Statistics in Biosciences
https://doi.org/10.1007/s12561-019-09253-3

Testing for Mediation Effect with Application to Human Microbiome Data

Haixiang Zhang[1] · Jun Chen[2] · Zhigang Li[3] · Lei Liu[4]

Received: 3 January 2019 / Revised: 22 May 2019 / Accepted: 19 July 2019

Abstract

Mediation analysis has been commonly used to study the effect of an exposure on an outcome through a mediator. In this paper, we are interested in exploring the mediation mechanism of microbiome, whose special features make the analysis challenging. We consider the isometric logratio transformation of the relative abundance as the mediator variable. Then, we present a de-biased Lasso estimate for the mediator of interest and derive its standard error estimator, which can be used to develop a test procedure for the interested mediation effect. Extensive simulation studies are conducted to assess the performance of our method. We apply the proposed approach to test the mediation effect of human gut microbiome between the dietary fiber intake and body mass index.

Keywords Compositional mediators · High-dimensional data · Isometric logratio transformation · Joint significance test · Mediation analysis

1 Introduction

Mediation models were first proposed in the social science literature [4] to study the effect of an intermediate variable, termed "mediator," on the path from an exposure to an outcome. There have been substantial recent interests in mediation analysis methodology developments and applications. For example, MacKinnon et al. [25] compared methods to test the significance of the mediation effect via Monte Carlo studies. MacKinnon et al. [26] proposed to use the distribution of product and resampling methods to test an indirect effect. Preacher and Hayes [32] provided an overview of simple

✉ Lei Liu
lei.liu@wustl.edu

[1] Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

[2] Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905, USA

[3] Department of Biostatistics, University of Florida, Gainesville, FL 32610, USA

[4] Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, USA

Published online: 27 July 2019

and multiple mediation and explored three approaches to testing the mediation effect. Coffman and Zhong [12] presented marginal structural models with inverse propensity weighting for assessing mediation. Boca et al. [6] proposed a permutation approach to testing multiple biological mediators simultaneously. Gu et al. [16] proposed a state space modeling approach to mediation analysis. Fritz et al. [15] studied the combined effect of measurement errors and omitted confounders in the single-mediator model. More details on mediation analysis are referred to the review by MacKinnon [27] and Preacher [33].

There is also a challenge in estimation and inferential procedures for mediation analysis in the high-dimensional setting. Zhang et al. [49] estimated and tested the high-dimensional mediation effects using the sure independent screening (SIS; [14]) and minimax concave penalty (MCP; [47]) techniques in the selective inference framework. However, if a mediator is screened out in the first stage, we are not able to make inference for this mediator anymore. That is, inference is only considered for those selected variables; all non-selected variables are treated as non-significant with p-values set to 1. Zhao and Luo [51] proposed a sparse high-dimensional mediation model by introducing a new penalty called Pathway Lasso, but they could not conduct tests for mediation effects. Barfield et al. [3] examined the indirect effect under the null for genome-wide mediation analyses with high-dimensional mediators via marginal models, while neither the family-wise error rate (FWER; [20]) nor the false discovery rate (FDR; [5]) for multiple testing was considered. Sampson et al. [35] proposed a multiple comparison procedure to control the FWER or FDR when testing multiple mediators. However, their procedure is only based on marginal models and the selected markers may not all be true biological mediators, which are thus called “probable mediators” but not “true mediators.” Of note, the above literature cannot be adopted directly to make inference for a specific mediator in the presence of high-dimensional nuisance confounders. We therefore propose an approach to estimate and test a mediator of interest among a large number of mediators via the de-biased Lasso technique [48].

In this paper, we are interested in exploring the mediation mechanism of microbiome on the path from an exposure to a health outcome. Our motivating example is a human gut microbiome study. Gut microbiota were obtained on 98 healthy subjects using fecal 16S sequencing [45]. We thus have the abundance (count) of each taxon in the microbiome. Zhang et al. [50] showed a significant negative association between the fiber intake and body mass index (BMI). A question arises as whether the association between the fiber intake and BMI is mediated by the gut microbiota. Of note, since the number of taxa varied greatly across samples, these count data were transformed into compositions after zero counts were replaced by 0.5 [9]. Moreover, the number of taxa (1234) considered is high-dimensional, and much larger than the number of samples (98), i.e., $p \gg n$. The high-dimensional and compositional characteristics pose new challenges to existing mediation analysis methods. To solve these issues, we first adopt the isometric logratio (*ilr*)-based transformation on compositional mediators, and refit these *ilr*-transformed variables via standard linear regression models; next, we apply our testing method towards the first component in these *ilr* variables, where the interpretation of the first *ilr* variable is meaningful and straightforward [17].

The rest of this article is organized as follows. In Sect. 2, we apply our proposed mediation analysis method of microbiome data based on the *ilr* transformation. In Sect. 3, we employ the de-biased Lasso technique, together with the joint significance test method to evaluate the mediation effect of interest in the presence of a large number of nuisance variables. In Sect. 4, some simulation studies are conducted to examine the performance of the proposed method. In Sect. 5, we provide an application to the human microbiome study. Some concluding remarks are given in Sect. 6.

2 Methodology

2.1 Traditional Mediation Model

The goal of mediation analysis is to investigate the effect of an exposure X on an outcome Y through intermediate variables, referred to as "mediators" [4]. In the literature, mediation analysis can be roughly divided into two categories: the structural equation modeling framework [27] and the counterfactual framework [21]. Below, our method belongs to the first category, which focuses on the following regression equations:

$$\begin{aligned} Y &= c^* + \gamma^* X + \mathbf{Z}'\boldsymbol{\eta}^* + \zeta, \\ Y &= c + \gamma X + \mathbf{M}'\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\eta} + \epsilon, \\ M_k &= c_k + \alpha_k X + \mathbf{Z}'\boldsymbol{\theta}_k + e_k, \quad k = 1, \ldots, p. \end{aligned} \tag{2.1}$$

