{b}\in\mathcal{B}_p} -\left(\mathbf{a}^{(t)}\right)^\top\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}+\lambda_\mathbf{b}\|\mathbf{b}\|_{1,\mathbf{w}}$. *The solution to* (2) *is given by*

$$\mathbf{b}^{(t)} = \begin{cases}\breve{\mathbf{b}}, & \text{if } \|\breve{\mathbf{b}}\|_2\le 1,\\ \dfrac{g\left(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}^\top\mathbf{a}^{(t)},\lambda_\mathbf{b},\mathbf{w}\right)}{\|g\left(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}^\top\mathbf{a}^{(t)},\lambda_\mathbf{b},\mathbf{w}\right)\|_2}, & \text{if } \|\breve{\mathbf{b}}\|_2 > 1.\end{cases}$$

Remark 2.1. We employ the augmented Lagrangian method (ALM) to solve the optimization problem in $g(\mathbf{h},\lambda,\mathbf{w})$. Specifically, the ALM involves the following two steps of iterations

$$\begin{aligned}\mathbf{b}^{(t+1)} &\leftarrow \operatorname*{argmin}_{\mathbf{b}} \frac{1}{2}\|\mathbf{h}-\mathbf{b}\|_2^2+\lambda_\mathbf{b}\|\mathbf{b}\|_{1,\mathbf{w}}+\frac{\mu_1}{2}\left(\mathbf{1}^\top\mathbf{b}+d^{(t)}\right)^2,\\ d^{(t+1)} &\leftarrow d^{(t)}+\mu_2\mathbf{1}^\top\mathbf{b}^{(t+1)},\end{aligned}$$

where $\mu_1,\mu_2>0$ are the step sizes in dual gradient ascent, which are set to be 1 in our numerical studies. The optimization problem in the first step can be solved using coordinate descent in an inner loop by iterating across the following p components

$$b_j^{(t+1,r+1)} = \frac{1}{1+\mu_1}S\left(h_j-\mu_1\left(\sum_{i<j}b_i^{(t+1,r+1)}+\sum_{i>j}b_i^{(t+1,r)}+d^{(t)}\right),\lambda_\mathbf{b}w_j\right),$$

where $S(a,\lambda)=\operatorname{sign}(a)(|a|-\lambda)_+$ denotes the soft-thresholding operator and h_j is the jth component of $\mathbf{h}$.

Using similar arguments, we can show that the solution to (3) is given by

$$\mathbf{a}^{(t+1)} = \begin{cases}\breve{\mathbf{a}}, & \text{if } \|\breve{\mathbf{a}}\|_2\le 1,\\ \dfrac{S\left(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)},\lambda_\mathbf{a}\right)}{\left\|S\left(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)},\lambda_\mathbf{a}\right)\right\|_2}, & \text{if } \|\breve{\mathbf{a}}\|_2 > 1,\end{cases}$$

where $\breve{\mathbf{a}} = \operatorname{argmin}_{\mathbf{a}\in\mathbb{R}^q} -\mathbf{a}^\top\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)}+\lambda_\mathbf{a}\|\mathbf{a}\|_1$ and $S(\mathbf{a},\lambda)=(S(a_1,\lambda),\ldots,S(a_p,\lambda))^\top$. Set $\mathbf{v}=(v_1,\ldots,v_q)^\top=\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)}$. The objective function in the definition of $\breve{\mathbf{a}}$ becomes $\sum_{j=1}^{q}(-a_jv_j+\lambda_\mathbf{a}|a_j|)$. We see that $a_j^{(t+1)}=0$ if $\lambda_\mathbf{a}\ge v_j$, and $|a_j^{(t+1)}|=\infty$ if $\lambda_\mathbf{a}<v_j$. Therefore, we have

$$\mathbf{a}^{(t+1)} = \begin{cases}\mathbf{0}, & \text{if } \lambda_\mathbf{a}\ge\|\mathbf{v}\|_\infty,\\ \dfrac{S\left(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)},\lambda_\mathbf{a}\right)}{\left\|S\left(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)},\lambda_\mathbf{a}\right)\right\|_2}, & \text{if } \lambda_\mathbf{a}<\|\mathbf{v}\|_\infty.\end{cases}$$

As our goal is to find two directions $\mathbf{a}$ and $\mathbf{b}$ to maximize $\operatorname{Cov}(\mathbf{a}^\top\mathbf{Y},\mathbf{b}^\top\mathbf{Z})$, we only consider the updates when the l_2 constraints are binding, which leads to Algorithm 1 below.

```
1. Initialize $\mathbf{a}^{(0)}$ as the first left singular vector with
   unit $l_2$ norm from the singular value decomposition of
   $\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}$.
2. Update b: Fix $\mathbf{a}^{(t)}$ and update $\mathbf{b}$ through
      $\mathbf{b}^{(t)} \leftarrow \frac{g(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}^\top\mathbf{a}^{(t)},\lambda_\mathbf{b},\mathbf{w})}{\|g(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}^\top\mathbf{a}^{(t)},\lambda_\mathbf{b},\mathbf{w})\|_2},$      (4)
   where $g(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}^\top\mathbf{a}^{(t)},\lambda_\mathbf{b},\mathbf{w})$ can be obtained through the
   iterations described in Remark 2.1.
3. Update a: Fix $\mathbf{b}^{(t)}$ and update $\mathbf{a}$ through
      $\mathbf{a}^{(t+1)} \leftarrow \frac{S(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)},\lambda_\mathbf{a})}{\|S(\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)},\lambda_\mathbf{a})\|_2}.$      (5)
4. Iterate Steps 2 and 3 until convergence.
```

Algorithm 1. Compositional sCCA: compositional data *versus* non-compositional data.

2.3 Selecting tuning parameters

To select the regularization parameters $\lambda_\mathbf{a}$ and $\lambda_\mathbf{b}$, we consider a two-stage K-fold cross-validation (CV) method as motivated by Chen et al. (2013). We partition all the samples into M folds, and denote $\mathbf{Y}_k=\mathbf{Y}_{I_k}$ and $\mathbf{Z}_k=\mathbf{Z}_{I_k}$, where I_k are the indexes of samples in the kth fold for $k=1,\ldots,K$. The K-fold cross-validation criterion is

$$\operatorname{CV}(\lambda_\mathbf{a},\lambda_\mathbf{b}) = \frac{1}{K}\sum_{k=1}^{K}\operatorname{Corr}\left(\hat{\mathbf{a}}_{-k}(\lambda_\mathbf{a},\lambda_\mathbf{b})^\top\mathbf{Y}_k,\hat{\mathbf{b}}_{-k}(\lambda_\mathbf{a},\lambda_\mathbf{b})^\top\mathbf{Z}_k\right) \quad (6)$$

where $\hat{\mathbf{a}}_{-k}(\lambda_\mathbf{a},\lambda_\mathbf{b})$ and $\hat{\mathbf{b}}_{-k}(\lambda_\mathbf{a},\lambda_\mathbf{b})$ are the solutions to the compositional sCCA problem based on the samples

$(\cup_{k=1}^{K} I_k)\backslash I_k$ with the tuning parameters $(\lambda_{\mathbf{a}}, \lambda_{\mathbf{b}})$. The parameter selection based on the CV criterion can be influenced by shrinkage problems arising from the sparsity penalty. To avoid this bias, we adopt the coefficients estimated from a two-stage approach when evaluating the CV criterion. In the first stage, we implement Algorithm 1 with the given tuning parameter pair $(\lambda_{\mathbf{a}}, \lambda_{\mathbf{b}})$ and exclude variables with zero coefficients. In the second stage, we recalculate the coefficients by applying Algorithm 1 with a tuning parameter pair (0,0). The optimal tuning parameter pair is chosen as the one that maximizes the CV value with these recalculated coefficients. Our approach accounts for the compositional structure in the second stage by restricting the coefficients of compositional data to $\mathcal{B}_p$. As will be shown below, the K-fold cross-validation performs reasonably well with $K = 5$ in our numerical studies.

2.4 Compositional data *versus* compositional data

We briefly describe an extension to the case, where both **X** and **Y** are compositional. For example, we want to associate the composition of the bacterial taxa with that of the fungi taxa. Let $\mathbf{Y} = (Y_1, \ldots, Y_q)^\top$ be the relative abundances of another set of compositional features. In this case, we let $\mathbf{U} = (\log(Y_1), \ldots, \log(Y_q))^\top$ and define $\hat{\boldsymbol{\Sigma}}_{\mathbf{UZ}}$ as the sample covariance between **U** and **Z**. Following the derivation in Section 2.1, we formulate the two-sided compositional sCCA problem as

$$\min_{\mathbf{a}\in\mathbb{R}^q, \mathbf{b}\in\mathbb{R}^p} -\mathbf{a}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{UZ}}\mathbf{b} + \lambda_{\mathbf{a}}\|\mathbf{a}\|_{1,\mathbf{w}_1} + \lambda_{\mathbf{b}}\|\mathbf{b}\|_{1,\mathbf{w}_2} \quad \text{s.t.} \quad \|\mathbf{a}\|_2 \le 1,\ \|\mathbf{b}\|_2 \le 1,\ \mathbf{a} \in \mathcal{B}_q, \mathbf{b} \in \mathcal{B}_p,$$

where $\mathbf{w}_j = (w_{1j}, \ldots, w_{pj})$ for $j = 1, 2$ are non-negative weights. This problem can be solved by Algorithm 2 below. We use the two-stage CV criterion to select the tuning parameters in the same way as described in Section 2.3.

```
1. Initialize $\mathbf{a}^{(0)}$ as the first left singular vector with unit $l_2$ norm from the singular value decomposition of $\hat{\boldsymbol{\Sigma}}_{\mathbf{UZ}}$.
2. Update b: Fix $\mathbf{a}^{(t)}$ and update b through
       $\mathbf{b}^{(t)} \leftarrow \frac{g(\hat{\boldsymbol{\Sigma}}_{\mathbf{UZ}}^\top \mathbf{a}^{(t)}, \lambda_{\mathbf{b}}, \mathbf{w}_2)}{\|g(\hat{\boldsymbol{\Sigma}}_{\mathbf{UZ}}^\top \mathbf{a}^{(t)}, \lambda_{\mathbf{b}}, \mathbf{w}_2)\|_2}.$   (7)
3. Update a: Fix $\mathbf{b}^{(t)}$ and update a through
       $\mathbf{a}^{(t+1)} \leftarrow \frac{g(\hat{\boldsymbol{\Sigma}}_{\mathbf{UZ}} \mathbf{b}^{(t)}, \lambda_{\mathbf{a}}, \mathbf{w}_1)}{\|g(\hat{\boldsymbol{\Sigma}}_{\mathbf{UZ}} \mathbf{b}^{(t)}, \lambda_{\mathbf{a}}, \mathbf{w}_1)\|_2}.$   (8)
4. Iterate Steps 2 and 3 until convergence.
```

Algorithm 2. Compositional sCCA: compositional data *versus* compositional data.

