# LOG-CONTRAST REGRESSION WITH FUNCTIONAL COMPOSITIONAL PREDICTORS: LINKING PRETERM INFANTS' GUT MICROBIOME TRAJECTORIES TO NEUROBEHAVIORAL OUTCOME

BY ZHE SUN[1,*], WANLI XU[2,‡], XIAOMEI CONG[2,§], GEN LI[3] AND KUN CHEN[1,†]

[1]*Department of Statistics, University of Connecticut, *zhe.sun@uconn.edu; †kun.chen@uconn.edu*
[2]*School of Nursing, University of Connecticut, ‡wanli.xu@uconn.edu; §xiaomei.cong@uconn.edu*
[3]*Department of Biostatistics, Columbia University, gl2521@cumc.columbia.edu*

The neonatal intensive care unit (NICU) experience is known to be one of the most crucial factors that drive preterm infants' neurodevelopmental and health outcome. It is hypothesized that stressful early life experience of very preterm neonate is imprinting gut microbiome by the regulation of the so-called brain-gut axis, and, consequently, certain microbiome markers are predictive of later infant neurodevelopment. To investigate, a preterm infant study was conducted; infant fecal samples were collected during the infants' first month of postnatal age, resulting in functional compositional microbiome data, and neurobehavioral outcomes were measured when infants reached 36–38 weeks of postmenstrual age. To identify potential microbiome markers and estimate how the trajectories of gut microbiome compositions during early postnatal stage impact later neurobehavioral outcomes of the preterm infants, we innovate a sparse log-contrast regression with functional compositional predictors. The functional simplex structure is strictly preserved, and the functional compositional predictors are allowed to have sparse, smoothly varying and accumulating effects on the outcome through time. Through a pragmatic basis expansion step, the problem boils down to a linearly constrained sparse group regression, for which we develop an efficient algorithm and obtain theoretical performance guarantees. Our approach yields insightful results in the preterm infant study. The identified microbiome markers and the estimated time dynamics of their impact on the neurobehavioral outcome shed lights on the linkage between stress accumulation in early postnatal stage and neurodevelpomental process of infants.

**1. Introduction.** Over the past decade, advances in neonatal care have contributed to a dramatic increase in survival among very preterm birth infants (born before 32 weeks' gestation) from 15% to over 90% (Fanaroff, Hack and MC (2003), Stoll, Hansen and Bell (2010)). With this cheerful gain in survival, recent research has shifted focus to the investigation of the increase in neurological morbidity and long-term adverse outcomes related to immature neuroimmune systems and stressful early life experience (Mwaniki et al. (2012)). In particular, the neonatal intensive care unit (NICU) experience is found to be one of the most crucial factors that drive preterm infant neurodevelopmental and health outcomes. Accumulated infant stress at NICU arises from numerous causes, such as repeated painful procedures, daily clustered care, maternal separation, among others. Mwaniki et al. (2012) showed that these neonatal insults were associated with a much escalated risk of long-term neurological morbidity, for example, 39.4% of NICU survivors had at least one neurodevelopmental deficit. However, the onset of the altered neuroimmune progress induced by infant stress/pain is often insidious, and the mechanism of this association, which holds the key for reducing costly health consequences of prematurity, remain largely unclear. Expanding research evidence

supports that a functional communication exists between the central nervous system and gastrointestinal tract, the brain-gut axis, in which the gut microbiome plays a key role in early programming and later responsivity of the stress system (Dinan and Cryan (2012)).