Here Y is an outcome variable, X is an exposure, $\mathbf{M} = (M_1, \ldots, M_p)'$ is a vector of mediators, $\mathbf{Z} = (Z_1, \ldots, Z_q)'$ represents the covariates such as age, sex, and weight; γ^* represents the total effect of the exposure X on the outcome Y adjusting for the effects of covariates; α_k represents the relation between X and M_k; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the regression coefficient vector with β_k representing the relation between M_k and Y adjusting for the effects of X and $\mathbf{Z}$; γ represents the direct effect of X on Y adjusting for the effects of $\mathbf{M}$ and $\mathbf{Z}$; $\boldsymbol{\eta}^*$, $\boldsymbol{\eta}$, and $\boldsymbol{\theta}_k$ denote the regression coefficients of $\mathbf{Z}$; c^*, c, and c_k represent regression intercepts; ζ, ϵ, and e_k are the error terms, $k = 1, \ldots, p$. By (2.1), it is straightforward to deduce that the total effect of X on Y can be written as $\gamma^* = \gamma + \sum_{i=1}^{p} \alpha_k \beta_k$. In this case, the term $\alpha_k \beta_k$ denotes the mediation effect by M_k, $k = 1, \ldots, p$. Following the framework of mediation analysis, our basic task is to make inference on the *product coefficient* $\alpha_k \beta_k$ [27] for $k = 1, \ldots, p$.

2.2 Isometric Logratio Transformation for Microbiome Data

Suppose there are p taxa in the microbiome for each sample, whose relative abundances are denoted by a vector $\mathbf{M} = (M_1, \ldots, M_p)'$. The p-part composition $\mathbf{M}$ lies in a space termed the "simplex" [1], which is given as

$$\mathcal{S}^p = \left\{ \mathbf{x} = (x_1, \ldots, x_p)' : x_k > 0, k = 1, \ldots, p; \sum_{k=1}^{p} x_k = 1 \right\}.$$

Compositions are subject to two constraints: the components are non-negative in (0, 1), and sum up to one. Thus, classical regression models in the real Euclidean space cannot be used to analyze the relative abundance directly [2]. For example, Hron et al. [17] indicated that the naive approach for traditional regression with the original explanatory variables would lead to misleading results, due to the fact that any p−1 variables may contain the same information as all p variables.

One key point about the statistical analysis of $\mathbf{M}$ is to express it in orthonormal coordinates with respect to the Aitchison geometry, then we can apply the well-established statistical methods in the Euclidean space. For this purpose, Egozcue et al. [13] suggested the isometric logratio (*ilr*) transformation technique by transforming the compositional data from the simplex $\mathcal{S}^p$ to the Euclidean space $\mathbb{R}^{p-1}$ in a distance-preserving manner. We can use the new *ilr* coordinates in a standard linear regression model. The details are given below.

Step 1: Conduct *ilr*-based transformation on the compositional mediators $M_1, \ldots, M_p$,

$$\tilde{M}_k = \sqrt{\frac{p-k}{p-k+1}} \ln \frac{M_k}{\sqrt[p-k]{\prod_{j=k+1}^{p} M_j}}, k = 1, \ldots, p-1. \tag{2.2}$$

Step 2: Refit a linear regression model as (2.1) in the Euclidean space.

$$\begin{aligned} Y &= c + \gamma X + \beta_1 \tilde{M}_1 + \cdots + \beta_{p-1} \tilde{M}_{p-1} + \mathbf{Z}'\boldsymbol{\eta} + \epsilon, \\ \tilde{M}_k &= c_k + \alpha_k X + \mathbf{Z}'\boldsymbol{\theta}_k + e_k, \quad k = 1, \ldots, p-1, \end{aligned} \tag{2.3}$$

Step 3: Testing for mediation effect towards $\tilde{M}_1$ based on the method in Sect. 3,

$$H_0 : \alpha_1\beta_1 = 0 \text{ vs. } H_1 : \alpha_1\beta_1 \neq 0.$$

2.3 Interpretation of the *ilr*-Transformed Variables

By (2.2), the *ilr*-transformed mediator $\tilde{M}_1$ is a scaled sum of all logratios of the original composition part M_1 and the other parts $M_2, \ldots, M_p$, where the linear relationship is described as

$$\tilde{M}_1 = \frac{1}{\sqrt{p(p-1)}} \left(\ln \frac{M_1}{M_2} + \cdots + \ln \frac{M_1}{M_p} \right) = \sqrt{\frac{p-1}{p}} \ln \frac{M_1}{\sqrt[p-1]{\prod_{j=2}^{p} M_j}}. \tag{2.4}$$

It is straightforward to realize that $\tilde{M}_1$ is formed by a logratio between the compositional part M_1 and the geometric mean of the remaining parts in the composition. Therefore, the *ilr* variable $\tilde{M}_1$ represents a measure of dominance of the compositional part M_1 with respect to the other parts [17]. In this case, the interpretation of the parameter β_1 in (2.4) indicates how much the response variable Y changes in average by a unit change of the $\tilde{M}_1$, which is the logarithm of the ratio between M_1 and the geometric mean of the $M_2, \ldots, M_p$ in the composition. Due to the compositional framework with $\sum_{i=1}^{p} M_i = 1$, it is not feasible to simultaneously vary one composition and keep other compositions unchanged. To be more specific, the working mechanism of microbiome (e.g., taxa) in human body is complex and the composition of microbiome is dynamic. In practice, it may be reasonable to research the "relative effect" rather than the "absolute effect" of each composition. The *ilr*-variable $\tilde{M}_1$ in (2.4) can play the role of "relative effect" from the composition M_1. In the literature, the above-mentioned interpretations with "relative effect" of compositions have also been adopted in the fields of macroeconomics [18], epidemiology [30], market shares [31], etc.

Based on the above clarification, if the hypothesis test $H_0 : \alpha_1\beta_1 = 0$ is rejected towards the *ilr*-transformed mediator $\tilde{M}_1$, we can say that the original composition M_1 has a significant mediation effect in comparison to the rest of compositions. In other words, the mediation effect $\alpha_1\beta_1$ of $\tilde{M}_1$ can reflect the relative mediation transmission capacity of the original composition M_1 by expression (2.4). However, the remaining *ilr*-variables $\tilde{M}_2, \ldots, \tilde{M}_{p-1}$ from (2.2) are not easy to interpret, because the original composition part M_1 is not contained therein. Therefore, if we are interested in exploring the mediation effects for other taxa M_ℓ, $\ell \in \{2, \ldots, p\}$, we can reorder M_ℓ to play the role of M_1 as $(M_\ell, M_1, \ldots, M_{\ell-1}, M_{\ell+1}, \cdots, M_p)'$, then run Steps 1–3 in Sect. 2.2 again to interpret the corresponding mediation effect of composition M_ℓ.