3 Structure-adaptive compositional sCCA

In this section, following the strategy proposed in Pramanik and Zhang (2020), we aim to incorporate the prior structure information robustly in the compositional sCCA procedure. The prior structure information could be the grouping structure or the phylogenetic tree structure among the taxa. The idea is to define a set of constraints that encode the prior structure information and use the constraints together with the data to estimate the weights **w** in an iterative fashion. It is worth mentioning that our constraint is “soft” as compared to the “hard” constraints used by traditional approaches such as the group Lasso or fused Lasso. As a result, our method provides significant improvement when an external structure is informative while maintaining robustness against a misspecified structure.

3.1 Structure-adaptive weights

Based on the setups described in Section 2.1, our procedure is to translate the auxiliary information into different penalization strengths through the weights **w**. Our framework is general enough to incorporate different types of external structures. For instance, co-expressed genes can be classified into the same group to reflect their biological relationships. In our study, we focus on leveraging the taxonomic grouping structure among taxa. Taxonomically related taxa, such as multiple species within the same genus, tend to share biological traits. Consequently, we anticipate that these taxa will exhibit similar relationships with omics features, leading to the expectation of comparable CCA coefficients. We translate the grouping structure information into different restrictions on the weights. Specifically, we divide the taxa into different groups according to their taxonomy such as phylum, family, and genus. We next consider the set of weights:

$$\mathcal{M}_{\text{Group}} = \left\{\mathbf{w} \in [0, C_U]^p \colon w_i = w_j \text{ if } i, j \in S_d \text{ for } i, j \in \{1, 2, \ldots, p\} \text{ and } d \in \{1, 2, \ldots, D\}\right\},$$

where D represents the number of groups and C_U denotes an upper bound on the weights.

3.2 Structure-adaptive compositional sCCA

Following Pramanik and Zhang (2020), we impose a penalty term on the weights and propose an algorithm to jointly estimate weights and parameters. Specifically, we define

$$h(w_j; \gamma) = \begin{cases} \exp\left\{w_j^{1-\frac{1}{\gamma}}\left(1 - \frac{1}{\gamma}\right)\right\}, & \text{if } 0 < \gamma < 1, \\ w_j, & \text{if } \gamma = 1. \end{cases}$$

We estimate $(\mathbf{a}, \mathbf{b})$ and **w** jointly by solving the following problem

TABLE 1 Performance of sCCA for the association between compositional data and non-compositional data ($\sigma_v = 4$). Numbers in the parentheses represent the corresponding standard deviations.

Setup	$p = q$	Method	$\hat{a}$				$\hat{b}$			
			TPR	FPR	MCC	Precision	TPR	FPR	MCC	Precision
S1	100	sCCA	0.99 (0.03)	0.13 (0.13)	0.68 (0.16)	0.55 (0.19)	0.97 (0.11)	0.19 (0.14)	0.57 (0.15)	0.43 (0.18)
		C-sCCA	0.99 (0.03)	0.12 (0.12)	0.69 (0.16)	0.55 (0.19)	0.99 (0.09)	0.11 (0.06)	0.68 (0.11)	0.54 (0.14)
		AC-sCCA	0.98 (0.06)	0.07 (0.12)	0.79 (0.17)	0.70 (0.21)	0.92 (0.18)	0.05 (0.06)	0.78 (0.16)	0.74 (0.19)
		SAC-sCCA	0.98 (0.05)	0.07 (0.11)	0.78 (0.16)	0.69 (0.21)	0.98 (0.13)	0.01 (0.03)	0.94 (0.13)	0.93 (0.14)
	200	sCCA	0.98 (0.05)	0.06 (0.04)	0.70 (0.12)	0.54 (0.16)	0.74 (0.38)	0.06 (0.05)	0.55 (0.17)	0.61 (0.28)
		C-sCCA	0.98 (0.04)	0.05 (0.04)	0.70 (0.12)	0.54 (0.16)	0.99 (0.04)	0.05 (0.04)	0.72 (0.16)	0.57 (0.22)
		AC-sCCA	0.96 (0.06)	0.03 (0.03)	0.80 (0.12)	0.70 (0.18)	0.93 (0.13)	0.02 (0.02)	0.83 (0.12)	0.78 (0.16)
		SAC-sCCA	0.96 (0.06)	0.03 (0.03)	0.80 (0.11)	0.71 (0.17)	1.00 (0.00)	0.00 (0.00)	0.99 (0.04)	0.97 (0.07)
S2	100	sCCA	0.99 (0.03)	0.13 (0.12)	0.67 (0.16)	0.54 (0.18)	0.99 (0.08)	0.11 (0.09)	0.62 (0.17)	0.46 (0.22)
		C-sCCA	0.99 (0.03)	0.13 (0.12)	0.68 (0.16)	0.54 (0.19)	1.00 (0.00)	0.04 (0.05)	0.81 (0.14)	0.70 (0.21)
		AC-sCCA	0.98 (0.05)	0.08 (0.12)	0.77 (0.17)	0.67 (0.21)	1.00 (0.02)	0.01 (0.03)	0.93 (0.12)	0.89 (0.18)
		SAC-sCCA	0.98 (0.05)	0.08 (0.12)	0.78 (0.16)	0.68 (0.21)	1.00 (0.00)	0.01 (0.01)	0.95 (0.07)	0.91 (0.12)
	200	sCCA	0.98 (0.04)	0.06 (0.04)	0.71 (0.14)	0.55 (0.19)	1.00 (0.00)	0.05 (0.04)	0.65 (0.15)	0.46 (0.19)
		C-sCCA	0.98 (0.04)	0.05 (0.04)	0.71 (0.13)	0.56 (0.18)	1.00 (0.00)	0.02 (0.02)	0.84 (0.13)	0.72 (0.19)
		AC-sCCA	0.97 (0.06)	0.03 (0.03)	0.82 (0.13)	0.72 (0.20)	0.99 (0.08)	0.00 (0.01)	0.95 (0.09)	0.91 (0.13)
		SAC-sCCA	0.97 (0.06)	0.03 (0.03)	0.82 (0.13)	0.73 (0.20)	1.00 (0.00)	0.00 (0.00)	0.96 (0.06)	0.93 (0.10)

$$\min_{\mathbf{a}\in\mathbb{R}^q,\mathbf{b}\in\mathbb{R}^p,\mathbf{w}\in\mathcal{M}} -\mathbf{a}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{YZ}}\mathbf{b} + \lambda_{\mathbf{a}}\|\mathbf{a}\|_1 + \lambda_{\mathbf{b}}\sum_{j=1}^{p}\left\{w_j|b_j| - \log h\left(w_j,\gamma\right)\right\}$$
$$\text{s.t.}\quad \|\mathbf{a}\|_2 \le 1,\ \|\mathbf{b}\|_2 \le 1,\ \mathbf{b}\in\mathcal{B}_p.$$

The design of the function h is to reduce our method to the classic (iterative) adaptive Lasso when there is no external information. The readers are referred to Pramanik and Zhang (2020) for more discussions on the motivation.

Next we introduce the algorithm to solve the above problem. We focus on the update for $\mathbf{w}$ as the updates for $\mathbf{a}$ and $\mathbf{b}$ remain the same as in Section 2.2. In particular, we update $\mathbf{w}$ through

$$\mathbf{w}^{(t+1)} \leftarrow \operatorname*{argmin}_{\mathbf{w}\in\mathcal{M}} \sum_{j=1}^{p}\left\{w_j|b_j^{(t)}| - \log h\left(w_j,\gamma\right)\right\}.$$

When $\mathcal{M} = \mathcal{M}_{\text{Group}}$, it is straightforward to verify that

$$w_j^{(t+1)} = \begin{cases} C_U & \text{if } b_j^{(t)} = 0 \text{ for all } j\in S_d,\\ \left(\dfrac{1}{|S_d|^{-1}\sum_{j\in S_d}|b_j^{(t)}|}\right)^{\gamma} & \text{otherwise.}\end{cases}$$

If we do not have any prior structural information on $\mathbf{b}$ that we can take advantage of, we take $\mathcal{M}$ to be $[0, C_U]^p$. In this case, we have

$$w_j^{(t+1)} = \begin{cases} C_U & \text{if } b_j^{(t)} = 0,\\ |b_j^{(t)}|^{-\gamma} & \text{otherwise.}\end{cases}$$

