SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Original Research Article

Statistical Methods in Medical Research
2022, Vol. 31(2) 361–371

DOI: 10.1177/09622802211061634
journals.sagepub.com/home/smm
SAGE

Functional response regression model on correlated longitudinal microbiome sequencing data

Bo Chen[1] **and Wei Xu**[1,2]

Abstract

Functional regression has been widely used on longitudinal data, but it is not clear how to apply functional regression to microbiome sequencing data. We propose a novel functional response regression model analyzing correlated longitudinal microbiome sequencing data, which extends the classic functional response regression model only working for independent functional responses. We derive the theory of generalized least squares estimators for predictors' effects when functional responses are correlated, and develop a data transformation technique to solve the computational challenge for analyzing correlated functional response data using existing functional regression method. We show by extensive simulations that our proposed method provides unbiased estimations for predictors' effect, and our model has accurate type I error and power performance for correlated functional response data, compared with classic functional response regression model. Finally we implement our method to a real infant gut microbiome study to evaluate the relationship of clinical factors to predominant taxa along time.

Keywords

Functional data analysis, functional response regression, human microbiome, longitudinal measures, generalized least squares estimation

1. Introduction

Microbiome is inherently dynamic in nature, attributing to the presence of interactions among microbes, microbes and the host, and with the environment. Researchers have shown that the microbiome can be altered over time, either transiently or long term, by infections or medical interventions such as antibiotics[1–3]. Recent advances in high-throughput experimental technologies are enabling researchers to measure dynamic behaviors of the microbiota at a large scale[4–6].

Comprehensive analyses of the microbiota over time provide insights into essential questions about microbiome dynamics, for example, how microbiome composition changes through infection/antibiotics and do changes in the microbiome cause or increase susceptibility and risk of certain diseases. Longitudinal data provides more information than single time point data because temporal information creates an inherent ordering in microbiome samples, and thereby they exhibit statistical dependencies that are a function of time[7–9]. These features enable discovery of rich information about microbiome data, including short and long-term trends. Therefore, it is imperative to analyse longitudinal microbiome studies for risk prediction. However, one of the major challenge with longitudinal microbiome data is the presence of uneven number of timepoints along the longitudinal timeline of different subjects[10], making it necessary for the use of appropriate computational techniques to address this issue.

[1]Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada;
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Corresponding author:
Wei Xu, 10-511, 610 University Avenue, Toronto, Ontario, M5G 2M9, Canada.
Email: wei.xu@uhnresearch.ca

To investigate factors associated with longitudinal microbiome composition, functional regression can be implemented, which considers longitudinal microbiome data as a continuous function for each subject. Functional regression is a well-developed method which has been used to model longitudinal data in different contexts. Morris[11] gave a comprehensive review on functional regression. There are a few reasons for choosing functional regression on longitudinal microbiome data. First, by modeling longitudinal data as continuous functions, uneven number of timepoints becomes not a problem. Next, depending on the research question, it may be more intuitive to consider microbiome data as function of time rather than discrete samples at single timepoints, so that the change patterns of microbiome dynamics can be illustrated by functional estimations. Last but not least, if there are large number of timepoints being observed, the predictor space can be at a very high dimension, where the traditional regression methods may become infeasible; in functional regression, the large number of timepoints becomes beneficial because it helps improving the estimation accuracy of the functional microbiome data for each subject.

However, to our best knowledge, the functional regression approach has not been implemented on microbiome sequencing data so far. Microbiome composition, which are usually quantified by operational taxonomic units (OTUs), may exhibit correlations between multiple OTUs[12]. The primary challenge in functional regression is to measure between-function OTU correlations. When considering the timepoints as discrete rather than functional, several different methods have been proposed in current literature to model multiple correlated OTUs. Briefly speaking, these methods can be categorized in three types. The most commonly used method is mixed effects model, which adds correlations of dependent variables by random effects[13–17]. Secondly, Dirichlet multinomial (DM) distribution and its extensions have been used to model the multivariate OTU data[18–22]. Lastly, the OTU correlations can be directly modeled by Generalized Estimating Equation (GEE) approach[6].

There are three types of functional regression models in general: scalar-on-function, function-on-scalar and function-on-function. For analyzing longitudinal microbiome data, we focus on the second type, where longitudinal microbiome data is modeled as functional response and predictors are time-invariant scalars. There has been only limited methodology developments in current literature considering for correlated functional responses when performing function-on-scalar regression. Functional mixed effects model is a common solution[23–26], but similar to the classic mixed effects model on scalar responses, the curve level random effects may only induce a positive correlation, while the OTU correlations can be both positive and negative[12]. Thus, the functional mixed effects model may not be appropriate for microbiome composition data.

In this paper, we focus on developing a novel functional regression model with correlated functional responses. Instead of using random effects to account for OTU correlations, correlation structure between multiple OTUs is constructed allowing for both positive and negative correlations, and accordingly we can model the correlated functional responses by generalized least squares estimations. In Section 2, We present the theoretical estimation of predictors' functional effects when functional responses are correlated. Based on our developed theory, we then propose a data transformation method on both predictors and functional responses data, so that our model can be implemented computationally effective in practice. In Section 3, we check the unbiasedness of predictors' effects estimation and statistical testing accuracy of our proposed model by simulation studies, and compare it with classic functional response regression model assuming independent functional responses. In Section 4, we apply our model to a real microbiome sequencing data with longitudinal measures. We finally discuss the limitations and further extensions of our method in Section 5.

2. Methodology

2.1 Functional regression theory overview

In functional data analysis, the functional data needs to be represented by linear combination of a finite number of known independent basis functions. The most commonly used basis functions are B-splines, Fourier series, principle components and wavelets[11]. Different functional regression methods were proposed in existing literature with each type of these basis functions[27–29,24]. In this paper, we focus on extending the classic functional response regression model using B-spline basis introduced by Ramsay and Silverman[27] to correlated functional response data. Additional works are required for modeling correlated functional responses under other basis representations.

