

Microbiome data analysis using semiparametric regression model

Duye Liu and Ao Yuan

Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University,
Washington, DC 20057, USA.

ARTICLE HISTORY

Compiled 20 June 2026

ABSTRACT

In microbiome data analysis, the interest is to infer the relationship among response, covariates and a large number of microbiomes. To model the larger number of microbiomes, we treat them as functional data, and use a semiparametric regression model for their relationship with the response, in which the effects of the covariates are specified as regression coefficients, and that of the microbiome is specified as a functional parameter, via the B-spline, and a logistic link for the outcome. We also infer the correlations among some selected microbiomes. Simulation study is conducted to evaluate the performance of the method, and the model is applied to analyze a real operational taxonomic microbiome data.

KEYWORDS

Microbime, binary outcome, semiparametric regression, B-spline, correlation.

1. Introduction

Recent advancements in bio-medical research indicate that microbime plays a critical role in determining susceptibility to some diseases. It was found that the composition and function of the microbiome is useful for investigating disease prevention and treatment responses (Lynch et al., 2016), and that microbiome is susceptible to chronic diseases, including cardiovascular, gastrointestinal, respiratory, and immune-related conditions (Aldirawi et al., 2023). In microbiome studies, operational taxonomic units (OTUs) are used to measure bacterial counts for each individual which is a long sequence of measurements. Additionally, different bacterial taxa may exhibit correlations. However, the analysis of microbiome data pose significant analytical challenges due to their unique characteristics - they are high-dimensional, compositional, sparse, and exhibit a high proportion of zero-inflated values. These complexities make standard modeling approaches insufficient for capturing the robust statistical relationships between the microbiome compositions and host-specific outcomes.

Given the compositional and high-dimensional nature of microbiome data, many regression models have been proposed to explore how microbiome composition together with other baseline covariates, relates to specific health outcomes, such as disease status or treatment response. One of the popular approaches for high-dimensional covariates is the log-contrast regression model. To accommodate high-dimensional data

with compositional covariates Shi et al.(2016) and Combettes et al.(2020) proposed linear log-contrast models to identify bacterial genera associated with BMI. Hu et al.(2020) proposed linear decomposition model (LDM) for microbiome data analysis, providing a unified framework for testing both global associations and operational taxonomic unit (OTU)-specific effects. Chen et al. (2016) proposed a two-part mixed-effects model combining logistic and beta regressions for testing the association between microbial abundance and clinical covariates for longitudinal microbiome data. This approach effectively handles the excessive zeros and compositional nature of microbiome data, facilitating the identification of microbial taxa associated with clinical outcomes over time; In addition, Lu et al. (2019) introduced generalized linear models (GLMs) with linear constraints on regression coefficients to identifies bacterial species associated with inflammatory bowel disease. While these models are primarily designed for continuous responses, logistic regression methods are frequently employed when the outcome of interest is binary, such as the presence or absence of disease.

Here we focuses on logistic regression analysis for microbiome compositional data to identify microbial taxa associated with binary disease outcomes. Previously, Park et al. (2016) suggested a robust penalized logistic regression based on weighted likelihood methodology; however, this method does not satisfy the zero-sum constraint. To improve this, Monti et al.(2021) proposed a Robust Logistic Zero-Sum Regression(RobLZS) estimator, which is used for binary outcomes with high-dimensional compositional covariates. Xia et al.(2013) introduced a logistic normal multinomial(LNM) regression model which addressed challenges such as over-dispersion, zero observations, and high dimensionality typical of microbiome studies. This approach performed better in capturing the variability in microbiome data compared to other models.

In this paper, we present a semiparametric model to study the association between a binary disease outcome, microbiome data, and other covariates. To analyze this data, we apply parametric regression for covariates, utilize B-splines to model the microbiome data, and employ a logistic regression model to capture the relationship between the outcome and covariates. This model offers greater flexibility and robustness compared to traditional linear models. We also infer the correlations among some selected microbiomes. In the results section, we describe the simulation studies to evaluate the performance of the method. Then the method is applied to analyze a real operational taxonomic microbiome data with binary outcome.