Algorithm 3 summarizes the implementation details of the structure adaptive Compositional sCCA. The selection of tuning parameter pair $(\lambda_{\mathbf{a}}, \lambda_{\mathbf{b}}, \gamma)$ follows a similar approach as described in Section 2.3, with the difference of using $(\lambda_{\mathbf{a}}, \lambda_{\mathbf{b}}, \gamma)$ in the first stage and $(0, 0, \gamma)$ in the second stage.

```
1. Initialize $\mathbf{a}^{(0)}$ as the first left singular vector with unit $l_2$ norm from the singular value decomposition of $\hat{\boldsymbol{\Sigma}}_{\mathbf{YZ}}$.
2. Update b: Fix $\mathbf{a}^{(t)}$ and update $\mathbf{b}$ through
       $$\mathbf{b}^{(t)} \leftarrow \frac{g\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{YZ}}^\top\mathbf{a}^{(t)}, \lambda_{\mathbf{b}}, \mathbf{w}^{(t)}\right)}{\|g\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{YZ}}^\top\mathbf{a}^{(t)}, \lambda_{\mathbf{b}}, \mathbf{w}^{(t)}\right)\|_2}, \qquad (9)$$
   where $g(\hat{\boldsymbol{\Sigma}}_{\mathbf{YZ}}^\top\mathbf{a}^{(t)}, \lambda_{\mathbf{b}}, \mathbf{w}^{(t)})$ can be obtained through the iterations described in Remark 2.1.
3. Update w: Fix $\mathbf{b}^{(t)}$ and update $\mathbf{w}$ through
       $$\mathbf{w}^{(t+1)} \leftarrow \operatorname*{argmin}_{\mathbf{w}\in\mathcal{M}} \sum_{j=1}^{p}\left\{w_j|b_j^{(t)}| - \log h\left(w_j,\gamma\right)\right\}.$$
4. Update a: Fix $\mathbf{b}^{(t)}$ and update $\mathbf{a}$ through
       $$\mathbf{a}^{(t+1)} \leftarrow \frac{S\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)}, \lambda_{\mathbf{a}}\right)}{\left\|S\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{YZ}}\mathbf{b}^{(t)}, \lambda_{\mathbf{a}}\right)\right\|_2}. \qquad (10)$$
5. Iterate Steps 2-4 until convergence.
```

Algorithm 3. Structure-Adaptive Compositional sCCA: compositional data *versus* non-compositional data.

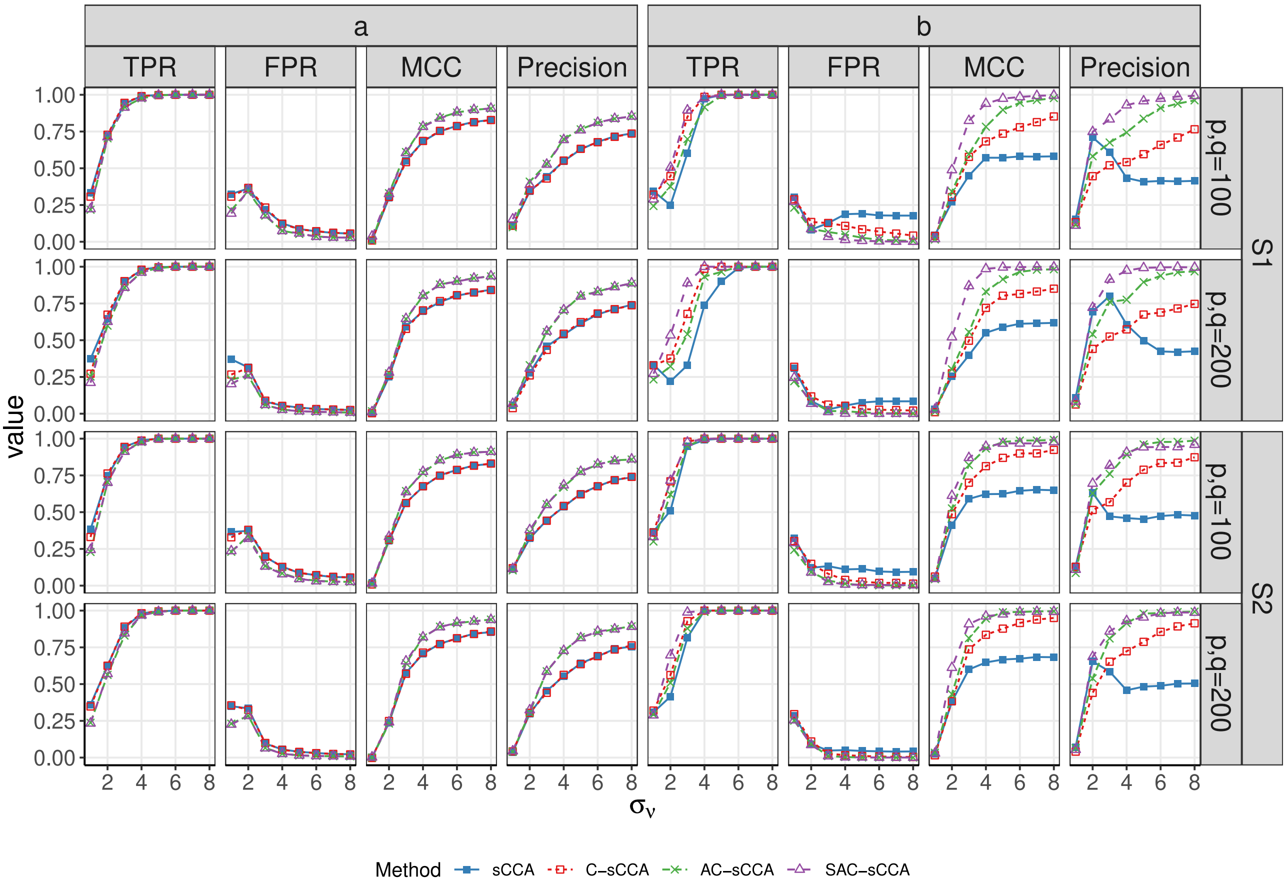

FIGURE 1
TPR, FPR, MCC, and Precision of sCCA for the association between compositional data and non-compositional data across association strength. Here, the range of σ_ν is $\{1, 2, \ldots, 8\}$. Line with solid blue squares: sCCA; line with open red squares: C-sCCA; line with green crosses: AC-sCCA; line with open purple triangles: SAC-sCCA.

4 Simulation studies

In this section, we evaluate the finite sample performance of the proposed methods through numerical simulations.

4.1 Compositional data *versus* non-compositional data

We first consider the CCA problem between compositional data (i.e., microbiome data) and non-compositional data (e.g., metabolomics data) following a similar setting considered in Chen et al. (2013). To capture the dependence between the two sets of high-dimensional data, we use a latent variable model to generate the compositional variables $\{\mathbf{X}_i\}$ (log scale) and non-compositional variables $\{\mathbf{Y}_i\}$ (original scale), where the dependence between these two sets of variables is governed by a latent variable ν. Specifically, we assume that

$$\log(\mathbf{X}_i) = \nu_i \boldsymbol{\omega}_{\mathbf{X}} + \boldsymbol{\varepsilon}_{\mathbf{X},i}, \quad \mathbf{Y}_i = \nu_i \boldsymbol{\omega}_{\mathbf{Y}} + \boldsymbol{\varepsilon}_{\mathbf{Y},i},$$

where $\nu_i \sim N(0, \sigma_\nu^2)$ and $\boldsymbol{\varepsilon}_{X,i}, \boldsymbol{\varepsilon}_{Y,i}$ follow $N(\mathbf{0}_p, \sigma_\varepsilon^2 \mathbf{I}_{p\times p})$ and $N(\mathbf{0}_q, \sigma_\varepsilon^2 \mathbf{I}_{q\times q})$, respectively. The coefficients $\boldsymbol{\omega}_{\mathbf{X}} \in \mathbb{R}^p$ and $\boldsymbol{\omega}_{\mathbf{Y}} \in \mathbb{R}^q$ control the relative contributions of individual variables to the overall association. The ratio $\sigma_\nu/\sigma_\varepsilon$ determines the overall association strength between $\log(\mathbf{X})$ and $\mathbf{Y}$, with a larger value indicating stronger association. For the dimensions, we set $(p, q) = (100, 100), (200, 200)$. We consider two setups for $\boldsymbol{\omega}_{\mathbf{X}}$.

$$\text{S1} \ \boldsymbol{\omega}_{\mathbf{X}} = \tfrac{0.85}{10} \times (1, 1, 1, 1, 1, 1, 1, 1, 1 - 9, \mathbf{0}_{p-10})^\top;$$
$$\text{S2} \ \boldsymbol{\omega}_{\mathbf{X}} = \tfrac{0.85}{6} \times (1, 1, 1, 0, 0, 1, 1, -5, 0, 0, \mathbf{0}_{p-10})^\top;$$

where $\mathbf{1}_p^\top \boldsymbol{\omega}_X = 0$ for both setups. The constraints imply that the association between $\mathbf{X}$ and $\mathbf{Y}$ is mediated through the log ratios for $\mathbf{X}$. We focus on the group structure (i.e., $\mathcal{M} = \mathcal{M}_{\text{group}}$) and assume that the p taxa form 20 groups, with the group size equal to 5 for $p = 100$ and equal to 10 for $p = 200$. For example, in Setup S2 with $p = 100$, the grouping is defined as

$$\boldsymbol{\omega}_{\mathbf{X}} = \frac{0.85}{6} \times \left(\underbrace{1, 1, 1, 0, 0}_{\text{Group 1}}, \underbrace{1, 1, -5, 0, 0}_{\text{Group 2}}, \underbrace{\mathbf{0}_{p-10}}_{\text{Groups 3−20}} \right)^\top$$