The classic functional response regression assumes independent functional responses and estimates predictors' effects by ordinary least squares estimations[27]. As the functional microbiome data may be correlated, we extend the classic estimation framework to generalized least squares estimations with a correlation matrix added in estimating equations representing OTU correlations. The idea of using generalized least squares estimations for correlated functional data has been implemented to estimate within-function correlations (correlations between timepoints)[30]. In Section 2.2, we use the similar idea but propose a novel correlated functional response regression model which estimates the predictors' effects in theory after accounting for the between-function OTU correlations.

2.2 Correlated functional response regression model

Suppose the OTU data consists of N samples and K OTUs. Each OTU of each sample is a continuous function of time. Let $\boldsymbol{y}(t)$ represents the collection of all functional OTU data. Then $\boldsymbol{y}(t)$ is a vector of length N_K where N_K denotes the product of N and K, and each of its element $y_i(t)$ is a single OTU function of time t for $i = 1, \ldots, N_K$. Let $\boldsymbol{X}$ be an $N_K \times q$ design matrix representing $q - 1$ predictors which are not functional. We assume the following functional response regression model:

$$\boldsymbol{y}(t) = \boldsymbol{X\beta}(t) + \boldsymbol{\epsilon}(t)$$

$\boldsymbol{\epsilon}(t)$ follows multivariate normal distribution, assuming the relative abundances (RAs) observations of OTU data follow log-normal distribution.

The functional data $\boldsymbol{y}(t)$ and $\boldsymbol{\beta}(t)$ are represented by basis functions:

$$\boldsymbol{y}(t) = \boldsymbol{C\phi}(t), \boldsymbol{\beta}(t) = \boldsymbol{B\theta}(t)$$

$\boldsymbol{\phi}(t)$ and $\boldsymbol{\theta}(t)$ are prespecified B-spline basis functions of length M_y and M_β. $\boldsymbol{C}$ and $\boldsymbol{B}$ are $N_K \times M_y$ and $q \times M_\beta$ coefficient matrices. $\boldsymbol{B}$ is unknown and needs to be estimated. Our target is to estimate $\boldsymbol{\beta}(t)$ via finding the generalized least squares estimation of $\boldsymbol{B}$. Differing from the classic functional response regression model, K OTUs may be correlated. The OTU correlations are measured in $\boldsymbol{W}$, which is the correlation matrix of $\boldsymbol{\epsilon}(t)$. We note that $\boldsymbol{W}$ may vary along time, but we assume time invariant $\boldsymbol{W}$ for simplicity in our theoretical work below. Estimating $\boldsymbol{B}$ with functional $\boldsymbol{W}(t)$ is theoretically more challenging and requires future investigation. After including $\boldsymbol{W}$, the generalized least squares is

$$\int [\boldsymbol{y}(t) - \boldsymbol{X\beta}(t)]' \boldsymbol{W} [\boldsymbol{y}(t) - \boldsymbol{X\beta}(t)] dt$$

Regularization of basis functions is the key idea for global smoothing in functional data analysis. B-spline basis functions are usually smoothed by roughness penalties[11]. Similar to the classic functional response regression framework[27], we use a linear differential operator $\mathcal{L}$ to define a roughness penalty for $\boldsymbol{\beta}$:

$$\lambda \int [\mathcal{L}\boldsymbol{\beta}(t)]' [\mathcal{L}\boldsymbol{\beta}(t)] dt$$

where λ is a smoothing parameter that measures the rate of smoothness of the fit.

In contrast to the regression spline smoothing which depends on the number of basis functions selected, spline smoothing by roughness penalties fix the number of basis to be $N_K + 2$ using order four B-splines, and choosing the degree of roughness by λ is equivalent to choosing the number of basis in functional models without a penalty term. The generalized cross-validation or GCV criterion is often used to select an appropriate smoothing parameter value, by finding the smoothing parameter that minimizes GCV. We adopt this approach to choose λ in our simulation and application study.

To estimate $\boldsymbol{B}$, we are trying to minimize the penalized least squares, which combines the generalized least squares with the roughness penalty. The penalized least squares with basis representations of $\boldsymbol{y}(t)$ and $\boldsymbol{\beta}(t)$ is

$$PLS(\boldsymbol{y}(t)|\boldsymbol{\beta}(t)) = \int [\boldsymbol{C\phi}(t) - \boldsymbol{XB\theta}(t)]' \boldsymbol{W} [\boldsymbol{C\phi}(t) - \boldsymbol{XB\theta}(t)] dt + \lambda \int [\mathcal{L}\boldsymbol{B\theta}(t)]' [\mathcal{L}\boldsymbol{B\theta}(t)] dt$$

For notation simplicity, we define the following four matrices:

$$\boldsymbol{J}_{\phi\phi} = \int \boldsymbol{\phi}(t)\boldsymbol{\phi}'(t) dt$$

$$\boldsymbol{J}_{\theta\theta} = \int \boldsymbol{\theta}(t)\boldsymbol{\theta}'(t) dt$$

$$\boldsymbol{J}_{\phi\theta} = \int \boldsymbol{\phi}(t)\boldsymbol{\theta}'(t) dt$$

$$\boldsymbol{R} = \int [\mathcal{L}\boldsymbol{\theta}(t)][\mathcal{L}\boldsymbol{\theta}(t)]' dt$$