2. The method

The observed longitudinal data is $D_n = \{(y_i, \mathbf{x}_i, \mathbf{m}_i) : i = 1, \dots, n\}$ from n independent individuals, where y_i is the binary indicator of disease status from the i -th individual, $\mathbf{x}_i$ is the covariates of i -th individual, and $\mathbf{m}_i$ is the sequence of microbiome counts of the i -th individual, it has length up to thousands or tens of thousands. For our data, the sample size is moderate, of hundreds. The goal is to infer the association between the response variable Y and covariates $\mathbf{X}$ and microbiome $\mathbf{m}$. For convenience, Hu and Satten (2020) convert the counts into frequencies by dividing the library size.

2.1. Model and estimation

It is more convenient to treat the microbiome data sequence as continuous, as $m_i(t)$ on $[0, T]$ for some $0 < T < \infty$. Consider the following semiparametric logistic regression model, which is a variation of the semiparametric linear model in Chen et al. (2021)

$$P(y_i = 1 \mid \mathbf{x}_i, \mathbf{m}_i) = g\left(\mathbf{x}_i^\top \boldsymbol{\alpha} + \int_0^T m_i(t)\beta(t) dt\right), \quad (i = 1, \dots, n), \quad (1)$$

where g is a given function, typically the logistic function; $\boldsymbol{\alpha}$ is the effects of the $\mathbf{x}_i$'s, and $\beta(\cdot)$ is the effects of the $m_i(\cdot)$'s. Now the goal becomes to estimate the unknown parameters $\boldsymbol{\alpha}$ and the functional parameter $\beta(\cdot)$. There are a number of ways to estimate a functional parameter, common methods including kernel smoothing, spline, reproducing kernel Hilbert space method, and nonparametric maximum likelihood estimate, etc. Each method has its own advantages and disadvantages. We use the B-spline of degree k to model $\beta(\cdot)$, which is an extensively studied method with broad applications. With this method, it needs to divide $[0, T]$ into $r + k - 1$ intervals $[b_j, b_{j+1})$'s, for some chosen number r of knots b_j 's to be addressed. The B-spline has r known basis functions $B_1(\cdot), \dots, B_r(\cdot)$, each is a polynomial with degree k . Typically $k = 3$. Then $\beta(\cdot)$ is well approximated as

$$\beta(t) = \sum_{j=1}^r \beta_j B_j(t), \quad t \in [0, T],$$

where each $B_j(\cdot)$ is a polynomial of degree $\leq k$, is defined on the j -th interval (and is 0 outside this interval), and the β_j 's are the coefficients to be estimated from the observed data. The above representation of $\beta(\cdot)$ is called a B-spline. There are different ways to choose the set of knots and the number r of knots. For random designed t , often the knots are chosen as the $r/(r + k + 1)$ -quantiles of t ($r = 1, \dots, r + k + 1$). For our case of roughly uniform design, we just choose them uniformly over $[0, T]$. For the number of knots, often $r = O(n^{1/3})$ or $o(n^{1/2})$ (Gu, 2002; Huang and Liu, 2006; Sun et al., 2008; Yoshida and Naito, 2012; Yuan et al., 2020). Once $[0, T]$, the number d of knots, and the set of knots are given, the B-spline basis $B_j(\cdot)$'s can be conveniently obtained using the R-package *splines*(), and only the β_j 's are to be estimated. Fan et al. (2013), Chiu et al. (2019) and Chen et al. (2021) used similar models in their genetic analysis with non-longitudinal observations.