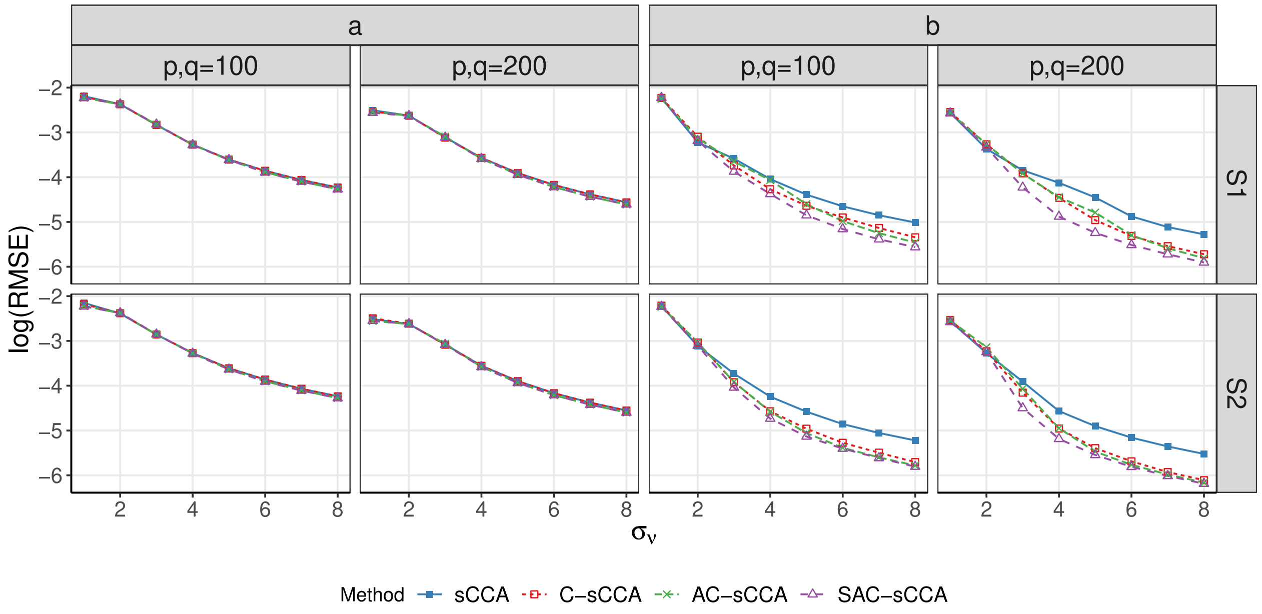

FIGURE 2
RMSE of sCCA for the association between compositional data and non-compositional data across association strength. Here, the range of σ_ν is $\{1, 2, \ldots, 8\}$. Line with solid blue squares: sCCA; line with open red squares: C-sCCA; line with green crosses: AC-sCCA; line with open purple triangles: SAC-sCCA.

TABLE 2 Performance of sCCA for the association between compositional data and non-compositional data ($\sigma_\nu = 4$, $p = 100$, $q = 200$). Numbers in the parentheses represent the corresponding standard deviations.

Setup	Method	$\hat{a}$				$\hat{b}$			
		TPR	FPR	MCC	Precision	TPR	FPR	MCC	Precision
S1	sCCA	0.98 (0.04)	0.05 (0.05)	0.74 (0.15)	0.61 (0.22)	0.92 (0.21)	0.16 (0.11)	0.58 (0.15)	0.48 (0.22)
	C-sCCA	0.99 (0.04)	0.05 (0.05)	0.75 (0.15)	0.61 (0.22)	0.98 (0.09)	0.10 (0.05)	0.70 (0.10)	0.56 (0.13)
	AC-sCCA	0.97 (0.06)	0.02 (0.03)	0.85 (0.12)	0.77 (0.19)	0.93 (0.16)	0.04 (0.04)	0.80 (0.15)	0.75 (0.18)
	SAC-sCCA	0.97 (0.06)	0.02 (0.03)	0.85 (0.12)	0.78 (0.19)	0.98 (0.13)	0.01 (0.02)	0.95 (0.12)	0.94 (0.12)
S2	sCCA	0.98 (0.04)	0.05 (0.04)	0.73 (0.13)	0.58 (0.18)	0.99 (0.08)	0.11 (0.08)	0.63 (0.17)	0.47 (0.22)
	C-sCCA	0.98 (0.04)	0.05 (0.04)	0.73 (0.13)	0.58 (0.18)	1.00 (0.00)	0.04 (0.04)	0.82 (0.15)	0.71 (0.23)
	AC-sCCA	0.97 (0.06)	0.02 (0.02)	0.84 (0.11)	0.75 (0.18)	1.00 (0.00)	0.01 (0.02)	0.95 (0.09)	0.92 (0.14)
	SAC-sCCA	0.97 (0.06)	0.02 (0.02)	0.84 (0.12)	0.75 (0.19)	1.00 (0.00)	0.01 (0.01)	0.96 (0.07)	0.93 (0.11)

The first two groups contain both zero and nonzero entries reflecting the fact that the external structure information is imperfect and noisy. We set $\boldsymbol{\omega}_{\mathrm{Y}} = 0.85 \times (0.08, 0.084, 0.089, \ldots, 0.12, \mathbf{0}_{q-10}^{\top})^{\top}$. Next, we fix $\sigma_\varepsilon = 1$ and vary σ_ν within $\{1, 2, \ldots, 8\}$ to control the strength of the canonical correlation. We report the true positive rate (TPR), false positive rate (FPR), Matthew's correlation coefficient (MCC), and Precision to measure the performance of different methods. Here,

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}},$$

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

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{TP}},$$

where TP, FP, TN, and FN represent the true positives, false positives, true negatives, and false negatives, respectively. The TPR, FPR, MCC, and Precision are computed by averaging over 100 simulation replicates. Denote the estimated canonical coefficients by $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. Their estimation targets are $\boldsymbol{\omega}_{\mathrm{X}}/\|\boldsymbol{\omega}_{\mathrm{X}}\|_2$ and $\boldsymbol{\omega}_{\mathrm{Y}}/\|\boldsymbol{\omega}_{\mathrm{Y}}\|_2$, respectively, where this normalization is to ensure comparability. The estimation accuracy is evaluated using the root mean square error (RMSE).

TABLE 3 Performance of sCCA for the association between two compositional datasets ($\sigma_\nu = 4$). Numbers in the parentheses represent the corresponding standard deviations.

Setup	$p = q$	Method	$\hat{\mathbf{a}}$				$\hat{\mathbf{b}}$			
			TPR	FPR	MCC	Precision	TPR	FPR	MCC	Precision
S3	100	sCCA	1.00 (0.02)	0.19 (0.12)	0.57 (0.13)	0.41 (0.13)	0.99 (0.09)	0.18 (0.09)	0.57 (0.11)	0.41 (0.13)
		C-sCCA	0.99 (0.02)	0.10 (0.06)	0.71 (0.11)	0.56 (0.15)	1.00 (0.01)	0.11 (0.08)	0.70 (0.13)	0.56 (0.17)
		AC-sCCA	0.99 (0.03)	0.03 (0.03)	0.87 (0.10)	0.80 (0.15)	1.00 (0.01)	0.03 (0.03)	0.88 (0.10)	0.80 (0.15)
		SAC-sCCA	1.00 (0.01)	0.01 (0.02)	0.96 (0.07)	0.94 (0.11)	1.00 (0.00)	0.01 (0.03)	0.97 (0.08)	0.96 (0.12)
	200	sCCA	0.97 (0.14)	0.09 (0.05)	0.59 (0.11)	0.41 (0.17)	0.92 (0.25)	0.08 (0.06)	0.58 (0.13)	0.46 (0.21)
		C-sCCA	0.99 (0.03)	0.05 (0.02)	0.69 (0.07)	0.52 (0.09)	1.00 (0.00)	0.06 (0.03)	0.69 (0.10)	0.52 (0.13)
		AC-sCCA	0.97 (0.05)	0.02 (0.02)	0.86 (0.10)	0.79 (0.15)	0.99 (0.03)	0.02 (0.01)	0.88 (0.09)	0.80 (0.15)
		SAC-sCCA	1.00 (0.01)	0.00 (0.00)	0.98 (0.04)	0.97 (0.07)	1.00 (0.00)	0.00 (0.01)	0.98 (0.06)	0.97 (0.09)
S4	100	sCCA	1.00 (0.00)	0.10 (0.07)	0.63 (0.14)	0.45 (0.17)	1.00 (0.00)	0.10 (0.07)	0.64 (0.14)	0.46 (0.17)
		C-sCCA	1.00 (0.00)	0.03 (0.03)	0.86 (0.12)	0.77 (0.19)	1.00 (0.00)	0.03 (0.06)	0.85 (0.16)	0.77 (0.23)
		AC-sCCA	1.00 (0.00)	0.00 (0.01)	0.97 (0.06)	0.94 (0.10)	1.00 (0.00)	0.01 (0.01)	0.96 (0.08)	0.94 (0.12)
		SAC-sCCA	1.00 (0.00)	0.01 (0.01)	0.93 (0.08)	0.89 (0.13)	1.00 (0.00)	0.01 (0.01)	0.95 (0.07)	0.92 (0.11)
	200	sCCA	1.00 (0.00)	0.05 (0.04)	0.66 (0.14)	0.47 (0.19)	1.00 (0.00)	0.05 (0.04)	0.66 (0.16)	0.48 (0.21)
		C-sCCA	1.00 (0.00)	0.02 (0.03)	0.86 (0.15)	0.76 (0.22)	1.00 (0.00)	0.01 (0.01)	0.87 (0.12)	0.78 (0.20)
		AC-sCCA	1.00 (0.00)	0.00 (0.00)	0.98 (0.05)	0.96 (0.09)	1.00 (0.00)	0.00 (0.00)	0.97 (0.05)	0.95 (0.09)
		SAC-sCCA	1.00 (0.00)	0.00 (0.01)	0.96 (0.07)	0.94 (0.11)	1.00 (0.00)	0.00 (0.00)	0.97 (0.05)	0.95 (0.10)

We compare the performance of the following four methods.

1. sCCA: sCCA without considering the compositional effect;
2. C-sCCA: compositional sCCA;
3. AC-sCCA: adaptive compositional sCCA, i.e., $\mathcal{M} = [0, C_U]^p$.
4. SAC-sCCA: structure adaptive compositional sCCA, i.e., $\mathcal{M} = \mathcal{M}_{\text{Group}}$.

For AC-sCCA and SAC-sCCA, we also apply adaptive weights on **a** with $\mathcal{M} = [0, C_U]^p$ in implementation. The value of C_U is set to 10^5.