3 Inference on the *ilr*-Transformed Mediation Effect

Motivated by the above compositional taxa data, we may face the problem to estimate and test a specific mediator of interest in the presence of high-dimensional mediators (Fig. 1). Furthermore, as described in Sect. 2.2, the *ilr* transformation will be used for compositional mediators. In this section, we will give the estimation and inference procedures for a specific mediator after the *ilr* transformation.

Without loss of generality, assume we are interested in testing the first mediator $\tilde{M}_1$. Here $\alpha_1\beta_1$ is the parameter of interest, and $\boldsymbol{\theta} = (\alpha_2\beta_2, \ldots, \alpha_{p-1}\beta_{p-1})'$ is the vector of "nuisance" parameters which need to be adjusted for. Our aim is to estimate $\alpha_1\beta_1$ and construct the p-value for testing $H_0 : \alpha_1\beta_1 = 0$ vs. $H_1 : \alpha_1\beta_1 \neq 0$.

Denote $(X_i, \tilde{\mathbf{M}}_i, Y_i)$ as the triplet sample, where $\tilde{\mathbf{M}}_i = (\tilde{M}_{i1}, \ldots, \tilde{M}_{i(p-1)})'$ is the mediator vector, $i = 1, \ldots, n$. For α_1, the ordinary least squares (OLS) estimator is denoted by $\hat{\alpha}_1$, and its corresponding variance estimate is $\hat{\sigma}^2_{\alpha_1}$. As we know, the OLS estimator of β_1 is not unique when the number of mediators p is larger than the sample size n. To solve this problem, we employ the de-biased Lasso technique [48] to derive the estimator of β_1. For convenience, we assume the intercepts c and c_k in (2.3) are zeros. Let

$$(\tilde{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\eta}}) = \arg\min_{\gamma,\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \gamma X_i - \sum_{j=1}^{p-1} \beta_j \tilde{M}_{ij} - \sum_{j=1}^{q} \eta_j Z_{ij} \right)^2 + \lambda \sum_{j=1}^{p-1} |\beta_j| \right\}, \tag{3.1}$$

where $\lambda > 0$ is the Lasso penalty parameter [39]. The de-biased Lasso estimator of β_1 is given by

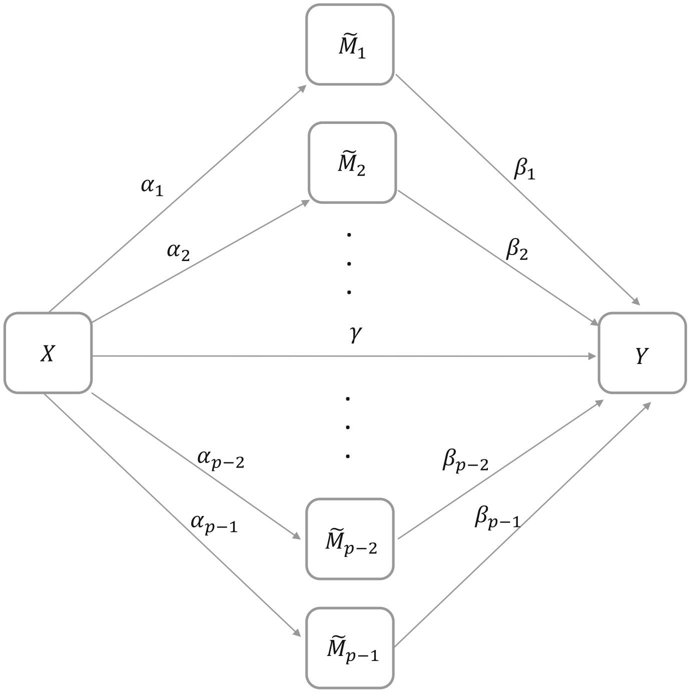

Fig. 1 A scenario of high-dimensional *ilr*-transformed mediation model with omitted confounders

$$\hat{\beta}_1 = \tilde{\beta}_1 + \frac{\sum_{i=1}^n R_i(Y_i - \tilde{\gamma}X_i - \sum_{j=1}^{p-1}\tilde{\beta}_j\tilde{M}_{ij} - \sum_{j=1}^q \tilde{\eta}_j Z_{ij})}{\sum_{i=1}^n R_i\tilde{M}_{i1}}, \tag{3.2}$$

where $\tilde{\gamma}$ and $\tilde{\boldsymbol{\beta}}$ are defined in (3.1); $R_i = \tilde{M}_{i1} - \hat{\phi}_1 X_i - \sum_{j=2}^{p-1}\hat{\phi}_j\tilde{M}_{ij} - \sum_{j=1}^q \hat{\phi}_{p-1+j}Z_{ij}$ is the residual from a Lasso regression of $\tilde{M}_{i1}$ versus X_i, $\mathbf{Z}_i$ and $\tilde{\mathbf{M}}_i$, $i = 1, \cdots, n$, and $\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \cdots, \hat{\phi}_{p+q-1})'$ is the Lasso solution from

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \left\{ \frac{1}{2n}\sum_{i=1}^n \left(\tilde{M}_{i1} - \phi_1 X_i - \sum_{j=2}^{p-1}\phi_j\tilde{M}_{ij} - \sum_{j=1}^q \phi_{p-1+j}Z_{ij} \right)^2 + \lambda^* \sum_{j=1}^{p+q-1} |\phi_i| \right\},$$

where $\lambda^* > 0$ is the Lasso penalty parameter. From (3.2), $\hat{\beta}_1$ is Lasso plus a one-step bias correction, and hence it is named “de-biased Lasso.”