Now let $h_{ij} = \int_0^T m_i(t)B_j(t) dt$, $\mathbf{h}_i = (h_{i1}, \dots, h_{ir})^\top$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_r)^\top$. Model (1) is rewritten as

$$P(y_i = 1 \mid \mathbf{x}_i, \mathbf{m}_i) = g(\mathbf{x}_i^\top \boldsymbol{\alpha} + \mathbf{h}_i^\top \boldsymbol{\beta}), \quad (i = 1, \dots, n). \quad (2)$$

Such a semiparametric logistic model is also used in Yuan et al. (2020). The log-likelihood is

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid D_n) = \sum_{i=1}^n \{y_i \log g(\mathbf{x}_i^\top \boldsymbol{\alpha} + \mathbf{h}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log(1 - g(\mathbf{x}_i^\top \boldsymbol{\alpha} + \mathbf{h}_i^\top \boldsymbol{\beta}))\}, \quad (3)$$

and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is estimated by the maximum likelihood estimate (MLE)

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \arg \max_{(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in (A, B)} \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid D_n). \quad (4)$$

2.2. Inference of microbiome correlation

The correlations among the microbiome is a topic of interest, however it is impossible to compute and compare all pairs of microbiomes as there are large numbers of them. Let $t_1, \dots, t_N$ be the points at which the microbiomes are measured, here we assume the locations t_j 's are arranged with some biological meaning. We only need to compare a relatively small number of selected microbiomes. For this we need to choose a relatively small number of locations t_{j_r} 's to compute and compare the correlations.

To investigate locations where the mean microbiome curve $\bar{m}(t)$ and its smoothed version $\hat{m}(t)$ differ the most, we first computed the mean curve $\bar{m}(\cdot)$ and its smoothed version $\hat{m}(\cdot)$ via kernel smoothing with Gaussian weights as

$$\bar{m}(t_j) = \frac{1}{n} \sum_{i=1}^n m_i(t_j), \quad \hat{m}(t_j) = \frac{\sum_{r=1}^N \phi((t_r - t_j)/h) \bar{m}(t_r)}{\sum_{r=1}^N \phi((t_r - t_j)/h)}, \quad (j = 1, \dots, N),$$

where $\phi(\cdot)$ is the density function of $N(0, 1)$ and $h = O(N^{-1/5})$ is a positive constant. Then compute the distance D_j between $\bar{m}(t_j)$ and $\hat{m}(t_j)$ at all points,

$$D_j = |\bar{m}(t_j) - \hat{m}(t_j)|, \quad (j = 1, \dots, N),$$

and for some pre-specified $k = 10 (\ll N)$, select the k largest distances $\{D_{j_r} : t_{j_1}, \dots, t_{j_r}\}$. These $t_{j_1}, \dots, t_{j_r}$ are the locations at which $\bar{m}(t_j)$ and $\hat{m}(t_j)$ differ the most, and are targets of investigation. Then compute the correlations among these pairs of locations (t_r, t_s) , and denote the sample correlation as $\hat{\rho}_{r,s}$. For each pair (t_r, t_s) , we test the null hypothesis $H_{r,s}^0 : \rho_{r,s} = 0$ vs the alternative $H_{r,s}^a : \rho_{r,s} \neq 0$. Since $\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{D} N(0, (1 - \rho^2)^2)$, under $H_{r,s}^0$, $\sqrt{n}\hat{\rho}_{r,s} \xrightarrow{D} N(0, 1)$. So the p -value of $\hat{\rho}_{r,s}$ is $p_{r,s} = 2(1 - \Phi(\sqrt{n}|\hat{\rho}_{r,s}|))$. If $p_{r,s} < 0.05$, $H_{r,s}^0$ is rejected.

3. Simulation study and application

3.1. Simulation study

We will simulate data, then compute the estimations from (4) via model (3). The MLE of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ from the above log-likelihood can be obtained using R-code.

We generate $n = 300, 500, 800, 1000$ independent observations $\{(Y_i, X_i, M_i(t)) : i = 1, \dots, n\}$ on the time domain $[0, 1]$ discretized at $T = 50$ equally spaced points. For each subject, X_i was generated from a multivariate normal distribution with covariance implemented by Cholesky factorization.