Next we re-express each component in $PLS(\boldsymbol{y}(t)|\boldsymbol{\beta}(t))$ by its trace. Note that each component is a scalar, and the trace of a scalar is simply itself. With some matrix algebra we achieve

$$\int \boldsymbol{\phi}'(t)\boldsymbol{C}'\boldsymbol{W}\boldsymbol{C}\boldsymbol{\phi}(t) = tr(\boldsymbol{C}'\boldsymbol{W}\boldsymbol{C}\boldsymbol{J}_{\boldsymbol{\phi\phi}})$$

$$\int \boldsymbol{\theta}'(t)\boldsymbol{B}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{B}\boldsymbol{\theta}(t) = tr(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{B}\boldsymbol{J}_{\boldsymbol{\theta\theta}}\boldsymbol{B}')$$

$$\int \boldsymbol{\phi}'(t)\boldsymbol{C}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{B}\boldsymbol{\theta}(t) = tr(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{C}\boldsymbol{J}_{\boldsymbol{\phi\theta}}\boldsymbol{B}')$$

$$\int [\mathcal{L}\boldsymbol{B}\boldsymbol{\theta}(t)]'[\mathcal{L}\boldsymbol{B}\boldsymbol{\theta}(t)] = tr(\boldsymbol{B}\boldsymbol{R}\boldsymbol{B}')$$

The penalized least squares then becomes

$$PLS(\boldsymbol{C}|\boldsymbol{B}) = tr(\boldsymbol{C}'\boldsymbol{W}\boldsymbol{C}\boldsymbol{J}_{\boldsymbol{\phi\phi}}) + tr(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{B}\boldsymbol{J}_{\boldsymbol{\theta\theta}}\boldsymbol{B}') - 2tr(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{C}\boldsymbol{J}_{\boldsymbol{\phi\theta}}\boldsymbol{B}') + \lambda tr(\boldsymbol{B}\boldsymbol{R}\boldsymbol{B}')$$

Taking derivative with respect to $\boldsymbol{B}$ and setting the result to 0, we find the generalized least squares estimate $\hat{\boldsymbol{B}}$ satisfies

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{B}}\boldsymbol{J}_{\boldsymbol{\theta\theta}} + \lambda\hat{\boldsymbol{B}}\boldsymbol{R} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{C}\boldsymbol{J}_{\boldsymbol{\phi\theta}}$$

$\hat{\boldsymbol{B}}$ can be expressed explicitly in conventional matrix algebra if we use Kronecker products. Let $vec(\hat{\boldsymbol{B}})$ indicates the vector obtained by writing matrix $\hat{\boldsymbol{B}}$ as a vector column-wise. We can rewrite the above equation as

$$[\boldsymbol{J}_{\boldsymbol{\theta\theta}} \otimes (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}) + \boldsymbol{R} \otimes \lambda\boldsymbol{I}]vec(\hat{\boldsymbol{B}}) = vec(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{C}\boldsymbol{J}_{\boldsymbol{\phi\theta}})$$

where $\otimes$ denotes the Kronecker product. So $\hat{\boldsymbol{B}}$ is solved by

$$vec(\hat{\boldsymbol{B}}) = [\boldsymbol{J}_{\boldsymbol{\theta\theta}} \otimes (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}) + \boldsymbol{R} \otimes \lambda\boldsymbol{I}]^{-1}vec(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{C}\boldsymbol{J}_{\boldsymbol{\phi\theta}})$$

Multiplying $\hat{\boldsymbol{B}}$ to the prespecified basis function $\boldsymbol{\theta}(t)$ provides a theoretical estimation of $\boldsymbol{\beta}(t)$ assuming correlated functional responses with correlation matrix $\boldsymbol{W}$.

2.3 Eliminating correlation by Cholesky decomposition

When the correlation matrix $\boldsymbol{W}$ is an identity matrix, the generalized least squares estimation reduces to the ordinary least squares estimation in classic functional response regression model, where statistical softwares, such as the `fda` package in R can be used. However, for correlated functional response data, despite the theoretical derivation in Section 2.2, the generalized least squares estimation may remain a computational challenge to most researchers without a statistical software. To fill this gap, we propose a data transformation technique to eliminate $\boldsymbol{W}$ in the estimation equation, so that existing functional data analysis softwares can be applied directly on the correlated functional response data.

We apply Cholesky decomposition to $\boldsymbol{W}$, such that $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}$ is a lower triangular matrix. Suppose we have another functional response data $\boldsymbol{y}^*(t) = \boldsymbol{L}'\boldsymbol{y}(t)$ and another design matrix $\boldsymbol{X}^* = \boldsymbol{L}'\boldsymbol{X}$, where $\boldsymbol{y}^*(t)$ are independent functional samples. Let the coefficient matrix $\boldsymbol{C}^* = \boldsymbol{L}'\boldsymbol{C}$. The prespecified basis functions $\boldsymbol{\phi}(t)$, $\boldsymbol{\theta}(t)$ and smoothing parameter λ remain the same, so $\boldsymbol{y}^*(t) = \boldsymbol{L}'\boldsymbol{C}\boldsymbol{\phi}(t) = \boldsymbol{C}^*\boldsymbol{\phi}(t)$. Therefore, the penalized least squares estimate $\tilde{\boldsymbol{B}}$ under classic functional response regression model is[27]

$$vec(\tilde{\boldsymbol{B}}) = [\boldsymbol{J}_{\boldsymbol{\theta\theta}} \otimes (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*) + \boldsymbol{R} \otimes \lambda\boldsymbol{I}]^{-1}vec(\boldsymbol{X}^{*\prime}\boldsymbol{C}^*\boldsymbol{J}_{\boldsymbol{\phi\theta}})$$

Following $\boldsymbol{X}^* = \boldsymbol{L}'\boldsymbol{X}$, $\boldsymbol{C}^* = \boldsymbol{L}'\boldsymbol{C}$ and $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{L}'$, it is straightforward to show

$$vec(\tilde{\boldsymbol{B}}) = vec(\hat{\boldsymbol{B}})$$

It implies that if we apply the transformation matrix $\boldsymbol{L}'$ on both $\boldsymbol{y}(t)$ and $\boldsymbol{X}$ and run a classic functional response regression model assuming independence, we will achieve exact same coefficient estimation of $\boldsymbol{B}$. This notably simplifies the computational challenge caused by correlated functional responses, because we can simply find the equivalent independent functional responses using data transformation technique, and the correlated functional response regression question reduces to a classic functional response regression question which can be implemented by existing softwares.