It has been shown by Zhang and Zhang [48] that $(\hat{\beta}_1 - \beta_1)/\sigma_{\beta_1} \xrightarrow{\mathcal{D}} N(0, 1)$, where $\hat{\beta}_1$ is the de-biased Lasso estimator in (3.2), $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution, and the estimation of the standard error is given as

$$\hat{\sigma}_{\beta_1} = n^{-1/2} \frac{\hat{\sigma}_\epsilon \sqrt{\sum_{i=1}^n R_i^2/n}}{|\sum_{i=1}^n R_i \tilde{M}_{i1}/n|}, \tag{3.3}$$

where $\hat{\sigma}_\epsilon^2 = \sum_{i=1}^n (Y_i - \tilde{\gamma} X_i - \sum_{j=1}^{p-1} \tilde{\beta}_j \tilde{M}_{ij} - \sum_{j=1}^q \tilde{\eta}_j Z_{ij})^2/(n - \hat{s})$ is based on the recommendation of Reid et al. [34], and $\hat{s}$ is the number of non-zero coefficients in the Lasso estimator $\tilde{\boldsymbol{\beta}}$.

To test the mediation effect $\alpha_1\beta_1$, we will adopt the *joint significance test* as in our previous work [49]. Specifically, the p-value is given by $P_{joint} = \max\{P_a, P_b\}$, with $P_a = 2(1-\Phi(|\hat{\alpha}_1|/\hat{\sigma}_{\alpha_1}))$ and $P_b = 2(1-\Phi(|\hat{\beta}_1|/\hat{\sigma}_{\beta_1}))$, where $\Phi(x)$ is the distribution function of $N(0, 1)$; $\hat{\alpha}_1$ and $\hat{\sigma}_{\alpha_1}$ are based on the OLS method; $\hat{\beta}_1$ and $\hat{\sigma}_{\beta_1}$ are defined in (3.2) and (3.3), respectively. Note that besides the joint significance test, other tests for the indirect effect can be considered: (a) methods based on the distribution of the product of two normal random variables, and (b) resampling methods. First, the product of the two normal random variables is not normal, but a Bessel function of the second kind. However, even the Bessel function does not work well in finite samples [26]. On the other hand, the resampling methods, e.g., the bias-corrected bootstrap, can provide better inference results, at the price of computational burden.

Remark Sohn and Li [37] proposed a compositional mediation framework to investigate the mediated effect of gut microbiome between fat intake and body mass index. There are three different aspects from our proposed method. First, Sohn and Li [37] established a compositional mediation model directly in the simplex space, while we use the ilr-transformed mediators to construct a high-dimensional mediation model in the Euclidean space. Second, Sohn and Li [37] adopted the additive logratio (alr) transformation on the composition $\mathbf{M}$. They focused on understanding the mediation effect of individual composition, and we study the relative mediation effect denoted by a specific composition in contrast to the rest of compositions. Third, Sohn and Li [37] used the Sobel test [36] to identify the significant compositional mediators, while we employ the joint significance test [25] in our method.

4 Simulation Study

In this section, we conduct simulations to examine the performance of our proposed method. Of note, the isometric logratio transformation (2.2) is needed for compositional data [17]. Without loss of generality, we will only focus on the performance of testing the first ilr-transformed mediator via simulation. For this goal, we generate data from Model (2.3) using R software, where the exposure X follows from $N(0, 1.5)$, the covariate Z follows from $N(0, 2)$, and ϵ is generated from $N(0, 1)$, together with $c = 0$, $\gamma = 0.5$, $\eta = 0.5$, $\beta = (\beta_1, 0.25, 0.30, 0.35, 0.55, 0, \cdots, 0)'$ with $p = 1234$ (the dimension of taxa in Sect. 5). Here we set $\beta_1 = 0$, 0.15, 0.25, 0.35, respectively. For the mediators $\tilde{M}_k$, we set $c_k = 0$, $\theta_k = 0.5$, and $\alpha = (\alpha_1, 0.15, 0.25, 0.35, 0.55, 0, \cdots, 0)'$ with $\alpha_1 = 0, 0.15, 0.25, 0.35$, respectively. We consider three cases for the generation of error term $\mathbf{e} = (e_1, \cdots, e_{p-1})'$ as follows:

Case I: e_i are independent and identically distributed $t(5)$ random variables, $i = 1, \ldots, p-1$;

Case II: $\mathbf{e} \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\Sigma_{ij})$ with $\Sigma_{ij} = 0.3^{|i-j|}$ for all $i, j = 1, \ldots, p-1$.

Case III: $\mathbf{e} \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the sample covariance matrix of the *ilr*-transformed mediators $\tilde{\mathbf{M}}$ in Sect. 5, and $\tilde{M}_1$ is corresponding to the taxon with ID = 14531.

For comparison, we also fit the data using marginal regression

$$Y = c + \gamma X + \beta_1 \tilde{M}_1 + \eta Z + \epsilon$$

with only one mediator $\tilde{M}_1$ (Naive). As pointed out by Preacher and Hayes [32], multiple mediators contribute to the outcome Y (as shown in Fig. 1). Thus it is imperative to adjust for other mediators in such analysis, especially given the potential correlations between different mediators. Furthermore, it is not feasible to predict Y using only one mediator in this naive model [49].

Of note, since we are only interested in 1 mediator (the first one), no multiple testing adjustment is needed in all three settings. Also, P_a, the p-value for exposure–mediator association is the same in these methods. So only P_b, the p-value from the mediator to the outcome is different, which impacts the overall p-value in the joint significance test. For the estimation of α_1, β_1, and $\alpha_1\beta_1$, we report the bias (BIAS) given by the sample mean of an estimate minus the true value, and the mean standard error (MSE) of an estimate in Tables 1, 2, and 3. We report the size and power of the test methods in Tables 4, 5, and 6. All results are based on 1000 replications with sample size $n = 100$ and 300, respectively.

It can be seen from Tables 1, 2, and 3 that our method is unbiased in all three cases. In contrast, the Naive method is unbiased only in Case I with independent mediators, and biased in the cases of correlated mediators (Cases II and III) towards the estimation of β_1. The estimation results of α_1 from our method and the Naive approach are exactly the same, since these two methods consider the same exposure–mediator association. Moreover, the BIAS and MSE of $\hat{\alpha}_1$ do not rely on the true value of parameter α_1, because we employ the ordinary least squares (OLS) to estimate α_1. Thus, the BIAS and MSE of $\hat{\alpha}_1$ are exactly the same for different values of α_1. Similarly, $\hat{\beta}_1$ for the naive method does not vary with β_1. However, $\hat{\beta}_1$ changes with β_1 for our proposed method. Finally, $\hat{\alpha}_1\hat{\beta}_1 - \alpha_1\beta_1 = (\hat{\alpha}_1 - \alpha_1)\hat{\beta}_1 + \alpha_1(\hat{\beta}_1 - \beta_1) = (\hat{\beta}_1 - \beta_1)\hat{\alpha}_1 + \beta_1(\hat{\alpha}_1 - \alpha_1)$ varies with different α_1 and β_1 for both methods.