Table 1 summarizes the results for the above four methods when fixing $\sigma_\nu = 4$. For both Setups S1 and S2, C-sCCA outperforms sCCA in terms of all four measures, especially in reducing the false positive rates and increasing the precision in identifying relevant compositional components, demonstrating the advantage of taking into account the compositional constraint. Compared to the first two methods, AC-sCCA and SAC-sCCA further reduce the FPR and thus lead to higher MCC in estimating **a** and **b**. For identifying **b** in Setup S1, SAC-sCCA outperforms the other three methods by exhibiting higher TPR, nearly zero FPR, and thus higher MCC because of incorporating grouping information. Figure 1 is in general consistent with these findings. As association strength increases, the TPR, MCC, and Precision of C-sCCA, AC-sCCA, and SAC-sCCA increase, whereas their FPRs show a declining trend. When $\sigma_\nu \geq 3$, the precision and FPR in estimating **b** of sCCA becomes worse as association strength increases, which means sCCA identifies more true variables at the cost of including more false variables. Figure 2 presents the RMSE in estimating the canonical coefficients, which decreases as the association strength σ_ν increases. By accounting for the compositional effect, the C-sCCA, AC-sCCA, and SAC-sCCA outperform sCCA, with SAC-sCCA providing the most accurate estimation. This demonstrates that methods accounting for the compositional nature yield more accurate estimations than those that do not, and considering structural information can further enhance performance.

Finally, we examine the scenario where $p = 100$ and $q = 200$ to assess the performance of our method on unbalanced datasets. As presented in Table 2, the four methods exhibit similar performance to that observed in the case where $p = q$. The results indicate that our method successfully handles unbalanced dimensions.

4.2 Compositional data *versus* compositional data

In this section, we study the performance of the proposed compositional sCCA for the association between two compositional datasets, for example, bacterial taxa abundance vs fungi taxa abundance. We modify the setting in Section 4.1 by considering the following models

$$\log(\mathbf{X}_i) = \nu_i \boldsymbol{\omega}_{\mathbf{X}} + \boldsymbol{\varepsilon}_{\mathbf{X},i}, \quad \log(\mathbf{Y}_i) = \nu_i \boldsymbol{\omega}_{\mathbf{Y}} + \boldsymbol{\varepsilon}_{\mathbf{Y},i},$$

where $\nu_i \sim N(0, \sigma_\nu^2)$ and $\boldsymbol{\varepsilon}_{X,i}, \boldsymbol{\varepsilon}_{Y,i}$ follow $N(\mathbf{0}_p, \sigma_\varepsilon^2 \mathbf{I}_{p\times p})$ and $N(\mathbf{0}_q, \sigma_\varepsilon^2 \mathbf{I}_{q\times q})$, respectively. We again consider two setups for $\boldsymbol{\omega}_{\mathbf{X}}$ and $\boldsymbol{\omega}_{\mathbf{Y}}$.

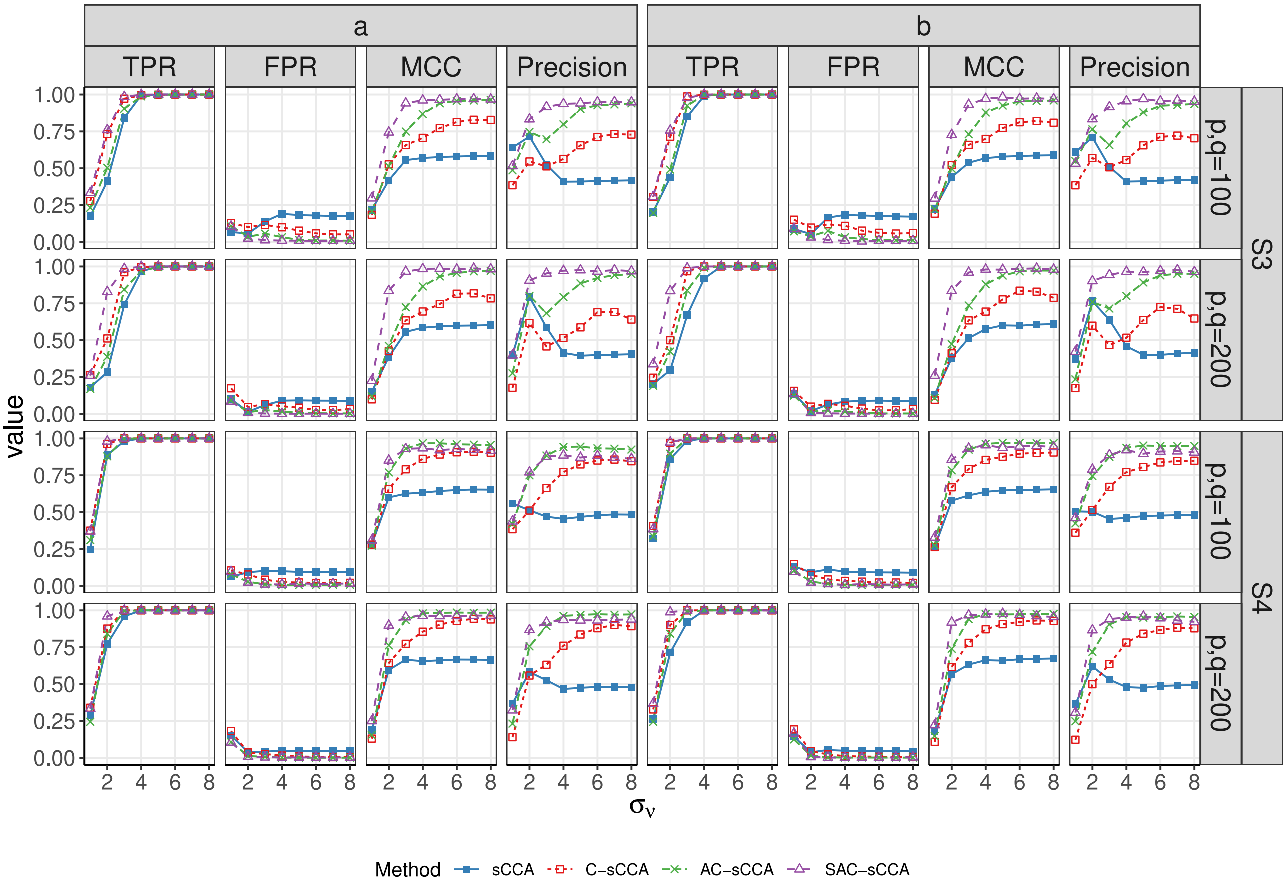

FIGURE 3
TPR, FPR, MCC, and Precision of sCCA for the association between two compositional datasets across signal strength. Here, the range of σ_v is $\{1, 2, \ldots, 8\}$. Line with solid blue squares: sCCA; line with open red squares: C-sCCA; line with green crosses: AC-sCCA; line with open purple triangles: SAC-sCCA.

S3 $\boldsymbol{\omega}_X = \frac{0.85}{10} \times (\mathbf{1}_9^\top, -9, \mathbf{0}_{p-10})^\top$ and

$$\boldsymbol{\omega}_Y = 0.85 \times (0.08, 0.085, 0.09, \ldots, 0.12, -0.9, \mathbf{0}_{q-10}^\top)^\top;$$

S4 $\boldsymbol{\omega}_X = \frac{0.85}{6} \times (1, 1, 1, 0, 0, 1, 1, -5, 0, 0, \mathbf{0}_{p-10})^\top$ and

$$\boldsymbol{\omega}_Y = \frac{0.85}{6} \times (1, 1, 1, 0, 0, 1, 1, -5, 0, 0, \mathbf{0}_{q-10})^\top;$$

where $\mathbf{1}_p^\top \boldsymbol{\omega}_X = 0$ and $\mathbf{1}_q^\top \boldsymbol{\omega}_Y = 0$ for both setups. We group $\boldsymbol{\omega}_\mathbf{X}$ and $\boldsymbol{\omega}_\mathbf{Y}$ using the same strategy applied to $\boldsymbol{\omega}_\mathbf{X}$ in Section 4.1. The other setups are the same as those in Section 4.1. Table 3 summarizes the empirical results. For Setups S3 and S4, C-sCCA often results in higher TPR, higher Precision, lower FPR and therefore higher MCC in estimating both **a** and **b** compared to sCCA, emphasizing the necessity of including compositional constraints again. For Setup S3, SAC-sCCA significantly outperforms the other methods in terms of higher TPR and MCC. Our structure-adaptive approach is robust against a misspecified structure (Setup S4) as the performance of SAC-sCCA is comparable to that of AC-sCCA. Figures 3, 4 show patterns similar to Figures 1, 2.

5 Real application

We applied sCCA, C-sCCA, AC-sCCA and SAC-sCCA to examine the association between gut bacterial composition and gut metabolism in a colorectal adenoma study conducted at the Mayo Clinic. The study utilized both gut microbiome and gut metabolomics data from 241 fecal samples selected from a frozen stool archive. The fecal samples were collected following a standard protocol and metabolomics profiling was conducted by Metabolon, Inc. (Durham, NC, United States) using a UPLCMS/MS platform, as detailed in Kim et al. (2020). Metabolic sub-pathway abundances were calculated by averaging the scaled abundances of metabolites within each sub-pathway, which are grouped into super-pathways. Bacterial DNA extraction and 16S rRNA gene sequencing were described in Hale et al. (2017). Specifically, the sequencing library was prepared at the University of Minnesota Genomics Center, and sequencing was performed using the Illumina MiSeq system at the Mayo Clinic Medical Genome Facility. These sequences were processed through the IM-TORNADO bioinformatics pipeline, clustering them into OTUs based on a 97% identity threshold.