Let m_i be a sequence of length 50. To generate microbiome counts m_i , we define a functional model

$$m_{ik} = h(t_k) + \epsilon_{ik}, \quad k \in \{1, \dots, 50\}, \quad i \in \{1, \dots, n\},$$

where $h(t_k)$ is a smooth function of time t_k and ϵ_i is a random error term. Let $h(t) = t/(1 + |t|)$.

We define an arbitrary coefficient vector α , and apply the logistic (sigmoid) link function to the term $\mathbf{x}_i^\top \alpha + \int_0^T m_i(t)\beta(t) dt$. We then generate y_i from a Bernoulli distribution. To compute $h_{ij} = \int_0^T m_i(t)B_j(t) dt$, we generate the B-spline basis using the R-package *splines*. We then apply the logistic link function to the linear term and fit the logistic regression model using `glm()` in R to find the maximum likelihood estimates (MLE) of parameters α and β .

In this simulation, we consider the case where the sample size $N = 300, 500, 800, 1000$. Figure 1 illustrates the comparison between the true and estimated $\beta(t)$.

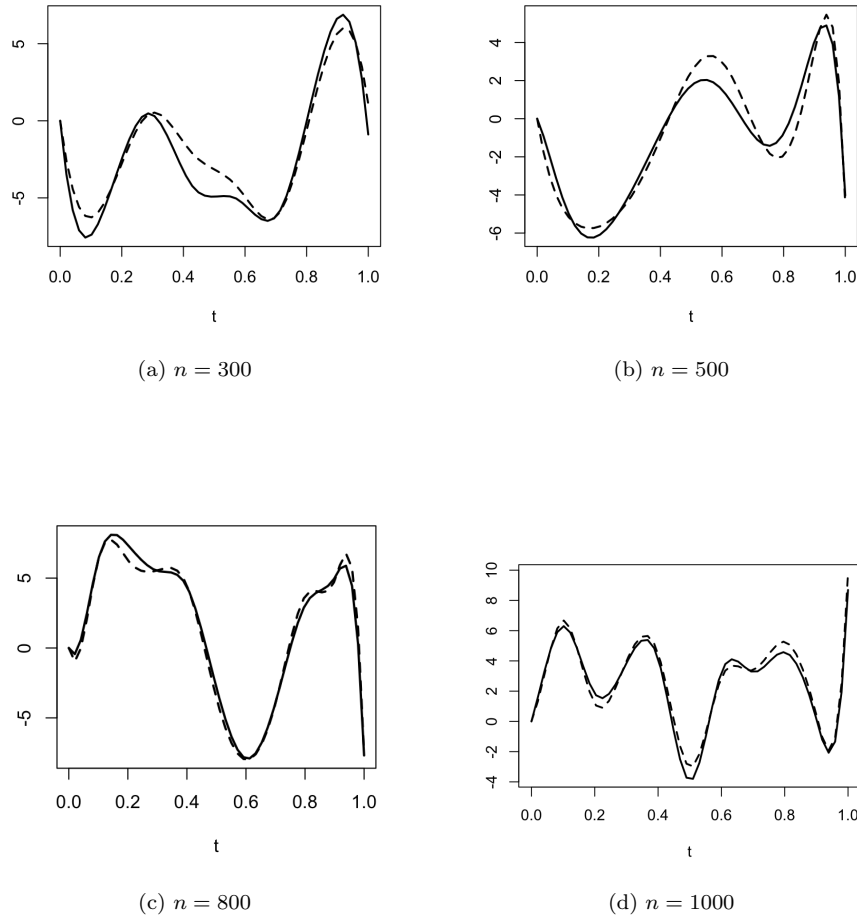


Figure 1. Comparison of true (solid line) and estimated (dashed line) $\beta(t)$ across sample sizes. As the sample size increases, the estimated $\beta(t)$ becomes more accurate and closely matches the true $\beta(t)$, reflecting the improved stability of the estimator.

For sample size $n = 300, 500, 800, 1000$, the corresponding dimension of $\mathbf{x}$ is 3, 5, 8, 10. Table 1 below shows the true and estimated α values and the estimated standard deviation (SD).