From Tables 5 and 6, the Naive estimate has inflated size when the mediators are correlated in the case of $(\alpha_1, \beta_1) = (0.15, 0)$, which will result in too many false discoveries. Thus, the Naive method is not appropriate for estimating and testing the mediator $\tilde{M}_1$. For ($\alpha_1 = 0, \beta_1 \neq 0$) or ($\alpha_1 \neq 0, \beta_1 = 0$), the sizes of our method are close to 0.05. For $\alpha_1 = \beta_1 = 0$, the sizes are more conservative, which is a common fact in mediation analysis. For example, such an effect is observed even in the single-mediator model [25].

Table 1 BIAS and MSE (in parenthesis) for the estimators in Case I

	(α_1, β_1)	Proposed			Naive		
		$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1\hat{\beta}_1$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1\hat{\beta}_1$
$n = 100$	(0, 0)	− 0.0025	0.0034	0.0008	− 0.0025	0.0012	0.0002
		(0.0878)	(0.0902)	(0.0072)	(0.0878)	(0.1135)	(0.0093)
	(0.15, 0)	− 0.0025	0.0223	0.0042	− 0.0025	0.0012	0.0004
		(0.0878)	(0.1351)	(0.0251)	(0.0878)	(0.1135)	(0.0197)
	(0, 0.15)	− 0.0025	− 0.0641	0.0010	− 0.0025	0.0012	− 0.0002
		(0.0878)	(0.1079)	(0.0110)	(0.0878)	(0.1135)	(0.0161)
	(0.15,0.15)	− 0.0025	− 0.0380	− 0.0047	− 0.0025	0.0012	0.0001
		(0.0878)	(0.1529)	(0.0301)	(0.0878)	(0.1135)	(0.0239)
	(0.25,0.25)	− 0.0025	− 0.0419	− 0.0098	− 0.0025	0.0012	− 0.0001
		(0.0878)	(0.1573)	(0.0475)	(0.0878)	(0.1135)	(0.0377)
	(0.35,0.35)	− 0.0025	− 0.0346	− 0.0119	− 0.0025	0.0012	− 0.0002
		(0.0878)	(0.1681)	(0.0693)	(0.0878)	(0.1135)	(0.0519)
$n = 300$	(0, 0)	− 0.0013	0.0027	0.0002	− 0.0013	− 0.0005	0.0001
		(0.0494)	(0.0384)	(0.0020)	(0.0494)	(0.0635)	(0.0031)
	(0.15, 0)	− 0.0013	0.0105	0.0016	− 0.0013	− 0.0005	− 0.0001
		(0.0494)	(0.0391)	(0.0061)	(0.0494)	(0.0635)	(0.0099)
	(0, 0.15)	− 0.0013	− 0.0193	0.0001	− 0.0013	− 0.0005	− 0.0002
		(0.0494)	(0.0495)	(0.0068)	(0.0494)	(0.0635)	(0.0079)
	(0.15,0.15)	− 0.0013	− 0.0076	− 0.0012	− 0.0013	− 0.0005	− 0.0003
		(0.0494)	(0.0502)	(0.0106)	(0.0494)	(0.0635)	(0.0123)
	(0.25,0.25)	− 0.0013	− 0.0058	− 0.0017	− 0.0013	− 0.0005	− 0.0005
		(0.0494)	(0.0505)	(0.0178)	(0.0494)	(0.0635)	(0.0202)
	(0.35,0.35)	− 0.0013	− 0.0043	− 0.0019	− 0.0013	− 0.0005	− 0.0006
		(0.0494)	(0.0505)	(0.0249)	(0.0494)	(0.0635)	(0.0281)

“Naive” is the marginal regression method. For the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, the bias of the OLS estimator $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$; thus, BIAS$(\hat{\alpha}_1) = \hat{\alpha}_1 - \alpha_1$ does not depend on the true value of α_1 for both methods. The same conclusion holds for $\hat{\beta}_1$ for the naive method but not for our proposed method. However, $\hat{\alpha}_1\hat{\beta}_1 - \alpha_1\beta_1 = (\hat{\alpha}_1 - \alpha_1)\hat{\beta}_1 + \alpha_1(\hat{\beta}_1 - \beta_1) = (\hat{\beta}_1 - \beta_1)\hat{\alpha}_1 + \beta_1(\hat{\alpha}_1 - \alpha_1)$ varies with different α_1 and β_1 for both methods

5 Application to Gut Microbiome Data

In this section, we apply our test procedure to a human gut microbiome data set, which includes 98 healthy subjects who were not on antibiotics for 3 months prior to data collection [45]. The subjects’ long-term diet information was gathered by food frequency questionnaire and converted to intake amounts of different nutrient categories. In this study, we consider the fiber intake assessed by percent calories from dietary fiber (square-root transformed as in Zhang et al. [50]) as the exposure. Body mass index (BMI) was measured as the outcome. The fiber intake demonstrates a significant negative association with BMI, and the gut microbiota is associated with