Although our proposed transformation method does not directly estimate $\hat{\boldsymbol{B}}$ using the estimating equation of $vec(\hat{\boldsymbol{B}})$ in Section 2.2, the theoretical work for estimating $\hat{\boldsymbol{B}}$ with correlated functional responses in Section 2.2 is the foundation of our method. Firstly, our transformation method relies on the knowledge of $\hat{\boldsymbol{B}}$. Without the derivation of $\hat{\boldsymbol{B}}$ in Section 2.2, we could still find $vec(\tilde{\boldsymbol{B}})$, but we could not show $vec(\tilde{\boldsymbol{B}}) = vec(\hat{\boldsymbol{B}})$ and thus justify the proposed transformation method achieves same coefficient estimation as directly estimating $\boldsymbol{B}$ following Section 2.2. Secondly, direct use of the estimating equation of $vec(\hat{\boldsymbol{B}})$ is also possible, although not as convenient as applying existing softwares to transformed data.

2.4 Estimating correlation matrix

We showed in Section 2.2 that the estimating equation for correlated functional response regression model depends on correlation matrix $\boldsymbol{W}$. However, in practice, the true $\boldsymbol{W}$ is usually unknown, and it needs to be estimated prior to estimating $\boldsymbol{B}$. For microbiome composition data, there may exist a specific correlation structure depending on the taxonomic structure of multiple OTUs[6]. Rather than using the naive approach assuming unstructured correlation estimation of $\boldsymbol{W}$, we adopt the Generalized Estimating Equation (GEE) approach[6] and the true $\boldsymbol{W}$ is estimated by $\hat{\boldsymbol{W}}$ according to the specific taxonomic structure of OTUs. redThis is a two-step estimation approach: in step 1, the non-functional parameter $\boldsymbol{W}$ is estimated under GEE model using iterative procedures at each timepoint, where both $\boldsymbol{W}$ and $\boldsymbol{\beta}$ are unknown. In step 2, the GEE estimator $\hat{\boldsymbol{W}}$ from step 1 is used to estimate $\boldsymbol{B}$ in the functional settings. Simultaneous estimation of both $\boldsymbol{W}$ and $\boldsymbol{B}$ under functional regression model requires further theoretical investigation.

The GEE approach may provide different estimations of $\boldsymbol{W}$ at different time t, and the time invariant estimator $\hat{\boldsymbol{W}}$ can be computed as the mean of $\hat{\boldsymbol{W}}(t)$ across the entire time interval. In practice, there may be only finite samples collected at a number of timepoints, and $\boldsymbol{W}$ may be estimated separately at each timepoint whenever samples are collected. The overall $\hat{\boldsymbol{W}}$ is then computed as the average of $\hat{\boldsymbol{W}}(t)$ at each timepoint.

In order to achieve the unbiased estimation, it needs to be noted that $\hat{\boldsymbol{W}}$ should not be estimated from the raw data $\boldsymbol{y}(t)$. Instead, $\hat{\boldsymbol{W}}$ must capture the correlation structure of residuals, which is $\boldsymbol{y}(t) - \boldsymbol{X\beta}(t)$. For this reason, we use the same design matrix $\boldsymbol{X}$ to fit the GEE model at each timepoint and estimate the residual correlations correspondingly. Appendix A of the supplementary materials shows that unbiased estimation of $\boldsymbol{W}$ can be achieved by estimating residual correlations after fitting $\boldsymbol{X}$ by GEE model regardless of true functional effect $\boldsymbol{\beta}(t)$. However, if the raw data $\boldsymbol{y}(t)$ are incorrectly used, the estimated $\hat{\boldsymbol{W}}$ can exhibit a significant bias from $\boldsymbol{W}$.

The resulting estimator $\hat{\hat{\boldsymbol{B}}}$ relying on $\hat{\boldsymbol{W}}$ is technically known as feasible generalized least squares. Unlike $\hat{\boldsymbol{B}}$, it may be less clear to evaluate the properties of $\hat{\hat{\boldsymbol{B}}}$ analytically. Alternatively, we use simulation studies in Section 3 to evaluate the unbiasedness of $\boldsymbol{\beta}(t)$ estimations when $\boldsymbol{W}$ is estimated by $\hat{\boldsymbol{W}}$.

3. Simulation

Firstly, simulation studies are designed to evaluate the unbiasedness of $\boldsymbol{\beta}(t)$ estimations. Besides, the accuracy of type I error and power performance for testing $\boldsymbol{\beta}(t)$ also need to be evaluated by simulation. There were only very limited theoretical work discussing statistical testing methods for the global predictor effect $\boldsymbol{\beta}(t)$ under functional response regression model. Zhang[31] showed that the test statistic followed an F-distribution under null hypothesis, but the degrees of freedom estimations are not trivial. Without an existing package, it is not easy to implement that theoretical work in our simulation studies, and we choose to use permutation tests instead, which can be conducted using R package `fda`. For both $\boldsymbol{\beta}(t)$ estimations and hypothesis testing, we also compare our proposed model to the classic functional response regression model which does not consider OTU correlations.

In our simulation settings, we generate a dataset with sample size $N = 100$ at 10 timepoints. For each sample, we assume that three OTUs are from the same taxon and correlated with each other, and specify the exchangeable correlation structure to represent their taxonomic structure. We assume the OTU RAs follow log-normal distribution. Then we simulate the log-transformed OTU RAs as a function of two covariates $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, where $\boldsymbol{x_1}$ is categorical and $\boldsymbol{x_2}$ is continuous.

Although the OTU correlations are assumed to be time invariant in our model, it may not be always true in practice. Thus, in addition to specifying a constant correlation (0.3 and -0.3) in simulation settings, we also specify the true correlations to be unequal at 10 timepoints, where $Cor(t) = 0.05 \times t$ for $t = 1, \ldots, 10$. The OTU correlations are assumed to be unknown in our model, and they are estimated by GEE[6] under the prespecified taxonomic structure. Taking unequal correlations along time into consideration, we estimate correlation at each timepoint separately by GEE, and the final estimation $\hat{\boldsymbol{W}}$ are computed as the average of correlation estimations from 10 timepoints.