Table 1. True values (α), estimates ($\hat{\alpha}$), and standard deviations (SD) for $n = 300, 500, 800, 1000$.

n		α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}
300	α	1.24	4.29	4.36							
	$\hat{\alpha}$	1.05	3.89	4.05							
	SD	0.29	0.62	0.62							
500	α	-4.19	-1.19	-3.85	4.48	-1.26					
	$\hat{\alpha}$	-4.30	-1.17	-3.68	4.83	-1.84					
	SD	0.58	0.26	0.47	0.64	0.31					
800	α	0.95	1.87	3.72	-4.73	-4.92	-4.78	-2.35	4.93		
	$\hat{\alpha}$	0.86	1.60	4.27	-4.69	-4.74	-4.67	-2.62	4.70		
	SD	0.22	0.29	0.55	0.57	0.56	0.59	0.37	0.55		
1000	α	4.05	-4.88	0.50	4.52	-2.66	1.83	0.44	4.75	-3.59	-0.87
	$\hat{\alpha}$	4.00	-5.01	0.69	4.41	-3.09	1.70	0.53	5.20	-3.51	-0.79
	SD	0.42	0.51	0.20	0.45	0.35	0.23	0.17	0.52	0.36	0.18

3.2. Correlation analysis

Table 2. Estimated correlations $\hat{\rho}_{r,s}$ and asymptotic p -values across different sample sizes N .

N	col_r	col_s	$\hat{\rho}_{r,s}$	p_{norm}
300	2	5	0.15	0.01
	3	4	-0.13	0.03
500	3	9	0.11	0.01
	2	7	0.10	0.03
800	4	8	0.10	0.007
	2	3	0.09	0.01
	3	6	0.09	0.02

From Table 2, we observe that for moderate sample sizes ($N = 300, 500, 800$), several pairs of locations exhibit statistically significant correlations, with estimated $\hat{\rho}_{r,s}$ values around 0.10–0.15 and p -values below 0.05. These results suggest that at selected locations where the mean curve and smoothed curve differ most, there is consistent correlation structure across subjects.

As N increases, the identified correlations tend to be smaller in magnitude, and at $N = 1000$ no statistically significant correlations are detected. This indicates that as the sample size grows, the discrepancies between $\bar{m}(t_j)$ and $\hat{m}(t_j)$ become less pronounced.

3.3. Application

We applied the proposed semiparametric logistic regression model to the *dietswitch* microbiome dataset obtained from the *microbiome* R package. The dataset contains operational taxonomic unit (OTU) profiles collected from subjects undergoing diet transition, along with demographic and clinical information. In this analysis, we used sex as the binary response variable, while subject type, BMI group, and nationality were included as covariates.

Prior to analysis, the OTU table was max-normalized, and taxa with zero or constant abundance across all subjects were removed, yielding 222 taxa for downstream inference. The covariates were modeled parametrically, whereas the microbiome abun-

dance profiles were treated as functional data and incorporated via the B-spline representation described in Section 2.

The parameter estimates for the covariates are shown in Table 3. The estimated coefficients indicate that subject type shows some associations with sex.

To investigate microbial interactions, correlations among taxa were estimated and tested. Table 4 summarizes statistically significant correlations. We observed a strong positive correlation between *Prevotella oralis et rel.* and *Prevotella melaninogenica et rel.* ($\hat{\rho} = 0.99, p < 0.001$), suggesting shared ecological response or metabolic similarity. In contrast, *Tannerella et rel.* displayed consistent negative correlations with several *Prevotella* species, indicating potential competitive interactions.