We focused the analysis on the overall association between the bacterial genera and metabolic sub-pathways. We followed Chen

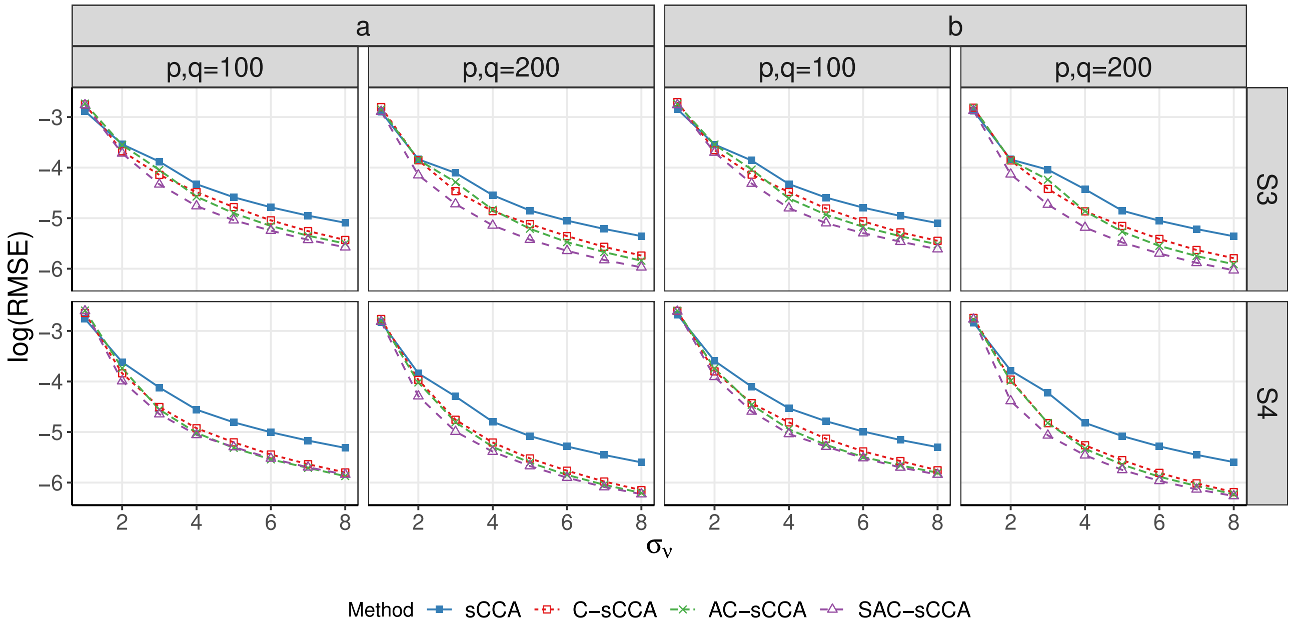

FIGURE 4
RMSE of sCCA for the association between two compositional datasets across association strength. Here, the range of σ_v is $\{1, 2, \ldots, 8\}$. Line with solid blue squares: sCCA; line with open red squares: C-sCCA; line with green crosses: AC-sCCA; line with open purple triangles: SAC-sCCA.

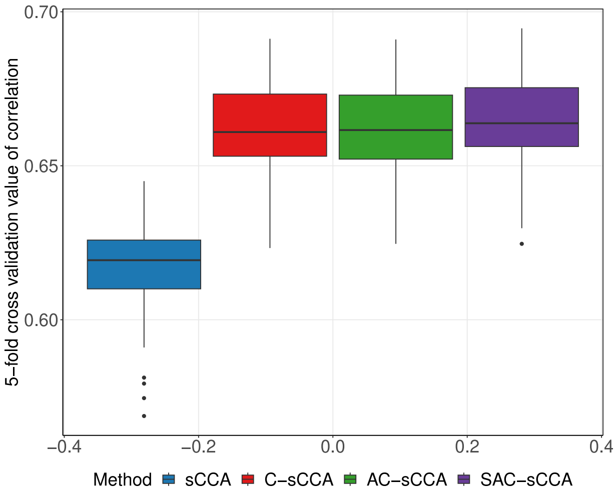

FIGURE 5
Boxplot of 5-fold cross-validated correlations for sCCA, C-sCCA, AC-sCCA, and SAC-sCCA across 100 replications, with tuning parameters determined by sample partition in each replication.

et al. (2013) to pre-process data. We excluded genera that were present in less than 1/4 samples and kept 63 relatively common genera, each belonging to a specific phylum. This approach ensures a balance between retaining sufficient taxa for meaningful analysis while filtering out rare genera that could introduce noise. Zeros were replaced with 0.5 in microbiome data to facilitate working on the log scale. Our final dataset is summarized as a metabolic sub-pathway abundance matrix $\mathbf{Y}_{241\times91}$ and a bacterial genus abundance matrix $\mathbf{X}_{241\times63}$. The group information, specified as super-pathway and phylum, respectively, is incorporated into our analysis. We applied logarithmic transformation to both matrices: for metabolic data, to normalize the distribution, and for genus abundance, to account for the compositional structure. Finally, we performed standardization to ensure that all variables have zero mean and unit variance.

We performed a two-stage five-fold CV described in Section 2.3 to identify the optimal tuning parameters across a range of models, from the most dense to the most sparse. To mitigate randomness, we conducted 100 replications of sample partitions. We selected the tuning parameter pair for each replication and recorded the corresponding CV values of four methods. As shown in Figure 5, sCCA has the lowest CV correlations, followed by C-sCCA and AC-sCCA, both of which yield comparable CV correlations, while SAC-sCCA achieves slightly higher CV correlations by incorporating grouping information. Therefore, by accounting for the compositional structure, we achieved a stronger association between the two datasets. The final parameters were determined by maximizing the CV values averaged across the 100 replications, with CV values of 0.6076 for sCCA, 0.6578 for C-sCCA, 0.6581 for AC-sCCA, and 0.6584 for SAC-sCCA.

Figure 6 shows the heatmap of pairwise spearman correlations between metabolic sub-pathways and genera selected by any of the four methods. The signs of the estimated coefficients align with the pairwise correlations. The selected metabolic sub-pathways belong to four super-pathways: *Carbohydrate*, *Lipid*, *Cofactors and Vitamins*, and *Amino Acid*. Hierarchical clustering analysis, using the complete linkage and Euclidean distance, was applied to cluster the bacterial genera. The coefficients estimated by C-sCCA, AC-sCCA, and SAC-sCCA for bacteria within the third cluster were mostly positive while the other two clusters showed an opposite trend. Interestingly, *Fatty Acid, Diacarboxylate (FA-DC)*, identified

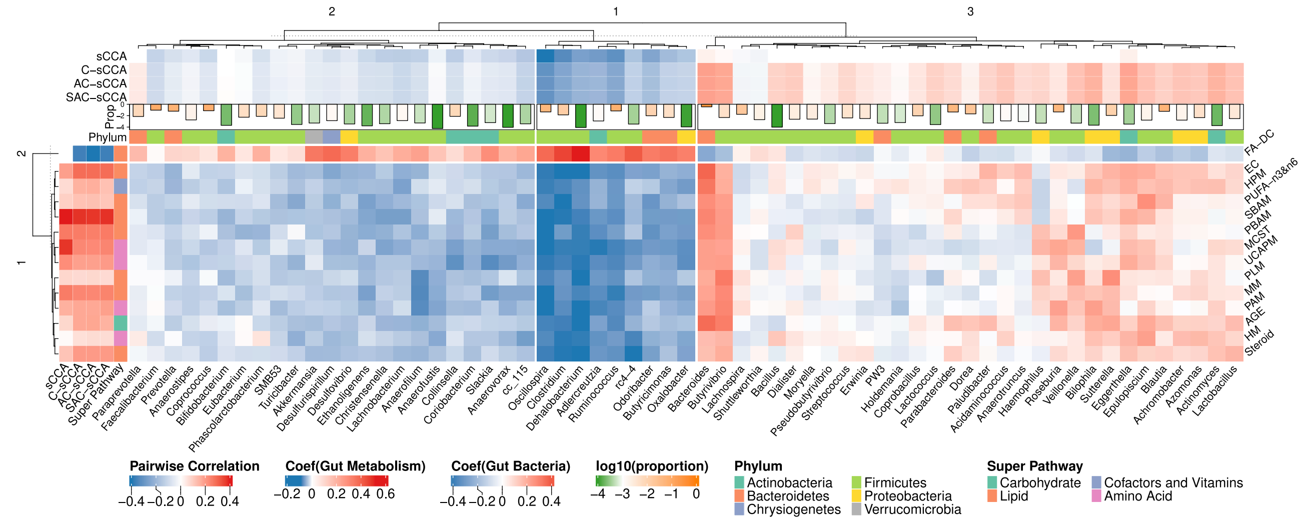

FIGURE 6
Heatmap of Spearman correlations between the bacterial genera and metabolic sub-pathways selected by either sCCA, C-sCCA, AC-sCCA, or SAC-sCCA. The color indicates the association direction, with red for positive correlations and blue for negative, varying in shade by strength. The bars at the top represent the average relative abundances of these genera on a log 10 scale, with orange indicating higher values and green indicating lower values. Abbreviations: AGE (Advanced Glycation End-product), EC (Endocannabinoid), FA-DC (Fatty Acid, Dicarboxylate), HPM (Hemoglobin and Porphyrin Metabolism), HM (Histidine Metabolism), MCST (Methionine, Cysteine, SAM and Taurine Metabolism), MM (Mevalonate Metabolism), PLM (Phospholipid Metabolism), PAM (Polyamine Metabolism), PUFA-n3&n6 (Polyunsaturated Fatty Acid, n3 and n6), PBAM (Primary Bile Acid Metabolism), SBAM (Secondary Bile Acid Metabolism), Steroid (Steroid), UCAPM (Urea cycle; Arginine and Proline Metabolism).