To check the unbiasedness of $\boldsymbol{\beta}(t)$ estimation, we specify $\beta_1(t) = \sqrt{t} \times 0.05$ and $\beta_2(t) = \sin(t) \times 0.05$ as the true functional effects of $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$. We then apply Cholesky decomposition to $\hat{\boldsymbol{W}}$ so that $\boldsymbol{y}(t)$, $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are transformed correspondingly. Lastly, we run the `fda` package in R on the transformed data. The estimation of $\beta_1(t)$ and $\beta_2(t)$ are based on the

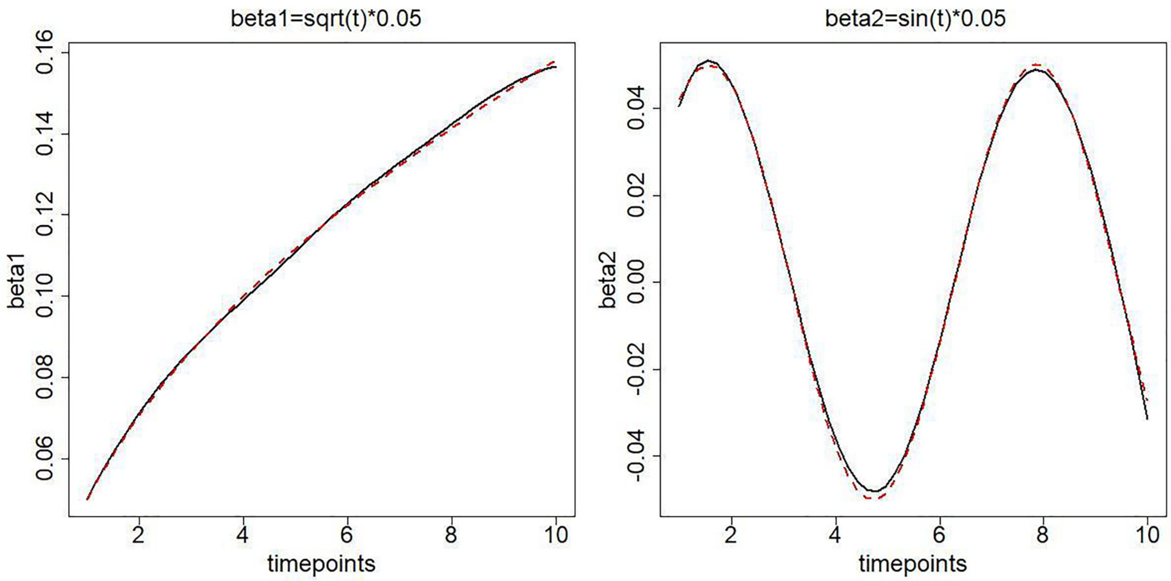

Figure 1. $\beta_1(t)$ and $\beta_2(t)$ estimation based on 1000 replications. Black solid curves are estimated values; red dash curves are true values.

average from 1000 replications. The true values and estimations are plotted in Figure 1 assuming true OTU correlation equal to 0.3. It shows that our proposed correlated functional response regression model provides unbiased estimation for $\boldsymbol{\beta}(t)$ estimation. When true OTU correlation is -0.3 or time variant, results are similar to Figure 1 and not shown.

Next, we check the type I error for testing $\boldsymbol{\beta}(t)$ by permutation test `Fperm.fd` in `fda` package. `Fperm.fd` only allows the covariate to be categorical, so we may only test the effect of $\boldsymbol{x_1}$. Although β_2 may not be tested because $\boldsymbol{x_2}$ is continuous, it may still be included in the model. Formally, we have the following null hypothesis:

$$H_0 : \beta_1(t) = 0 \ \forall t$$

For comparison, we also check type I error from the classic functional response regression model which assumes no OTU correlation. All type I errors are summarized in Table 1. Type I errors are estimated based on 10000 simulation replications with true $\alpha = 0.05$.

Table 1 shows that testing $\beta_1(t) = 0$ by our method provides accurate type I error when OTU correlations are constantly 0.3 or -0.3. When the true correlation is time variant, type I error may be slightly inflated, because the true unequal correlations are replaced by a constant correlation estimation in our model. On the other side, the classic functional response regression models provide inaccurate type I errors, which are significantly inflated (0.2053 and 0.1981) or deflated (0.0002) depending on the OTU correlations being positive or negative. Compared to the classic functional response regression model which incorrectly assumes OTUs are independent, the accurate type I error estimation indicates that p-values and test power estimations based on our model are much more reliable, and the small type I error inflation when the true correlation is time variant may be acceptable.

It needs to be noted that the permutation test adjustment cannot achieve accurate type I errors under classic functional response model when OTU correlations are present. The motivation behind permutation test is to adjust for the timepoints correlation rather than correlation between functional responses. Because the correlation between continuous timepoints is unknown, analytical form of the test statistics may not be available. With permutation test adjustment, type I errors are accurate if the functional responses are independent, regardless of the correlation between any timepoints. In our

Table 1. Comparison of type I error performance based on 10000 replications when the true OTU correlation is constantly 0.3, -0.3 or unequal (ranging from 0.05 to 0.5) at 10 timepoints, $\alpha = 0.05$.