As such, a central hypothesis is that the stressful early life experience of very preterm neonates is imprinting gut microbiome by the regulation of the brain-gut axis, and, consequently, certain microbiome markers are predictive of later infant neurodevelopment. To investigate, a study was conducted in a NICU in the northeast of the U.S., where stable preterm infants were recruited. Infant fecal samples were collected daily, when available, during the infant's first month of postnatal age. Bacterial DNA were isolated and extracted from each stool sample and through sequencing and processing, resulted in gut microbiome data. Gender, delivery type, birth weight, feeding type, among others, were also recorded for each infant. Infant neurobehavioral outcomes were measured when the infant reached 36–38 weeks of postmenstrual age, using the NICU Network Neurobehavioral Scale (NNNS). More details on the study and the data are provided in Section 2. The above scientific hypothesis can then be approached through a statistical analysis, by examining how the microbiome compositions collected over the early postnatal period predict or impact on the later NNNS score, after adjusting for the effects of relevant infant characteristics.

The gut microbiome data were processed and operationalized as compositions, as commonly done in the microbiome literature (Bomar et al. (2011), Cong et al. (2017)). Compositional data analysis is not an unfamiliar territory to statisticians. Data consisting of percentages or proportions of certain composition are commonly encountered in various scientific fields, including ecology, biology and geology. One unique attribute of compositional data is the unit-sum constraint, that is, the components of a composition are nonnegative and always sum up to one; this entails that the data live in a simplex and thus renders many statistical methods that comply with Euclidean geometry inapplicable. Much foundational work on the statistical treatment of compositional data was done by John Aitchison (Aitchison (1982), Aitchison and Bacon-Shone (1984)); see Aitchison (2003) for a thorough survey on the subject. Of particular interest to us is regression with compositional predictors, for which the *log-contrast models* (Aitchison and Bacon-Shone (1984)) have been very popular. A prominent feature of the model is that it enables the regression analysis to obey the so-called principle of subcompositional coherence, that is, the compositional data should be analyzed in a way that the same results can be obtained regardless of whether we analyze the entire composition or only a subcomposition (Aitchison and Egozcue (2005)). Recently, Lin et al. (2014) studied a sparse linear regression model with compositional covariates, extending the log-contrast model to high dimensions. The problem was nicely formulated as a constrained lasso regression (Tibshirani (1996)) with a zero-sum linear constraint on the regression coefficients. Shi, Zhang and Li (2016) further extended the sparse regression model to the case of multiple linear constraints for the analysis of microbiome subcompositions. A debiased procedure was adopted to obtain an asymptotically unbiased estimator of the regression coefficients and its asymptotic distribution; see Li (2015) for a recent comprehensive review on microbiome compositional data analysis. However, to our knowledge, regression method on handling high-dimensional compositional trajectories or series is still lacking.

Motivated by the needs in identifying potential microbiome markers and estimating how the trajectories of microbiome compositions along early postnatal stage impact later neurobehavioral outcome, we propose a *sparse log-contrast regression model with functional compositional predictors*. In our approach detailed in Section 3, longitudinal microbial compositions are treated as functional compositional predictors with time-varying effects on the outcome. We build a scalar-on-function regression model for the log-transformed predictors which naturally connects to the log-contrast regression. We particularly focus on the identification of important microbes using a sparsity-inducing regularization method. Section 4

concerns the computational issues. Some theoretical properties of the proposed estimator that are of practical concern are discussed in Section 5. In Section 6 simulation studies showcase the superior performance of the proposed approach over several competing methods. The data analysis of the preterm infant study is presented in Section 7. The identified microbiome markers are justifiable based on existing literature, and the estimated dynamic trajectories of their impact on the outcome shed new lights on the functional linkage between the accumulation of prenatal stress and neurodevelpment of infants. Some concluding remarks are given in Section 8. We have implemented the proposed methods and relevant tools in a user-friendly R package called `Compack` (Sun and Chen (2020)), which has been released to The Comprehensive R Archive Network (R Core Team (2020)) and can be downloaded at https://cran.r-project.org/web/packages/Compack/.