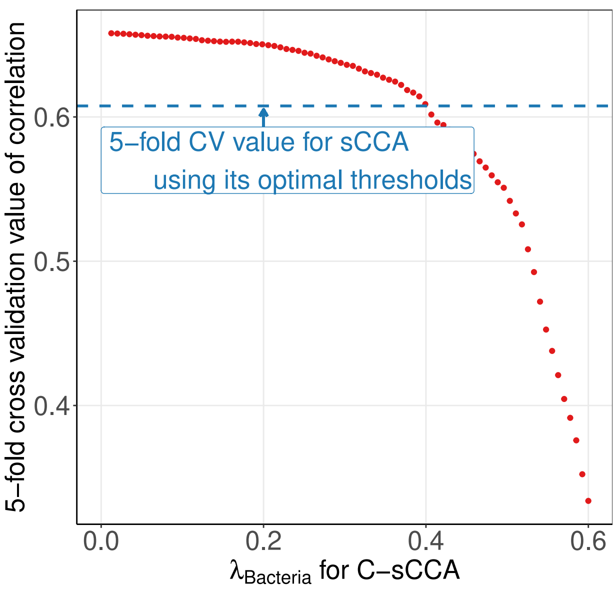

FIGURE 7
Averaged cross-validated correlations across 100 replications. The blue horizontal line represents the CV value of sCCA with its optimal threshold. The red points denotes the CV values of C-sCCA with varying tuning parameters for gut bacterial genera.

by C-sCCA, AC-sCCA, and SAC-sCCA but not by sCCA, was overall negatively correlated with bacterial genera in the third cluster, and positively correlated with those in the first/second clusters. Dicarboxylic acids can be produced by various bacteria through different metabolic pathways (Yu et al., 2018). For example, species in the genus *Clostridium*, which showed a strong correlation with FA-DC in our data, can produce succinic acid and other dicarboxylic acids as fermentation products (Koendjbiharie et al., 2018). As a comparison, other detected metabolic sub-pathways exhibited both negative and positive correlations with bacterial genera in the third cluster, and overall negative correlations with those in the first/second clusters.

Despite achieving higher cross-validated correlations, the three methods that accounted for compositional structure failed to induce a sparse structure for the bacterial genera. Although this may be the biological truth, as gut metabolic capabilities are contributed by a large number of bacteria collectively (Cox et al., 2022), to gain more insights into the benefits of using the compositional constraint, we reconsidered C-sCCA by fixing the final parameter for metabolic sub-pathways and varying the parameter for bacterial genera to achieve the same cross-validated correlation as sCCA. Figure 7 shows the average cross-validated correlation of C-sCCA for each parameter pair. We selected the tuning parameter of bacterial genera as 0.4 so that the averaged CV value of C-sCCA is almost the same as that of sCCA.

Figure 8 presents the heatmap of pairwise Spearman correlations between the metabolic sub-pathways and genera selected by any of sCCA and C-sCCA with newly determined tuning parameter pair. C-sCCA identifies 8 metabolism sub-pathways and 17 bacterial genera, while sCCA selects a broader set of 13 metabolic sub-pathways and 35 bacterial genera. By incorporating the compositional structure, C-sCCA achieves the comparable averaged CV value with a more focused selection of metabolic sub-pathways and bacterial genera.

Among the eight metabolic sub-pathways identified by C-sCCA, five belong to the *Lipid* super-pathway, two are part of the *Amino Acid* super-pathway, and only one, the *Advanced Glycation End-product*, is

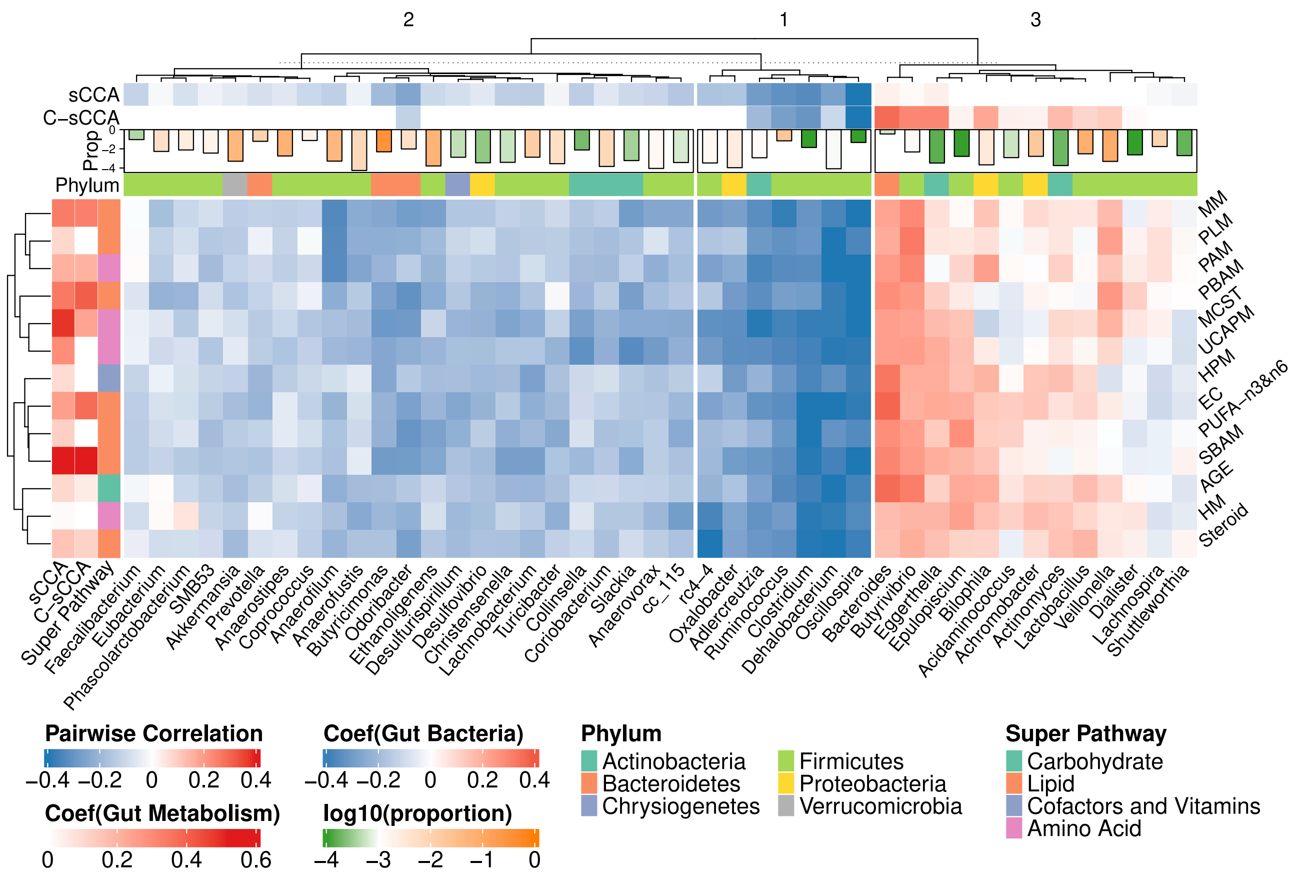

FIGURE 8
Heatmap of Spearman correlations between the bacterial genera and metabolic sub-pathways selected by either sCCA or C-sCCA. The CV values averaged across 100 replications are approximately 0.61 for both methods with the chosen tuning parameters. Other details are the same as in Figure 6.

associated with the *Carbohydrate* super-pathway. For the bacterial genera, a comparison between C-sCCA results and the association analysis of bacterial genera and metabolic sub-pathways by Kim et al. (2020) reveals interesting patterns. The bacterial genera identified by C-sCCA are grouped into three clusters. The first two clusters predominantly show a negative association with the selected sub-pathways. In the first cluster, *Oscillospira*, *Clostridium*, *Ruminococcus*, *Adlercreutzia*, and *Dehalobacterium* are consistently selected by C-sCCA and were also reported in the findings of Kim et al. (2020). In the second cluster, *Odoribacter* is detected by C-sCCA but was not identified by Kim et al. (2020). In the third cluster, *Bacteroides*, *Eggerthella*, and *Butyrivibrio* stand out with the largest C-sCCA coefficients, showing strong positive correlations with the selected metabolic sub-pathways. In Kim et al. (2020), the genera were classified into two clusters, with the second cluster consisting of *Bacteroides*, *Epulopiscium*, and *Butyrivibrio*. Our hierarchical tree further reveals that *Epulopiscium* and *Eggerthella* exhibit very similar patterns, as they are grouped together.

6 Discussion

In this study, we developed a compositional sparse canonical correlation analysis (C-sCCA) framework for association analysis between microbiome data and other high-dimensional datasets, accounting for the compositional nature of microbiome sequencing data. We introduced two variants of the C-sCCA method: one for compositional vs non-compositional data, and another for compositional vs compositional data. Our results show that by incorporating the compositional constraint, we achieved improved selection of relevant taxa, enhancing both power and precision. Additionally, we extended our framework to incorporate prior structural information, such as the grouping of bacterial taxa, among the compositional components. Application of C-sCCA to real microbiome data demonstrated that it produced results that were biologically more interpretable.

There are several potential extensions to our work. While we primarily focused on the grouping structure of bacterial taxa, we could also exploit the hierarchical grouping structure (phylum to genus) and the phylognetic relationship by devising appropriate constraints on the weights $\mathbf{w}$. For the hierarchical structure, we can derive a set of covariates each representing a hierarchical level, which can then be used to impose specific structures on the weights. Suppose for the jth taxa, we have a corresponding covariate $\boldsymbol{\xi}_j$. To incorporate such covariate information, we define the set of weights as

$$\mathcal{M}_{\text{Covariate}} = \left\{\mathbf{w} \in [0, C_U]^p \colon w_j = f\left(\boldsymbol{\xi}_j; \boldsymbol{\theta}\right) \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta}, j \in \{1, 2, \ldots, p\}\right\},$$

where $f(\cdot;\boldsymbol{\theta})$ is a prespecified class of functions parameterized by $\boldsymbol{\theta}$. We can also use the phylogenetic tree information by imposing a smoothness constraint, which depends on the pairwise patristic distances among the taxa. Suppose d_{ij} is the patristic distance between the taxa i and j, we can consider the set

$$\mathcal{M}_{\text{Smooth}} = \left\{ \mathbf{w} \in [0, C_U]^p : \sum_{1\le i<j\le p} \kappa\left(d_{ij}\right)|w_i - w_j| \le \epsilon \right\},$$

for some decreasing function $\kappa(\cdot)$ of the distance and tuning parameter $\epsilon > 0$.