Regression model	Correlation=0.3	Correlation=-0.3	Unequal correlations
Correlated functional response	0.0567	0.0540	0.0715
Classic functional response	0.2053	0.0002	0.1981

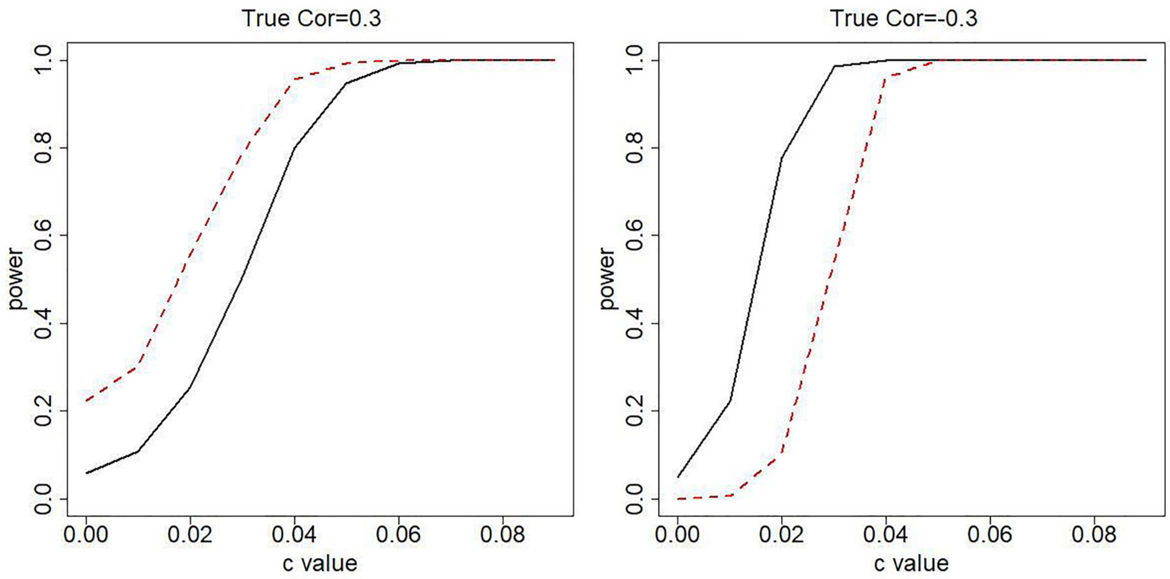

Figure 2. Power estimation for testing $\beta_1(t) = 0$ based on 1000 replications. Black solid curves represents powers under correlated functional response regression model; red dash curves represents powers under classic functional response regression model. *c* value, which represents the strength of the predictor effect, ranges from 0 to 0.09.

simulation, we apply permutation tests to both the classic functional response and correlated functional response model. As shown in Table 1, the classic functional response model can still have inaccurate type I errors due to OTU correlations. The OTU correlations need to be calibrated by Cholesky decomposition using our correlated functional response regression model.

Finally, we evaluate the power performance for testing $\beta_1(t) = 0$. We specify the true value as $\beta_1(t) = \sqrt{t} \times c$, where c value ranging from 0 to 0.09 represents the strength of the predictor effect. We first estimate test powers under our correlated functional response regression model. For comparison, we also evaluate test powers under classic functional response regression model. All powers are estimated based on 1000 replications and summarized in Figure 2 as a function of c value. True OTU correlations are set to 0.3 and -0.3, where the type I error estimation is accurate under correlated functional response regression model as shown in Table 1.

When type I errors are accurate, power estimations are also expected to be accurate under our correlated functional response regression model (black solid curves). Figure 2 further shows that the power performance under classic functional response regression model (red dash curves) departs from our model. The power difference can be dramatic, for instance, 0.777 vs. 0.104 when correlation is -0.3 and $\beta_1(t) = \sqrt{t} \times 0.02$, which indicates a huge power loss by using the classic functional response regression model. We suggest not using classic functional response regression model with correlated functional response data, as the test results can be totally misleading. We further show this point by an application study in Section 4.

4. Application

We illustrate our method by implementing it into a premature infant gut microbiome study[32]. There are 922 specimens from 58 infants with multiple specimens sequenced at different postconceptional ages for each infant, and three predominant taxa are identified, which are Bacilli, Clostridia and Gammaproteobacteria. The relationship of clinical factors to

Table 2. P-values for testing the association between three predominant taxa and four clinical factors: mode of birth (C-section), period of study - sampled after 01/01/2011 or not (Period), breast milk volume (Milk) and days of antibiotics (Antibiotics).

Regression model	C-section	Period	Milk	Antibiotics
Correlated functional response	**0.005**	0.295	**0.005**	0.745
Classic functional response	**0.025**	0.535	0.085	0.910

predominant taxa were evaluated using mixed model regression treating the longitudinal observations of three predominant taxa as repeated measures in their study. In contrast, we model the longitudinal observations as function of postconceptional ages and analyze three predominant taxa together after considering their correlations.