Table 2 BIAS and MSE (in parenthesis) for the estimators in Case II

	(α_1, β_1)	Proposed			Naive		
		$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1\hat{\beta}_1$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1\hat{\beta}_1$
$n = 100$	(0, 0)	−0.0014	0.0172	0.0005	−0.0014	0.1048	0.0002
		(0.0680)	(0.1053)	(0.0076)	(0.0680)	(0.1434)	(0.0120)
	(0.15, 0)	−0.0014	0.0334	0.0053	−0.0014	0.1048	0.0159
		(0.0680)	(0.1095)	(0.0186)	(0.0680)	(0.1434)	(0.0241)
	(0, 0.15)	−0.0014	−0.0457	0.0006	−0.0014	0.1048	0.0001
		(0.0680)	(0.1232)	(0.0109)	(0.0680)	(0.1434)	(0.0200)
	(0.15,0.15)	−0.0014	−0.0215	−0.0029	−0.0014	0.1048	0.0157
		(0.0680)	(0.1325)	(0.0242)	(0.0680)	(0.1434)	(0.0292)
	(0.25,0.25)	−0.0014	−0.0252	−0.0061	−0.0014	0.1048	0.0260
		(0.0680)	(0.1527)	(0.0429)	(0.0680)	(0.1434)	(0.0444)
	(0.35,0.35)	−0.0014	−0.0170	−0.0060	−0.0014	0.1048	0.0364
		(0.0680)	(0.1611)	(0.0621)	(0.0680)	(0.1434)	(0.0600)
$n = 300$	(0, 0)	−0.0001	0.0089	0.0001	−0.0001	0.1161	−0.0001
		(0.0395)	(0.0499)	(0.0019)	(0.0395)	(0.0779)	(0.0055)
	(0.15, 0)	−0.0001	0.0191	0.0029	−0.0001	0.1161	0.0173
		(0.0395)	(0.0508)	(0.0079)	(0.0395)	(0.0779)	(0.0126)
	(0, 0.15)	−0.0001	−0.0132	0.0002	−0.0001	0.1161	−0.0001
		(0.0395)	(0.0618)	(0.0057)	(0.0395)	(0.0779)	(0.0110)
	(0.15,0.15)	−0.0001	0.0020	0.0003	−0.0001	0.1161	0.0173
		(0.0395)	(0.0628)	(0.0115)	(0.0395)	(0.0779)	(0.0157)
	(0.25,0.25)	−0.0001	0.0045	0.0011	−0.0001	0.1161	0.0289
		(0.0395)	(0.0640)	(0.0191)	(0.0395)	(0.0779)	(0.0240)
	(0.35,0.35)	−0.0001	0.0070	0.0024	−0.0001	0.1161	0.0405
		(0.0395)	(0.0639)	(0.0266)	(0.0395)	(0.0779)	(0.0325)

Please refer to the footnotes in Table 1

both fiber intake and BMI [50]. It is of great clinical significance to know whether the association between the fiber intake and BMI is mediated by the gut microbiota.

In between exposure and outcome, subjects' stool samples were collected and the DNA samples were analyzed by Roche 454 pyrosequencing of 16S rDNA gene segments. We thus have the abundance (count) of each taxon in the microbiome. Of note, the number of taxa in the microbiome data set is high-dimensional. There is also sparsity in the data as many taxa are absent across samples [28]. Similar to Bokulich et al. [7] and Yun et al. [46], we removed a taxon if it appears in fewer than 8% of the samples, leaving 1234 taxa in 98 samples ($p \gg n$). Next, since the number of sequencing reads varied greatly across samples, these count data were transformed into compositions after zero counts were replaced by the maximum rounding error 0.5 [9,23]. Thus, the potential mediators ($\mathbf{M}$) are compositional abundances of 1234 taxa. To remove the compositional effects, we calculated the isometric logratio transformed $\tilde{\mathbf{M}}$ as in (2.2). For analysis, X and $\tilde{\mathbf{M}}$ are further standardized with mean 0 and variance 1.

Table 3 BIAS and MSE (in parenthesis) for the estimators in Case III

	(α_1, β_1)	Proposed			Naive		
		$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1\hat{\beta}_1$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1\hat{\beta}_1$
$n = 100$	(0, 0)	− 0.0012	0.0295	0.0005	− 0.0012	− 0.0898	0.0003
		(0.0660)	(0.2250)	(0.0147)	(0.0660)	(0.1375)	(0.0112)
	(0.15, 0)	− 0.0012	0.0469	0.0072	− 0.0012	− 0.0898	− 0.0131
		(0.0660)	(0.2183)	(0.0351)	(0.0660)	(0.1375)	(0.0235)
	(0, 0.15)	− 0.0012	− 0.0283	0.0004	− 0.0012	− 0.0898	0.0002
		(0.0660)	(0.2359)	(0.0175)	(0.0660)	(0.1375)	(0.0101)
	(0.15,0.15)	− 0.0012	− 0.0036	− 0.0003	− 0.0012	− 0.0898	− 0.0133
		(0.0660)	(0.2345)	(0.0393)	(0.0660)	(0.1375)	(0.0233)
	(0.25,0.25)	− 0.0012	− 0.0109	− 0.0024	− 0.0012	− 0.0898	-0.0224
		(0.0660)	(0.2471)	(0.0656)	(0.0660)	(0.1375)	(0.0376)
	(0.35,0.35)	− 0.0012	− 0.0030	− 0.0009	− 0.0012	− 0.0898	− 0.0315
		(0.0660)	(0.2498)	(0.0918)	(0.0660)	(0.1375)	(0.0525)
$n = 300$	(0, 0)	0.0005	0.0178	− 0.0001	0.0005	− 0.0852	− 0.0001
		(0.0406)	(0.0693)	(0.0030)	(0.0406)	(0.0741)	(0.0046)
	(0.15, 0)	0.0005	0.0272	0.0039	0.0005	− 0.0852	− 0.0129
		(0.0406)	(0.0698)	(0.0108)	(0.0406)	(0.0741)	(0.0119)
	(0, 0.15)	0.0005	− 0.0310	0.0001	0.0005	− 0.0852	0.0001
		(0.0406)	(0.0828)	(0.0057)	(0.0406)	(0.0741)	(0.0040)
	(0.15,0.15)	0.0005	− 0.0058	− 0.0010	0.0005	− 0.0852	− 0.0128
		(0.0406)	(0.0880)	(0.0146)	(0.0406)	(0.0741)	(0.0116)
	(0.25,0.25)	0.0005	0.0038	0.0009	0.0005	− 0.0852	− 0.0213
		(0.0406)	(0.0900)	(0.0244)	(0.0406)	(0.0741)	(0.0197)
	(0.35,0.35)	0.0005	0.0121	0.0042	0.0005	− 0.0852	-0.0297
		(0.0406)	(0.0891)	(0.0338)	(0.0406)	(0.0741)	(0.0280)

Please refer to the footnotes in Table 1

Table 4 Size and power at significance level 0.05 in Case I

(α_1, β_1)	$n = 100$		$n = 300$	
	Proposed	Naive	Proposed	Naive
(0, 0)	0.003	0.002	0.003	0.005
(0.15, 0)	0.026	0.022	0.044	0.034
(0, 0.15)	0.011	0.014	0.039	0.030
(0.15, 0.15)	0.169	0.123	0.761	0.549
(0.25, 0.25)	0.581	0.480	0.999	0.977
(0.35, 0.35)	0.901	0.843	1	1