We can also extend our approach to learn sub-spaces from multiple views, i.e., when we have multiple groups of measurements, $\mathbf{X}_{(i)} \in \mathbb{R}^{p_i}$, for $i = 1, \ldots, m$ on matching samples. This situation naturally arises in multi-omics studies, where methylomic, transcriptomic, metabolomic, and microbiome data are collected from a single group of individuals. A number of approaches for generalizing CCA to multiple views have been proposed in the literature, and some of these extensions are summarized in Kettenring (1971). We could adapt these multi-view methods to incorporate the compositional constraint.

There are several limitations to our framework. First, we did not simultaneously address the zero inflation commonly observed in microbiome data. We used a simple zero replacement strategy before running the C-sCCA. Although this strategy has been commonly used in microbiome data analysis at log scale (Zhou et al., 2022; Lin and Peddada, 2020), better methods can be developed such as replacing the log scale transformation by modified centered log-ratio transform designed for addressing zero inflation (Yoon et al., 2019), imposing another multinomial layer to account for sampling variability associated with the sequencing process (Chen and Li, 2013), or using more informative imputation methods such as mbDenoise (Zeng et al., 2022) and mbImpute (Jiang et al., 2021). Second, although our framework can select subsets of features that explain the largest correlation between the datasets, their detailed relationships can not be learned simultaneously. Developing methods that combine feature-level selection with the construction of feature-feature correlation networks is a promising area for future research. Third, we can enhance robustness to outliers through several strategies, including outlier detection and removal during data preprocessing, replacing empirical covariance estimators with robust estimators (Luo et al., 2024), and investigating the optimal choice of penalty functions (Chalise and Fridley, 2012), such as Huber loss, Tukey loss, or L_0 penalty (Lindenbaum et al., 2022). Fourth, while our work focuses on linear associations, future extensions of our composite sCCA framework could capture nonlinear relationships by integrating with kernel CCA (Akaho, 2001; Fukumizu et al., 2007), deep CCA (Andrew et al., 2013), or nonparametric CCA (Lancaster, 1958; Michaeli et al., 2016). Lastly, our framework assumes that the association is mediated through the ratios of the compositional components since only ratios are meaningful for compositional data. However, when the association is at the level of absolute abundance - where the total microbial load also matters - our method may not work well. This limitation is more inherent to the constraints of current sequencing technologies than to the method itself.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Genotypes and Phenotypes (https://www.ncbi.nlm.nih.gov/gap) with the study accession number phs001204.v1. p1.

Ethics statement

The studies involving humans were approved by Mayo Clinic's Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants and apos; legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

LD: Data curation, Formal Analysis, Writing–original draft. YT: Formal Analysis, Methodology, Writing–original draft. XZ: Conceptualization, Supervision, Writing–review and editing. JC: Conceptualization, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The authors gratefully acknowledge National Institute of Health R01GM144351 (JC & XZ), National Science Foundation DMS1830392, DMS2113359, DMS1811747 (XZ) and National Science Foundation DMS2113360 and Mayo Clinic Center for Individualized Medicine (JC).

Conflict of interest

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B Methodol.* 44, 139–160. doi:10.1111/j.2517-6161.1982.tb01195.x

Aitchison, J., and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* 71, 323–330. doi:10.1093/biomet/71.2.323

Akaho, S. (2001). "A kernel method for canonical correlation analysis," in *International meeting of psychometric society*, 1.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). "Deep canonical correlation analysis," in *International conference on machine learning* (Atlanta, GA: PMLR), 1247–1255.

Chalise, P., and Fridley, B. L. (2012). Comparison of penalty functions for sparse canonical correlation analysis. *Comput. Statistics and Data Analysis* 56, 245–254. doi:10.1016/j.csda.2011.07.012

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. doi:10.1093/biostatistics/kxs038

Chen, J., and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Statistics* 7. doi:10.1214/12-AOAS592

Chen, X., Han, L., and Carbonell, J. (2012). Structured sparse canonical correlation analysis. *Proc. Fifteenth Int. Conf. Artif. Intell. Statistics* 22, 199–207.

Chu, D., Liao, L.-Z., Ng, M. K., and Zhang, X. (2013). Sparse canonical correlation analysis: new formulation and algorithm. *IEEE Trans. Pattern Analysis Mach. Intell.* 35, 3050–3065. doi:10.1109/TPAMI.2013.104

Cox, T. O., Lundgren, P., Nath, K., and Thaiss, C. A. (2022). Metabolic control by the microbiome. *Genome Med.* 14, 80. doi:10.1186/s13073-022-01092-0

Friedman, J., and Alm, E. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8, e1002687. doi:10.1371/journal.pcbi.1002687

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *J. Mach. Learn. Res.* 8.

Hale, V. L., Chen, J., Johnson, S., Harrington, S. C., Yab, T. C., Smyrk, T. C., et al. (2017). Shifts in the fecal microbiota associated with adenomatous polyps. *Cancer Epidemiol. Prev. Biomarkers* 26, 85–94. doi:10.1158/1055-9965.EPI-16-0337EPI-16-0337

Hardoon, D. R., and Shawe-Taylor, J. (2011). Sparse canonical correlation analysis. *Mach. Learn.* 83, 331–353. doi:10.1007/s10994-010-5222-7

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 83. doi:10.1186/s13059-017-1215-1s13059-017-1215-1

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28, 321–377. doi:10.1093/biomet/28.3-4.321

Jiang, R., Li, W. V., and Li, J. J. (2021). mbImpute: an accurate and robust imputation method for microbiome data. *Genome Biol.* 22, 192. doi:10.1186/s13059-021-02400-4

Jiang, S., Xiao, G., Koh, A. Y., Kim, J., Li, Q., and Zhan, X. (2019). A bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* 22, 522–540. doi:10.1093/biostatistics/kxz050

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* 58, 433–451. doi:10.1093/biomet/58.3.433

Kim, M., Vogtmann, E., Ahlquist, D. A., Devens, M. E., Kisiel, J. B., Taylor, W. R., et al. (2020). Fecal metabolomic signatures in colorectal adenoma patients are associated with gut microbiota and early events of colorectal cancer pathogenesis. *Mbio* 11, e03186. doi:10.1128/mbio.03186-19

Koendjbiharie, J. G., Wiersma, K., and van Kranenburg, R. (2018). Investigating the central metabolism of Clostridium thermosuccinogenes. *Appl. Environ. Microbiol.* 84, e00363-18. doi:10.1128/AEM.00363-18

Lancaster, H. O. (1958). The structure of bivariate distributions. *Ann. Math. Statistics* 29, 719–736. doi:10.1214/aoms/1177706532

Lê Cao, K.-A., Martin, P. G., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinforma.* 10, 34–17. doi:10.1186/1471-2105-10-34

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H.-W., and Wang, Y.-P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinforma.* 14, 245. doi:10.1186/1471-2105-14-245

Lin, H., and Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* 11, 3514. doi:10.1038/s41467-020-17041-7

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* 101, 785–797. doi:10.1093/biomet/asu031

Lindenbaum, O., Salhov, M., Averbuch, A., and Kluger, Y. (2022). "l_0-sparse canonical correlation analysis," in *International conference on learning representations*.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662. doi:10.1038/s41586-019-1237-9

Luo, L., Wang, W., Bao, S., Peng, X., and Peng, Y. (2024). Robust and sparse canonical correlation analysis for fault detection and diagnosis using training data with outliers. *Expert Syst. Appl.* 236, 121434. doi:10.1016/j.eswa.2023.121434

Michaeli, T., Wang, W., and Livescu, K. (2016). "Nonparametric canonical correlation analysis," in *International conference on machine learning* (New York, NY: PMLR), 48, 1967–1976.

Mohammadi-Nejad, A.-R., Hossein-Zadeh, G.-A., and Soltanian-Zadeh, H. (2017). Structured and sparse canonical correlation analysis as a brain-wide multi-modal data fusion approach. *IEEE Trans. Med. Imaging* 36, 1438–1448. doi:10.1109/TMI.2017.2681966

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc.* 1, S119. doi:10.1186/1753-6561-1-S1-S119

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.* 8, Article 1. doi:10.2202/1544-6115.1406

Pramanik, S., and Zhang, X. (2020). *Structure adaptive lasso*. arXiv preprint arXiv:2006.02041.

Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., and Knight, R. (2021). The microbiome and human cancer. *Science* 371, eabc4552. doi:10.1126/science.abc4552

Wensel, C. R., Pluznick, J. L., Salzberg, S. L., and Sears, C. L. (2022). Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J. Clin. Investigation* 132, e154944. doi:10.1172/JCI154944

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. doi:10.1093/biostatistics/kxp008

Yang, L., and Chen, J. (2022). A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome* 10, 130. doi:10.1186/s40168-022-01320-0

Yoon, G., Gaynanova, I., and Müller, C. L. (2019). Microbial networks in spring-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. Genet.* 10, 516. doi:10.3389/fgene.2019.00516

Yu, J.-L., Qian, Z.-G., and Zhong, J.-J. (2018). Advances in bio-based production of dicarboxylic acids longer than C4. *Eng. Life Sci.* 18, 668–681. doi:10.1002/elsc.201800023

Zeng, Y., Li, J., Wei, C., Zhao, H., and Wang, T. (2022). mbDenoise: microbiome data denoising using zero-inflated probabilistic principal components analysis. *Genome Biol.* 23, 94. doi:10.1186/s13059-022-02657-3

Zhou, H., He, K., Chen, J., and Zhang, X. (2022). LinDA: linear models for differential abundance analysis of microbiome compositional data. *Genome Biol.* 23, 95. doi:10.1186/s13059-022-02655-5