**2. Preterm infant study and problem setup.** Data were collected at a Level IV NICU in the northeast region of the U.S. (Level IV NICUs provide the highest level, the most acute care.) Fecal samples of preterm infants were collected daily when available, mainly during the infant's postnatal age (PNA) of five to 28 days ($t \in [5, 28]$). Bacterial DNA were isolated and extracted from each stool sample (Bomar et al. (2011), Cong et al. (2017)); the V4 regions of the 16S rRNA gene were sequenced using the Illumina platform and clustered and analyzed using QIIME (Cong et al. (2017)), resulting in microbiome count data. Since the number of sequencing reads varied a lot across samples, we further normalize the data by calculating the ratio of each microbe in each sample. As a result, we obtain a compositional data matrix. To conduct log transformation in our model, following the convention in the literature we replace zeros by the maximum rounding error (i.e., 0.5) to avoid singularity (Aitchison (2003), Lin et al. (2014)). Due to the limited sample size, we mainly focus on $p = 22$ categories at the order level of the taxonomic ranks as a proof of concept. (We also perform a confirmative analysis at the genus level which has more than 60 categories.) Taxonomic rank is the relative level of a group of organisms in a taxonomic hierarchy in biological classification; the major ranks are species, genus, family, order, class, phylum, kingdom and domain. More details on the microbiome data are provided in Section 3 of the Supplemental Material (Sun et al. (2020)). In this study infants with less than five fecal samples were excluded which resulted in $n = 34$ infants. There was a total of 414 fecal samples, so the average number of daily fecal samples collected for each infant was 12.2. Figure 1(a) shows the histogram of the number of samples collected from each infant, and Figure 1(b)–(d) show examples of the observed profile of the time-varying compositions along the postnatal age for three different infants.

Infant neurobehavioral outcomes were measured when the infant reached 36–38 weeks of postmenstrual age or prior to hospital discharge, using the NICU Network Neurobehavioral Scale (NNNS). The NNNS is a standardized assessment of neonatal neurobehavioral outcomes that provides an appraisal of neurological integrity and behavioral function of the normal and at-risk/preterm infant. In particular, the stress/abstinence subscale (NSTRESS) measures signs of stress and includes 50 items. Each sign of stress/abstinence is scored as present or absent, and the composite NSTRESS score ranges between 0 and 1. A higher NSTRESS score demonstrates a more stressful behavioral performance. Cong et al. (2017) showed that the composite NSTRESS score is positively associated with painful/stressful experience in preterm infants. Other variables about birth and characteristics of an infant included gender, delivery type, premature rupture of membranes (PROM), score for Neonatal Acute Physiology–Perinatal Extension-II (SNAPPE-II), birth weight and percentage of feeding with mother's breast milk (%MBM).