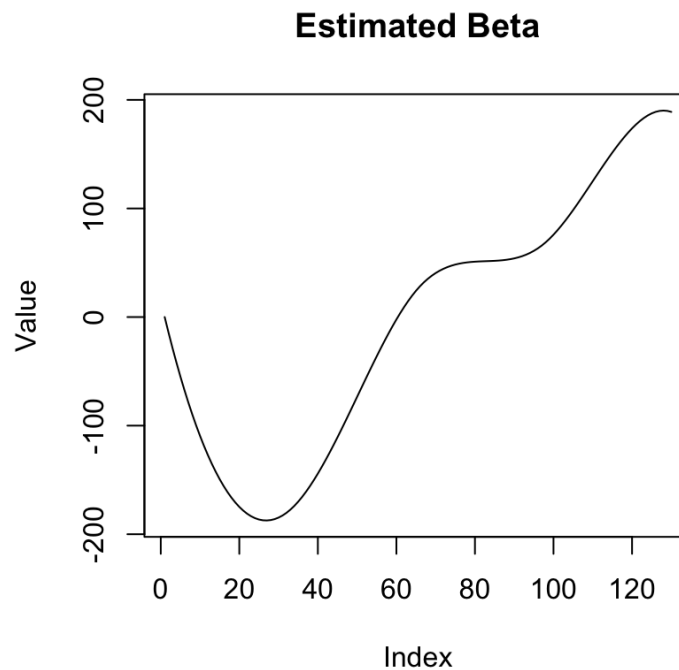


Figure 2. Application results on the *dietswitch* microbiome dataset.

The parameter estimates for the covariates are shown in Table 3 below. The estimated coefficients indicate that subject type shows some associations with sex.

Table 3. Estimated covariate effects, $\hat{\alpha}$, and standard deviations on the *dietswitch* data ($n = 222$).

n		α_1	α_2	α_3
222	$\hat{\alpha}$	-0.02	-1.92	-1.07
	SD	0.02	0.27	0.49

To investigate microbial interactions, correlations among taxa were estimated and tested. Table 4 below summarizes statistically significant correlations. We observed strong positive correlation between *Prevotella oralis et rel.* and *Prevotella melaninogenica et rel.* ($\rho = 0.99, p < 0.001$), suggesting shared ecological response or metabolic

similarity. In contrast, *Tannerella* et rel. displayed consistent negative correlations with several *Prevotella* species, indicating potential competitive interactions.

Table 4. Significant pairwise correlations among selected taxa in the *dietswitch* data.

col_r	col_s	$\hat{\rho}$	p-value
<i>Helicobacter</i>	<i>Campylobacter</i>	0.57	0.00
<i>Prevotella oralis</i> et rel.	<i>Prevotella melaninogenica</i> et rel.	0.99	0.00
<i>Tannerella</i> et rel.	<i>Prevotella melaninogenica</i> et rel.	-0.47	0.00
<i>Tannerella</i> et rel.	<i>Prevotella oralis</i> et rel.	-0.47	0.00
<i>Propionibacterium</i>	<i>Tannerella</i> et rel.	0.36	0.00
<i>Helicobacter</i>	<i>Propionibacterium</i>	0.35	0.00
<i>Clostridium ramosum</i> et rel.	<i>Propionibacterium</i>	0.35	0.00
<i>Helicobacter</i>	<i>Bacillus</i>	0.32	0.00
<i>Campylobacter</i>	<i>Bacillus</i>	0.31	0.00
<i>Campylobacter</i>	<i>Propionibacterium</i>	0.31	0.00
<i>Helicobacter</i>	<i>Megamonas hypermegale</i> et rel.	0.30	0.00
<i>Campylobacter</i>	<i>Clostridium ramosum</i> et rel.	0.24	0.00
<i>Bacillus</i>	<i>Propionibacterium</i>	0.23	0.00
<i>Propionibacterium</i>	<i>Prevotella melaninogenica</i> et rel.	-0.22	0.00
<i>Propionibacterium</i>	<i>Prevotella oralis</i> et rel.	-0.22	0.00
<i>Helicobacter</i>	<i>Clostridium ramosum</i> et rel.	0.21	0.00
<i>Clostridium ramosum</i> et rel.	<i>Tannerella</i> et rel.	0.21	0.00
<i>Vibrio</i>	<i>Tannerella</i> et rel.	-0.17	0.01
<i>Helicobacter</i>	<i>Vibrio</i>	0.17	0.01
<i>Helicobacter</i>	<i>Tannerella</i> et rel.	0.16	0.02
<i>Clostridium ramosum</i> et rel.	<i>Prevotella melaninogenica</i> et rel.	-0.14	0.03
<i>Bacillus</i>	<i>Clostridium ramosum</i> et rel.	0.13	0.05
<i>Clostridium ramosum</i> et rel.	<i>Prevotella oralis</i> et rel.	-0.13	0.05