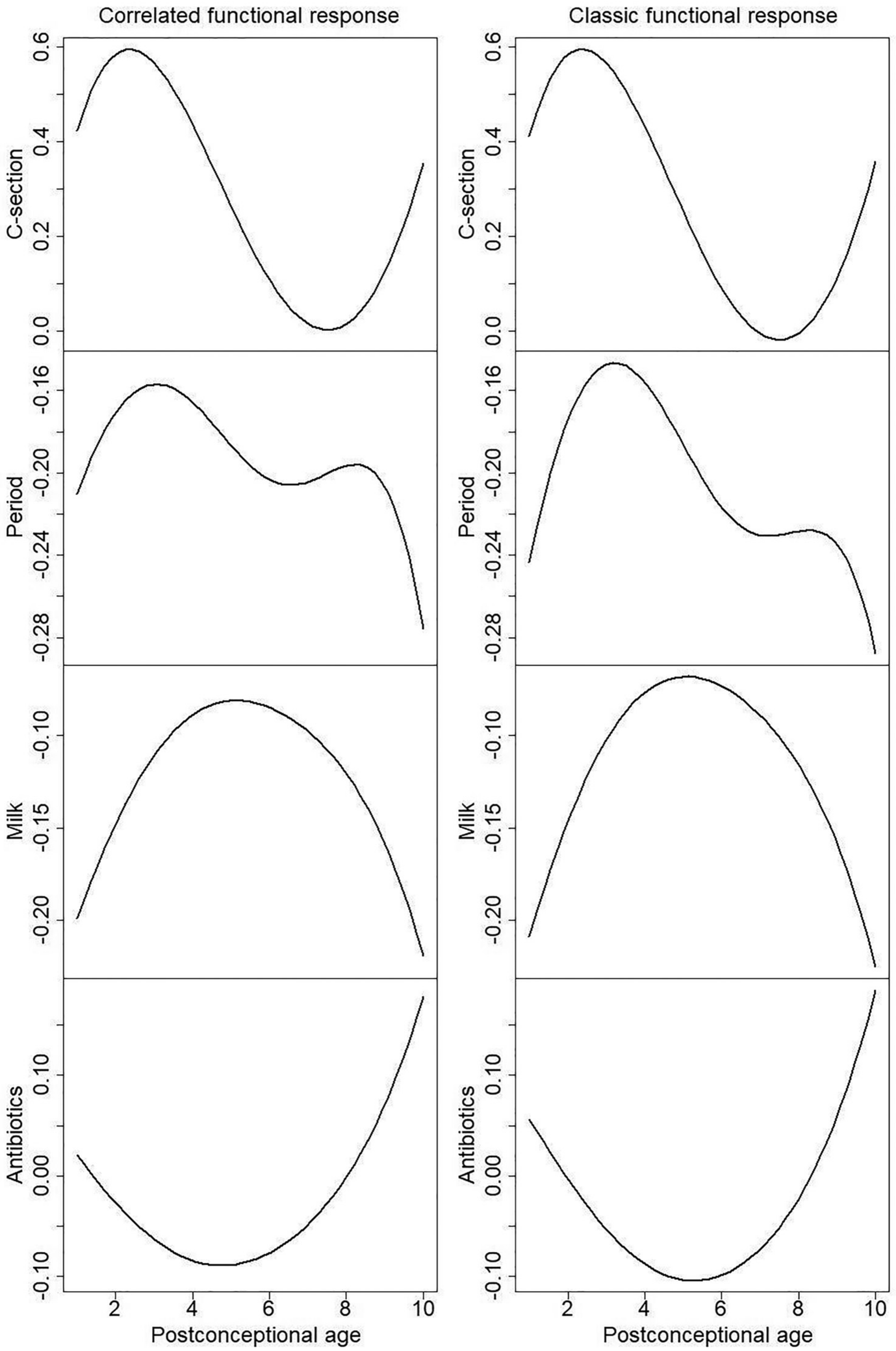

Figure 3. Effects of four clinical factors: mode of birth (C-section), period of study - sampled after 01/01/2011 or not (Period), breast milk volume (Milk) and days of antibiotics (Antibiotics) for predicting all three predominant taxa under correlated functional response regression model (left) and classic functional response regression model (right).

We note that the postconceptional age measurements for each infant are not balanced as the number of measurements may be different. In addition, each infant sample may have different starting and ending ages. For better illustration, we shift and scale the postconceptional ages of each sample to make all postconceptional ages on the same scale from 1 to 10. The converted data is then applied to our functional response regression model. Residual plots after fitting our model are presented in Appendix B of the supplementary materials for model diagnosis.

The correlations between taxa are unknown and we use GEE method described in Section 3 to estimate the correlation matrix $\boldsymbol{W}$:

$$\hat{\boldsymbol{W}} = \begin{pmatrix} 1 & -0.101 & -0.376 \\ -0.101 & 1 & -0.228 \\ -0.376 & -0.228 & 1 \end{pmatrix}$$

We find $\boldsymbol{L}$ as the Cholesky decomposition of $\hat{\boldsymbol{W}}$. The clinical predictors and three predominant taxa modeled as the functional responses are transformed by $\boldsymbol{L}$. We then estimate and test the effects of clinical predictors, including mode of birth, period of study, breast milk volume and days of antibiotics for predicting the three predominant taxa. Days of antibiotics is a continuous measurement and we convert it to binary ($>$ or $\leq$ its median) in order to perform the permutation test. Results for estimating predictors' effects are shown in Figure 3. Estimations under the classic functional response regression are also shown for comparison, and we find that both estimations have very similar patterns, indicating that both models can provide unbiased estimations of $\boldsymbol{\beta}(t)$.

Simulation results from Section 3 suggests that the classic functional response regression model assuming no correlation among taxa may have deflated type I error, given the three predominant taxa are negatively correlated. To confirm this, we show p-values under both our correlated functional response regression model and classic functional response regression model in Table 2. Due to the deflated type I error, we observe that the p-values under classic functional response regression model are consistently less significant. For example, milk effects to three predominant taxa can only be identified at $\alpha = 0.05$ by our correlated functional response regression model, and our model suggests more significant C-section effects, although significance can be identified by both models. Effects of Period and Antibiotics are not significant under both models. These results imply that p-values under classic functional response regression model can be too conservative, and we conclude not to use the classic functional response regression model to avoid misleading test results when the functional responses are correlated.

5. Discussion

In this paper, we propose a correlated functional response regression model which can evaluate the association between correlated longitudinal OTU observations with their predictors. We further propose a data transformation technique to make our method computationally effective by using existing functional data analysis softwares. Predictors' effects are theoretically derived and their properties including unbiasedness, type I error and testing power are evaluated by comprehensive simulations. Both simulations and application studies show that our model performance is superior to classic functional response regression model, and only our model can provide accurate type I errors, p-values and type I errors on correlated functional response data. Our proposed method is the first functional regression model on longitudinal microbiome data, which provides solid and effective computational tool on future clinical and biological research.

Despite the clear benefits of our method, there are also some limitations with our current model. First, we assume the RAs of OTU data follow log-normal distribution, which may not be true in practice. OTU data may be zero-inflated, and several methods have been proposed to deal with zero-inflated OTU data when OTU data is not functional[33,34]. It is our future work to incorporate these methods, e.g., two-part model, into functional regression framework. The major challenge is to extend the generalized linear model with binary responses to functional response situation, so that the longitudinal data of OTU prevalence may also be fitted as functional responses.