"Naive" is the marginal regression method

We evaluate the mediation test on individual taxon abundance one by one using the proposed approach in Sect. 3, where three taxa are significant with p-values smaller than 0.05. In Table 7, we give the estimates, standard errors, and p-values for those

Table 5 Size and power at significance level 0.05 in Case II

(α_1, β_1)	$n = 100$		$n = 300$	
	Proposed	Naive	Proposed	Naive
(0, 0)	0.003	0.003	0.001	0.012
(0.15, 0)	0.036	0.065	0.054	0.308
(0, 0.15)	0.008	0.026	0.028	0.043
(0.15, 0.15)	0.201	0.278	0.750	0.901
(0.25, 0.25)	0.591	0.680	0.988	0.995
(0.35, 0.35)	0.883	0.898	1	1

“Naive” is the marginal regression method

Table 6 Size and power at significance level 0.05 in Case III

	$n = 100$		$n = 300$	
(α_1, β_1)	Proposed	Naive	Proposed	Naive
(0, 0)	0.004	0.006	0.001	0.011
(0.15, 0)	0.048	0.054	0.023	0.184
(0, 0.15)	0.006	0.004	0.011	0.008
(0.15, 0.15)	0.108	0.061	0.377	0.127
(0.25, 0.25)	0.331	0.199	0.824	0.588
(0.35, 0.35)	0.529	0.499	0.984	0.936

“Naive” is the marginal regression method

Table 7 Estimates and p-values of potential mediating taxa (unadjusted p-value < 0.05)

ID	Phylum	Class	Order	Family	Genus	$\hat{\alpha}$ (SE)	$\hat{\beta}$ (SE)	P_{joint}
14477	F	C	C*	V	Other	− 0.2305 (0.0993)	1.2412 (0.4252)	0.0202
10485	F	C	C*	V	M	− 0.2258 (0.0994)	1.4136 (0.4516)	0.0231
6167	B	B	B*	B**	B***	− 0.2258 (0.0994)	2.4668 (0.7434)	0.0231

$P_{joint} = \max\{P_a, P_b\}$

F Firmicutes, *C* Clostridia, *C** Clostridiales, *V* Veillonellaceae, *M* Megasphaera, *B* Bacteroidetes, *B** Bacteroidales, *B*** Bacteroidaceae, *B**** Bacteroides

potential significant mediators. More specifically, Trompette et al. [40] found that dietary fermentable fiber content changed the composition of the gut microbiota, in particular by altering the ratio of *Firmicutes* to *Bacteroidetes*. Moreover, Ismail et al. [22] suggested that the proportions of the *Firmicutes* and *Bacteroidetes* may play an important role in the pathogenesis of obesity.

To adjust for multiple testing, we apply the FDR control. None of the taxa is significant under the FDR control, which is in line with the conclusion of Zhang et al. [50]. Although none of the associations survived the multiple testing correction, the identified nominally significant taxa, coupled with strong biological evidence, justified a future large sample study.

6 Conclusion and Remarks

In this paper, we have proposed an approach to estimating and testing a specific mediator of interest adjusting for other high-dimensional mediators and confounded covariates. Furthermore, we can employ the proposed method for high-dimensional compositional data based on the *ilr* transformation. The simulation and real data application indicate that the proposed method is feasible in practice.

A closely related topic is to study the combined or overall effects of high-dimensional mediators altogether rather than a specific mediator in the presence of high-dimensional confounders. For example, Huang and Pan [19] proposed a transformation model using spectral decomposition to evaluate the combined mediation effects of high-dimensional continuous mediators. Chén et al. [11] introduced a novel direction of mediation (DM) approach by linearly combining potential mediators into a smaller number of orthogonal components in the high-dimensional setting, where the components are ranked by the proportion of the likelihood. Zhang et al. [50] proposed a distance-based approach for testing the overall mediation effect of the human microbiome with multiple mediators.

For the application to microbiome data, we have considered the high-dimensional and compositional nature of bacterial taxa. Moreover, the microbiome data are structured in the sense that bacterial taxa are related to each other by a phylogenetic tree [38,44]. The adaption of our method to the tree-guided strategy merits further consideration. In addition, since the bacterial taxa are correlated to each other, we could employ the elastic net [52] penalized criterion in (3.2). The theoretical properties of this elastic net-based approach needs further careful research.

Another feature of the microbiome data is the presence of zero values. In the Application, we simply replaced zero values by 0.5. More rigorous consideration to dealing with zeros in compositional data using non-parametric imputation was given by Martín-Fernández et al. [29]. Furthermore, when there are a large portion of zero values, two part models, e.g., Chen and Li [10], Chai al. [8], and Liu et al. [24], can be used to separately model the odds of the presence of zero values and the amount of positive values.

As mentioned before, it is of interest to consider the multiple testing problem when the target is a set of mediators rather than a single mediator. Here a possible solution is to use Sampson et al. [35] multiple comparison procedure by replacing the p-value for mediator–outcome association with our de-biased Lasso-based p-value in Sect. 3. Details of this method remain to be explored.

Of note, in the joint significance test procedure, the calculation of P_a originates from the ordinary least squares estimation for linear models, and that of P_b is derived from de-biased lasso, which is also not based on the normal distribution assumption for the outcome variable and *ilr*-transformed mediators [48]. As pointed out by one reviewer, it is desirable to consider more robust inference method for future research.

Finally, in this paper, we adopted the structural equation modeling approach for mediation analysis. The counterfactual approach, originated from causal inference, can be used to define a causal effect. Examples of counterfactual mediation analysis include VanderWeele [42,43] and Imai et al. [21]. These approaches can decompose the total effect into direct and indirect effects without linear assumptions. Their application to the microbiome data analysis should be further pursued.

Acknowledgements The authors would like to thank the Editor, the Associate Editor, and two reviewers for their constructive and insightful comments and suggestions that greatly improved the manuscript. Research reported in this publication was supported by the NIH R21 AG063370 and UL1 TR002345. The content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH.