4. Discussion

We proposed a semiparametric logistic model for the analysis of binary outcome with microbine sequence covariates, which is modeled by a B-spline. We also proposed a method for the selection of locations on the microbine and to compute their correlations. Simulation studies are conducted to evaluate the performance of the proposed method which show it work well for such a high dimensional data. Then the method is applied to analyze a real microbiome data, and revealed some susceptible important relationships among some of the microbiome sites.

References

- [1] Aldirawi, H., Morales, F.G. (2023). Univariate and multivariate statistical analysis of microbiome data: An overview.
- [2] Chen, L., Tian, R., Chen, G., Yuan, A., Li, C.-M., Bently, A.R., Hoffman, H.J., Rotimi, C. (2022). Semiparametric partial linear modeling of risk factors for ear infections: The Early Childhood Longitudinal Study. *Journal of Applied Statistics*, published online 11/07/2022.
- [3] Chen, G., Zhou, L., Yuan, A., Rotimi, C. (2021). Longitudinal GWAS data analysis using semiparametric regression model. Manuscript.
- [4] Chen, E.Z., Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data.
- [5] Combettes, P.L., Müller, C.L. (2020). Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications.
- [6] Fan, R.Z., Wang, Y.F., Mills, J.L., Wilson, A.F., Bailey-Wilson, J.E., Xiong, M.M. (2013). Functional linear models for association analysis of quantitative traits. *Genetic Epidemiology*, **37**, 726–742.
- [7] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

- [8] Hu, Y.-J., Satten, G.A. (2020). Testing hypotheses about the microbiome using the linear decomposition model. Manuscript.
- [9] Huang, J., Liu, L. (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrika*, **62**(3), 269–276.
- [10] Lu, J., Shi, P., Li, H. (2019). Generalized linear models with linear constraints for microbiome compositional data.
- [11] Lynch, S.V., Pedersen, O. (2016). The human intestinal microbiome in health and disease. *New England Journal of Medicine*, **375**, 2369–2379.
- [12] Monti, G.S., Filzmoser, P. (2021). Robust logistic zero-sum regression for microbiome compositional data.
- [13] Park, H., Konishi, S. (2016). Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *Journal of Statistical Computation and Simulation*.
- [14] Qin, J., Garcia, T.P., Ma, Y., Tang, M.-X., Marder, K., Wang, Y. (2014). Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint. *Annals of Applied Statistics*, **8**, 1182–1208.
- [15] Robertson, T., Wright, F.T., Dykstra, R. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore.
- [16] Shi, P., Zhang, A., Li, H. (2016). Regression analysis for microbiome compositional data.
- [17] Sun, J., Kopciuk, K.A., Lu, X. (2008). Polynomial spline estimation of partially linear single-index proportional hazards regression models. *Computational Statistics and Data Analysis*, **53**(1), 176–188.
- [18] Xia, F. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis.
- [19] Yoshida, T., Naito, K. (2012). Asymptotics for penalized additive B-spline regression. *Journal of Japan Statistics*, **42**(1), 81–107.
- [20] Yuan, A., Piao, J., Ning, J., Qin, J. (2021a). Semiparametric isotonic regression model and estimation for group testing data. *Canadian Journal of Statistics*, **49**(3), 659–677.
- [21] Yuan, A., Yin, A., Tan, M.T. (2021b). Enhanced robust procedure for causal inference. *Statistics in Biosciences*, **13**(3), 454–478.