To formulate the statistical problem, let $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}} \in \mathbb{R}^n$ be consisting of the observed neurobehavioral outcomes of the preterm infants, that is, their NNNS scores. Let
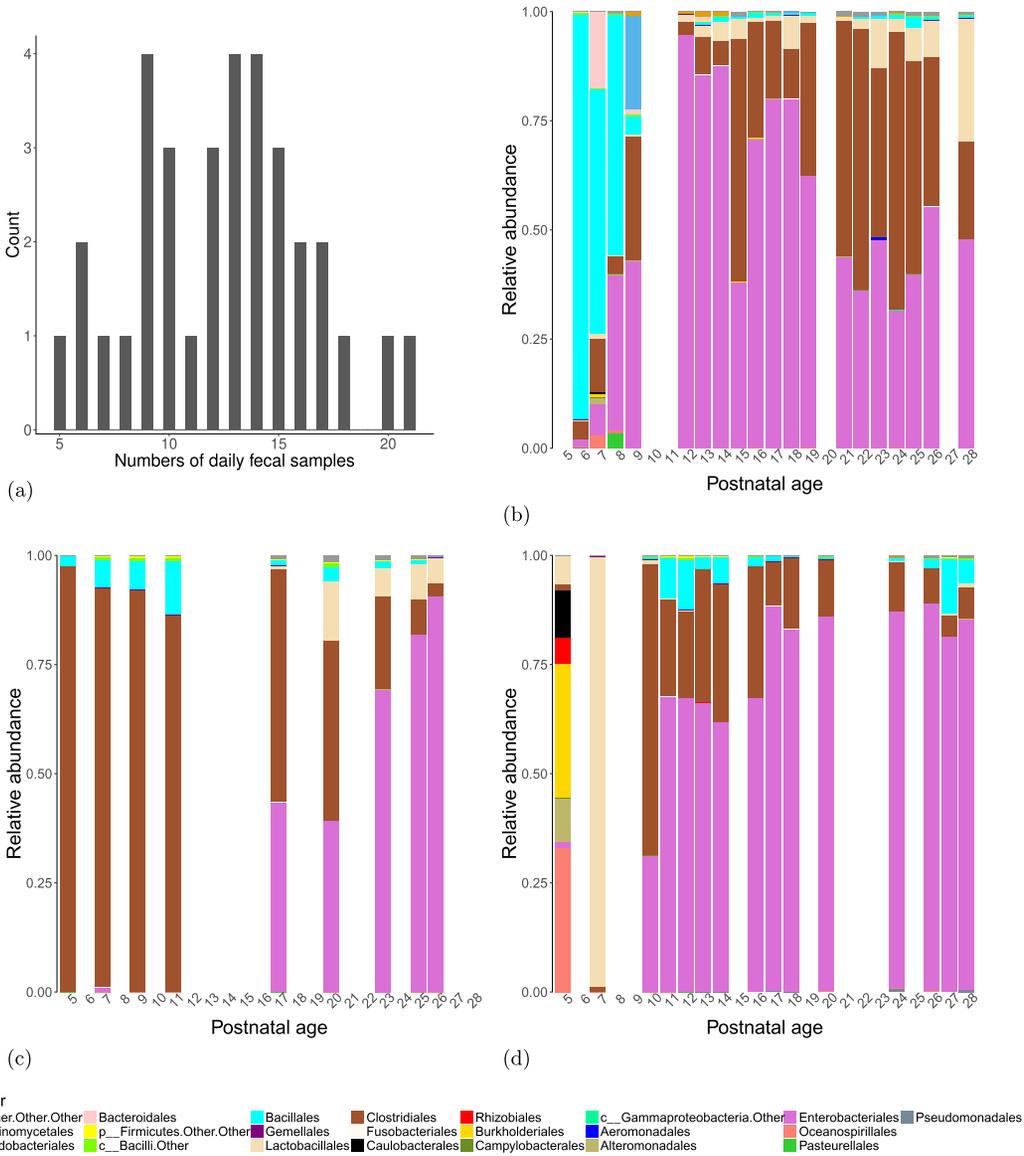
FIG. 1. (a) *Histogram of the number of samples collected from each infant.* (b)–(d) *Example profiles of time–varying compositional data along postnatal age from three infants.*

$\mathbf{x}_i(t) = [x_{i1}(t), \ldots, x_{ip}(t)]^{\mathrm{T}} \in \mathbb{S}^{p-1}$ be the gut microbiome compositions from the $i$th infant at time $t$. Here, we let $\mathbb{S}^{p-1} = \{[x_1, \ldots, x_p]^{\mathrm{T}} \in \mathbb{R}^p; x_j > 0, \sum_{j=1}^{p} x_j = 1.\}$, to denote the $(p-1)$-dimensional positive simplex lying in $\mathbb{R}^p$. Let $\mathbf{X}(t) = [\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)]^{\mathrm{T}} \in \mathbb{R}^{n \times p}$ be the matrix of the functional predictors at time $t$. The observed gut microbiome compositions during the early postnatal period can then be viewed as discrete observations from $\mathbf{X}(t)$. Also define $\mathbf{Z}_c \in \mathbb{R}^{n \times p_c}$, formed by data from the aforementioned time-invariant infant characteristics, for example, gender, delivery type, among others.