References

1. Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London
2. Aitchison J (1999) Logratios and natural laws in compositional data analysis. Math Geol 31:563–580
3. Barfield R, Shen J, Just A, Vokonas P, Schwartz J, Baccarelli A, VanderWeele T, Lin X (2017) Testing for the indirect effect under the null for genome-wide mediation analyses. Genet Epidemiol 41:824–833
4. Baron R, Kenny D (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical consideration. J Personal Soc Psychol 51:1173–1182
5. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300
6. Boca S, Sinha R, Cross A, Moore S, Sampson J (2014) Testing multiple biological mediators simultaneously. Bioinformatics 30:214–220
7. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R et al (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10:57–9
8. Chai HT, Jiang HM, Lin L, Liu L (2018) A marginalized two-part beta regression model for microbiome compositional data. PLoS Comput Biol 14:e1006329
9. Cao Y, Lin W, Li H (2018) Large covariance estimation for compositional data via composition-adjusted thresholding. J Am Stat Assoc. https://doi.org/10.1080/01621459.2018.1442340
10. Chen EZ, Li H (2016) A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics 32:2611–2617
11. Chén O, Crainiceanu C, Ogburn E, Caffo B, Wager T, Lindquist M (2018) High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics 19(2):121–136
12. Coffman D, Zhong W (2012) Assessing mediation using marginal structural models in the presence of confounding and moderation. Psychol Methods 17:642–664
13. Egozcue J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Math Geol 35:279–300
14. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J Royal Stat Soc 70:849–911
15. Fritz M, Kenny D, MacKinnon D (2016) The combined effects of measurement error and omitting confounders in the single-mediator model. Multivar Behav Res 51:681–697
16. Gu F, Preacher K, Ferrer E (2014) A state space modeling approach to mediation analysis. J Educ Behav Stat 39:117–143
17. Hron K, Filzmoser P, Thompson K (2012) Linear regression with compositional explanatory variables. J Appl Stat 39:1115–1128
18. Hrůzová K, Todorov V, Hron K, Filzmoser P (2016) Classical and robust orthogonal regression between parts of compositional data. Statistics 50:1261–1275
19. Huang Y, Pan W (2016) Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics 72:402–413
20. Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802
21. Imai K, Keele L, Tingley D (2010) A general approach to causal mediation analysis. Psychol Methods 15:309–334
22. Ismail N, Ragab S, ElBaky A, Shoeib A, Alhosary Y, Fekry D (2011) Frequency of Firmicutes and Bacteroidetes in gut microbiota in obese and normal weight Egyptian children and adults. Arch Med Sci 7:501–507
23. Lin W, Shi P, Feng R, Li H (2014) Variable selection in regression with compositional covariates. Biometrika 101:785–797

24. Liu L, Shih YCT, Strawderman RL, Zhang DW, Johnson B, Chai H (2019) Statistical analysis of zero-inflated continuous data: a review. Stat Sci 34:253–279
25. MacKinnon D, Lockwood C, Hoffman J, West S, Sheets V (2002) A comparison of methods to test mediation and other intervening variable effects. Psychol Methods 7:83–104
26. MacKinnon D, Lockwood C, Williams J (2004) Confidence limits for the indirect effect: distribution of the product and resampling methods. Multivar Behav Res 39:99–128
27. MacKinnon D (2008) Introduction to statistical mediation analysis. Erlbaum and Taylor Francis Group, New York
28. Mandal S, Treuren W, White R, Eggesbø M, Knight R, Peddada S (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial Ecol Health Dis 26(1):27663
29. Martín-Fernández J, Barceló-Vidal C, Pawlowsky-Glahnm V (2003) Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Math Geol 35:253–278
30. Mert M, Filzmoser P, Endel G, Wilbacher I (2018) Compositional data analysis in epidemiology. Stat Methods Med Res 27:1878–1891
31. Morais J, Thomas-Agnan C, Simioni M (2018) Using compositional and Dirichlet models for market share regression. J Appl Stat 45:1670–1689
32. Preacher K, Hayes A (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 40:879–891
33. Preacher K (2015) Advances in mediation analysis: a survey and synthesis of new developments. Annu Rev Psychol 66:825–852
34. Reid S, Tibshirani R, Friedman J (2016) A study of error variance estimation in lasso regression. Stat Sin 26:35–67
35. Sampson J, Boca S, Moore S, Heller R (2018) FWER and FDR control when testing multiple mediators. Bioinformatics 34:2418–2424
36. Sobel M (1982) Asymptotic confidence intervals for indirect effects in structural equation models. Sociol Methodol 13:290–312
37. Sohn M, Li H (2019) Compositional mediation analysis for microbiome studies. Ann Appl Stat 13:661–681
38. Tang Z, Chen G, Alekseyenko A, Li H (2017) A general framework for association analysis of microbial communities on a taxonomic tree. Bioinformatics 33:1278–1285
39. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B 58:267–288
40. Trompette A, Gollwitzer E, Yadava K et al (2014) Gut microbiota metabolism of dietary fiber influences allergic airway disease and hematopoiesis. Nat Med 20:159–166
41. Tsilimigras M, Fodor A (2016) Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol 26:330–335
42. VanderWeele T (2009) Marginal structural models for the estimation of direct and indirect effects. Epidemiology 20:18–26
43. VanderWeele T (2016) Mediation analysis: a practitioner's guide. Annu Rev Public Health 37:17–32
44. Wang T, Zhao H (2017) Constructing predictive microbial signatures at multiple taxonomic levels. J Am Stat Assoc 112:1022–1031
45. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R et al (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334:105–108
46. Yun Y, Kim H, Kim S et al (2017) Comparative analysis of gut microbiota associated with body mass index in a large Korean cohort. BMC Microbiol 17:151
47. Zhang C-H (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38:894–942
48. Zhang C-H, Zhang S (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B 76:217–242
49. Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, Zhang W, Schwartz J, Just A, Colicino E, Vokonas P, Zhao L, Lv J, Baccarelli A, Hou L, Liu L (2016) Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics 32:3150–3154
50. Zhang J, Wei Z, Chen J (2018) A distance-based approach for testing the mediation effect of the human microbiome. Bioinformatics 34:1875–1883

51. Zhao Y, Luo X (2016) Pathway Lasso: estimate and select sparse mediation pathways with high-dimensional mediators.arXiv:1603.07749v1, Preprint
52. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67:301–320