Another limitation is that the hypothesis testing approach relying on the `fda` package may only test categorical rather than numerical covariates. Besides that, when the predictor is categorical, e.g., sex, it is sometimes of interest to see the separate fitted response curves for each category (male and female). Although these curves can be easily plotted under classic functional response regression model, it becomes more challenging under our model due to our data transformation technique. Our data transformation keeps $\boldsymbol{\beta}(t)$ estimations invariant but not the predictors. The transformed predictor of sex may have more than two categories, which may not have a practical meaning. Plotting fitted response curves with the transformed data does not really show any pattern related to male or female. Additional methodology development is under way to deal with the interpretation issue of categorical predictors after data transformation.

Acknowledgements

The authors acknowledge and are grateful for the support of the Tomcyzk AI and Microbiome Working Group.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

BC was supported by the Tomcyzk AI and Microbiome Working Group and the Princess Margaret Cancer Foundation. WX was funded by Natural Sciences and Engineering Research Council of Canada (NSERC Grant RGPIN-2017-06672), Princess Margaret Cancer Foundation Award.

ORCID iD

Bo Chen https://orcid.org/0000-0002-5916-4443

References

1. Gilbert JA, Blaser MJ, Caporaso JG et al. Current understanding of the human microbiome. *Nat Med* 2018; **24**: 392–400.
2. Faust K, Lahti L, Gonze D et al. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr Opin Microbiol* 2015; **25**: 56–66.
3. Gonzalez A, King A, 2nd MSR et al. Characterizing microbial communities through space and time. *Curr Opin Biotechnol* 2012; **23**: 431–436.
4. Gerber GK. The dynamic microbiome. *FEBS Lett* 2014; **588**: 4131–4139.
5. Backhed F, Roswall J, Peng Y et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 2015; **17**: 690–703.
6. Chen B and Xu W. Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures. *PLoS Comput Biol* 2020; **16**: e1008108.
7. Kostic AD, Gevers D, Siljander H et al. The dynamics of the human infant gut microbiome in development and in progression towards type 1 diabetes. *Cell Host Microbe* 2015; **17**: 260–273.
8. Caporaso JG, Lauber CL, Costello EK et al. Moving pictures of the human microbiome. *Genome Biol* 2011; **12**: R50.
9. Morris A, Paulson JN, Talukder H et al. Longitudinal analysis of the lung microbiota of cynomolgous macaques during long-term shiv infection. *BMC Microbiome* 2016; **4**: 38.
10. Ridenhour BJ, Brooker SL, Williams JE et al. Modeling time-series data from microbial communities. *ISME J* 2017; **11**: 2526–2537.
11. Morris JS. Functional regression. *Annu Rev Stat Appl* 2015; **2**: 321–359.
12. Mandal S, Van Treuren W, White RA et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* 2015; **26**: 27663.
13. Su L, Tom BDM, Long DL et al. Two-part and related regression models for longitudinal data. *Annu Rev Stat Appl* 2017; **4**: 283–315.
14. Anthea M. Random effects modeling and the zero-inflated poisson distribution. *Communications in Statistics - Theory and Methods* 2014; **43**: 664–680.
15. Chen EZ and Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 2016; **32**: 2611–2617.
16. Zhang X, Mallick H, Tang Z et al. Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics* 2017; **18**: 4.
17. Zhang X, Pei YF, Zhang L et al. Negative binomial mixed models for analyzing longitudinal microbiome data. *Front Microbiol* 2018; **9**: 1683.
18. La Rosa PS, Brooks JP, Deych E et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE* 2012; **7**: e52078.
19. Chen J and Li H. Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *Ann Appl Stat* 2013; **7**: 418–442.
20. Tang ZZ, Chen G, Alekseyenko AV et al. A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics* 2017; **33**: 1278–1285.
21. Tang ZZ and Chen G. Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* 2018; **20**: 698–713.
22. Tang ZZ and Chen G. Robust and powerful differential composition tests for clustered microbiome data. *Statistics in Biosciences* 2021; **13**: 200–216.
23. Guo W. Functional mixed effects models. *Biometrics* 2002; **58**: 121–128.
24. Morris JS and Carroll RJ. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* 2006; **68**: 179–199.

25. Antoniadis A and Sapatinas T. Estimation and inference in functional mixed-effects models. *Computational Statistics and Data Analysis* 2007; **51**: 4793–4813.
26. Scheipl F, Staicu AM and Greven S. Functional additive mixed models. *J Comput Graph Stat* 2015; **24**: 477–501.
27. Ramsay JO and Silverman BW. Modelling functional responses with multivariate covariates. In *Functional Data Analysis*, 2nd ed. New York, NY: Springer, 2005. pp. 223–245.
28. Ratliffe SJ, Heller GZ and Leader LR. Functional data analysis with application to periodically stimulated foetal heart rate data. *Stat Med* 2002; **21**: 1103–1127.
29. Yao F, Muller HG and Wang JL. Functional linear regression analysis for longitudinal data. *The Annals of Statistics* 2005; **33**: 2873–2903.
30. Reiss PT, Huang L and Mennes M. Fast function-on-scalar regression with penalized basis expansions. *Int J Biostat* 2010; **6**: 28.
31. Zhang JT. Statistical inferences for linear models with functional responses. *Statistica Sinicas* 2011; **21**: 1431–1451.
32. Larosa PS, Warner BB, Zhou Y et al. Patterned progression of bacterial populations in the premature infant gut. *PNAS* 2014; **111**: 12522–12527.
33. Xu L, Turpin W, Paterson AD et al. Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE* 2015; **10**: e0129606.
34. Kaul A, Mandal S, Davidov O et al. Analysis of microbiome data in the presence of excess zeros. *Front Microbiol* 2017; **8**: 2014.