As the main objective is to identify the microbiome markers that are predictive of later infant neurodevelopment, we need to perform a regression analysis to examine how the outcome $y$, the NNNS score, is associated with $\mathbf{X}(t)$, the gut micorbiome trajectories, while controlling for the infant characteristics collected in $\mathbf{Z}_c$. The fact that $\mathbf{X}(t)$ is both functional and compositional makes the problem very challenging.

## 3. Regression with functional compositional predictors.

3.1. *Linear log-contrast model.* We first briefly review the existing regression approaches for dealing with a single set of compositional predictors. Suppose we observed $n$ independent observations of a response variable $y_i \in \mathbb{R}$ and a compositional predictor $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ such that $\mathbf{x}_i \in \mathbb{S}^{p-1}$. Denote $\mathbf{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$ as the response vector and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ as the design matrix.

Ignoring the simplex structure of $\mathbf{X}$ would lead to parameter identifiablity issue in the linear regression of $\mathbf{y}$ on $\mathbf{X}$. One naive "remedy" is to exclude an arbitrary component of the compositional vector in the regression, which, however, leads to a method that is not invariant to the choice of the removed component since it affects both of prediction and selection and, consequently, makes proper model interpretation and inference difficult. Ever since the pioneer work by John Aitchison (Aitchison (1982, 2003), Aitchison and Bacon-Shone (1984)) on the statistical treatments of compositional data, the so-called *log-contrast model* has gained much popularity in a variety of regression problems with compositional predictors. The main idea is to perform a log-ratio transformation of the compositional data, such that the transformed data admit the familiar Euclidean geometry in $\mathbb{R}^{p-1}$. Specifically, for each $i = 1, \ldots, n$, let $\widetilde{z}_{ij} = \log(x_{ij}/x_{ir})$, where $r \in \{1, \ldots, p\}$ is a chosen reference level, and $j = 1, \ldots, r-1, r+1, \ldots, p$, resulting in $\widetilde{\mathbf{Z}}_{\bar{r}} = [\widetilde{z}_{ij}] \in \mathbb{R}^{n \times (p-1)}$. Also define $z_{ij} = \log(x_{ij})$ and $\mathbf{Z} = [z_{ij}] \in \mathbb{R}^{n \times p}$. The linear log-contrast regression model is expressed as

$$(1) \qquad \mathbf{y} = \beta_0^* \mathbf{1}_n + \widetilde{\mathbf{Z}}_{\bar{r}} \beta_{\bar{r}}^* + \mathbf{e},$$

where $\beta_0^*$ is the intercept, $\beta_{\bar{r}}^* \in \mathbb{R}^{p-1}$ is the regression coefficient vector and $\mathbf{e} \in \mathbb{R}^n$ is the random error vector with zero mean. Interestingly, although it appears that the model in (1) depends on the choice of the reference level, it in fact admits a symmetric form. By simple algebra, model (1) can be equivalently expressed as

$$(2) \qquad \mathbf{y} = \beta_0^* \mathbf{1}_n + \mathbf{Z} \beta^* + \mathbf{e} \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j^* = 0,$$

where $\beta^*$ is the regression coefficient vector for design matrix $\mathbf{Z}$, and $\mathbf{e}$ and $\beta_0^*$ are the same as in model (1). It can be showed that $\beta_{\bar{r}}^* \in \mathbb{R}^{p-1}$ is a subvector of a regression coefficient vector $\beta^* \in \mathbb{R}^p$ by removing its $r$th component $\beta_r^*$.

Consequently, in classical regression setups the least squares estimation under model (1) is equivalent to the constrained least squares estimation under model (2). However, in high-dimensional scenarios, that is, when $p$ is much larger than $n$, the two model formulations could lead to discrepancies in regularized estimation. For example, the two corresponding lasso criteria (Tibshirani (1996)) are no longer equivalent:

$$(3) \qquad \min_{\beta_0, \beta_{\bar{r}}} \left\{ \frac{1}{2n} \|\mathbf{y} - \beta_0 \mathbf{1}_n - \widetilde{\mathbf{Z}}_{\bar{r}} \beta_{\bar{r}}\|^2 + \lambda \|\beta_{\bar{r}}\|_1 \right\},$$

$$(4) \qquad \min_{\beta_0, \beta} \left\{ \frac{1}{2n} \|\mathbf{y} - \beta_0 \mathbf{1}_n - \mathbf{Z} \beta\|^2 + \lambda \|\beta\|_1 \right\} \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j = 0,$$

where $\|\cdot\|$, $\|\cdot\|_1$ denote the $\ell_2$, $\ell_1$ norms, respectively, and $\lambda$ is a tuning parameter controlling the amount of regularization. Although (3) is simpler to compute, clearly its solution and, hence, its variable selection depend on the choice of the reference component. In contrast, (4) remains to be symmetric in all the $p$ components. Lin et al. (2014) proposed and studied (4) and showed that the estimator admits many desirable properties (Aitchison (2003)).

3.2. *Sparse functional log-contrast regression.* In the preterm infant study the compositional predictors are observed over a continuous domain, that is, time, and thus they should be treated as functional compositional data. Recall from Section 2 that $\mathbf{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{X}(t) \in \mathbb{R}^{n \times p}$ the matrix of the functional and compositional predictors at $t$ and $\mathbf{Z}_c \in \mathbb{R}^{n \times p_c}$ the matrix of time-invariant control variables. Here, to focus on the main idea, we assume $\mathbf{X}(t)$ is completely observed for $t \in \mathbb{T}$, and the discussion about handling discrete time data is deferred to Section 4.2. Similar as in Section 3.1, we define $\widetilde{\mathbf{Z}}_{\bar{r}}(t) \in \mathbb{R}^{n \times (p-1)}$, for $r = 1, \ldots, p$, and $\mathbf{Z}(t) = \log(\mathbf{X}(t)) \in \mathbb{R}^{n \times p}$.

Motivated by model (2), we propose a *functional log-contrast regression model*,

$$(5) \qquad \mathbf{y} = \beta_0^* \mathbf{1}_n + \mathbf{Z}_c \beta_c^* + \int_{t \in \mathbb{T}} \mathbf{Z}(t) \beta^*(t) \, dt + \mathbf{e} \quad \text{s.t. } \mathbf{1}_p^T \beta^*(t) = 0, \forall t \in \mathbb{T},$$

where $\beta_0^*$ is the intercept, $\beta_c^* \in \mathbb{R}^{p_c}$ is the regression coefficient vector corresponding to the control variables, $\beta^*(t) = [\beta_1^*(t), \ldots, \beta_p^*(t)]^T \in \mathbb{R}^p$ is the functional regression coefficient vector as a function of $t$ and the remaining terms are defined the same as in model (2). The proposed model allows the compositional predictors to have potentially different effects on the response through $\beta^*(t)$, and their aggregated effects on the response is then given by the integral of $\mathbf{Z}(t)$ weighted by $\beta^*(t)$ over time. Following Lin et al. (2014), here we adopt the symmetric form of the log-contrast model, in which the zero-sum constraints preserve the simplex structure over time while all the compositional components are treated equally.

To address the problems in the preterm infant study, we consider both sparsity and smoothness of $\beta^*(t)$. First, as it is believed that only a few compositional components are relevant to the prediction of the outcome, we assume the true coefficient curves are sparse, that is, $s^* = |\mathcal{S}| \ll p$, where $\mathcal{S}$ is the index set of the nonzero coefficient curves

$$\mathcal{S} = \{j; \beta_j^*(t) \neq 0 \text{ for some } t \in \mathbb{T}, j = 1, \ldots, p\}.$$

This sparsity assumption is the basis of component selection and is widely applicable, especially when $p$, the number of compositional components, is large. Second, since the effects of gut microbiome compositions on preterm infant's neurodevelopment evolves gradually over the postnatal period, we assume the coefficient curves are smooth over $t$ and adopt a truncated basis expansion approach (Ramsay and Silverman (2005)) to bring the infinite dimensional problem to finite dimensions. Specifically, we assume

$$(6) \qquad \beta^*(t) = \mathbf{B}^* \mathbf{\Phi}(t),$$

where $\mathbf{B}^* = [\beta_1^*, \ldots, \beta_p^*]^T \in \mathbb{R}^{p \times k}$ is a coefficient matrix, and $\mathbf{\Phi}(t) = [\phi_1(t), \ldots, \phi_k(t)]^T \in \mathbb{R}^k$ consists of basis with $\mathbf{J}_{\phi\phi} = \int_{t \in \mathbb{T}} \mathbf{\Phi}(t) \mathbf{\Phi}^T(t) \, dt$ being a positive definite (p.d.) matrix. Here, for simplicity, the same set of basis functions is used in the expansion of each $\beta_j(t)$, $j = 1, \ldots, p$, which usually suffices in practice, and the extension to use different basis for different $\beta_j(t)$ is straightforward. There are many choices of the basis functions, for example, Fourier basis, wavelet basis and spline basis; see Ramsay and Silverman (2005) for a detailed account on the truncated basis expansion approaches in functional regression.

Some discussions on the number of basis functions are in order. In classical least squares types of estimation, the choice of $k$ usually boils down to a bias-and-variance tradeoff. That is, while larger values of $k$ can lead to a better in-sample estimation at the risk of potential overfitting, smaller values of $k$ result in simpler estimators at the expense of missing interesting local oscillations. The issue can be resolved by echoing regularization, that is, taking a sufficiently large $k$ to ensure the flexibility of the model and performing regularized estimation to avoid overfitting. From a theoretical perspective we allow $k$ to grow with the sample size $n$, that is, the complexity of the functional curves that the method can potentially capture may increase when more data become available; see Section 5 for details. We also remark that, for

a nonparametric treatment, one can assume $\beta^*(t)$ satisfies the Hölder condition (Tsybakov (2009)) to control the approximate error induced by the basis truncation.

The functional sparsity in $\beta^*(t)$ now amounts to the row-sparsity of the coefficient matrix $\mathbf{B}^*$ in (6). The zero-sum constraint on $\beta^*(t)$, that is, $\mathbf{1}_p^{\mathrm{T}}\beta^*(t) = 0$ for all $t \in \mathbb{T}$, is now equivalent to $\mathbf{B}^{*\mathrm{T}}\mathbf{1}_p = \mathbf{0}$. To see this, note that $\mathbf{1}_p^{\mathrm{T}}\beta^*(t) = 0$ leads to

$$\int_{t\in\mathbb{T}} \mathbf{1}_p^{\mathrm{T}}\mathbf{B}^*\mathbf{\Phi}(t)\mathbf{\Phi}(t)^{\mathrm{T}}(\mathbf{1}_p^{\mathrm{T}}\mathbf{B}^*)^{\mathrm{T}}\,dt = \mathbf{1}_p^{\mathrm{T}}\mathbf{B}^*\mathbf{J}_{\phi\phi}(\mathbf{1}_p^{\mathrm{T}}\mathbf{B}^*)^{\mathrm{T}} = 0;$$

it follows that $\mathbf{B}^{*\mathrm{T}}\mathbf{1}_p = \mathbf{0}$ as $\mathbf{J}_{\phi\phi}$ is p.d. (The other direction holds trivially.) Further, the integral part in the model becomes

$$\int_{t\in\mathbb{T}} \mathbf{Z}(t)\beta^*(t)\,dt = \int_{t\in\mathbb{T}} \mathbf{Z}(t)\mathbf{B}^*\mathbf{\Phi}(t)\,dt$$

$$= \left\{\int_{t\in\mathbb{T}} \mathbf{Z}(t)\big(\mathbf{I}_p \otimes \mathbf{\Phi}(t)^{\mathrm{T}}\big)\,dt\right\} \mathrm{vec}(\mathbf{B}^{*\mathrm{T}}) = \mathbf{Z}\beta^*,$$

where, with some abuse of notations, we redefine $\beta^* = [\beta_1^{*\mathrm{T}}, \ldots, \beta_p^{*\mathrm{T}}]^{\mathrm{T}} = \mathrm{vec}(\mathbf{B}^{*\mathrm{T}}) \in \mathbb{R}^{pk}$ and

$$(7) \qquad \mathbf{Z} = \int_{t\in\mathbb{T}} \mathbf{Z}(t)\big(\mathbf{I}_p \otimes \mathbf{\Phi}(t)^{\mathrm{T}}\big)\,dt = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p] \in \mathbb{R}^{n\times(pk)}.$$

Each $\beta_j^* \in \mathbb{R}^k$ and $\mathbf{Z}_j \in \mathbb{R}^{n\times k}$ correspond to the coefficient vector and the covariate matrix for the $j$th compositional component, respectively. We remark that $\mathbf{Z}$ is usually not exactly computed since $\mathbf{Z}(t)$ may not be fully observed; we defer the discussion to Section 4.2.

The functional model in (5) then becomes a constrained sparse linear regression model

$$(8) \qquad \mathbf{y} = \beta_0^*\mathbf{1}_n + \mathbf{Z}_c\beta_c^* + \mathbf{Z}\beta^* + \mathbf{e} \quad \text{s.t.} \sum_{j=1}^p \beta_j^* = \mathbf{0},$$

where $\beta^*$ is expected to be sparse accordingly to the row-sparsity of $\mathbf{B}^*$. To enable the selection of the compositional components, we therefore propose to conduct model estimation by minimizing a linearly constrained group lasso criterion (Yuan and Lin (2006)),

$$(9) \qquad \min_{\beta_0,\beta_c,\beta} \left\{ \frac{1}{2n}\|\mathbf{y} - \beta_0\mathbf{1}_n - \mathbf{Z}_c\beta_c - \mathbf{Z}\beta\|^2 + \lambda \sum_{j=1}^p \|\beta_j\| \right\} \quad \text{s.t.} \sum_{j=1}^p \beta_j = \mathbf{0},$$

where $\lambda$ is a tuning parameter controlling the amount of regularization. We remark that the group lasso penalty is imposed on the coefficients for each microbiome category to encourage microbe selection.

The proposed estimator possesses several desirable invariance properties (Aitchison (2003), Lin et al. (2014)):

(I) Scale invariance: The estimator is invariant to the transformation $\mathbf{X}(t) \to \mathbf{S}\mathbf{X}(t)$ where $\mathbf{S} = \mathrm{diag}(\mathbf{s})$ is a diagonal matrix with diagonal elements $\mathbf{s} = [s_1, \ldots, s_n]^{\mathrm{T}}$ and all $s_i > 0$. That is, it does not matter whether the data vectors are scaled to have a unit sum; the method only cares about the relative proportions. This is simply because $\mathbf{Z}(t)\beta(t) = \{\log(\mathbf{X}(t)) + \log(\mathbf{s})\mathbf{1}_p^{\mathrm{T}}\}\beta(t) = \log(\mathbf{X}(t))\beta(t)$, due to the zero-sum constraints. In fact, this scale invariance continues to hold when the scaling factor $\mathbf{s}$ changes in time.

(II) Permutation invariance: Results of the analysis do not depend on the sequence by which the components are given or labeled.

(III) Subcomposition coherence: If we know in advance that some $\beta_j(t)$ curves are zero, the analysis is unchanged if we apply the procedure to the subcompositions formed by the components of $\mathbf{X}(t)$ corresponding to the other $\beta_j(t)$ curves. To see this, suppose $\beta_j(t) \equiv 0$

for $j \in \mathcal{S}^c$, where $\mathcal{S}^c$ is the complement of a set $\mathcal{S}$ on $\{1, \ldots, p\}$. Let $\mathbf{s}(t) = \{\mathbf{X}_{\mathcal{S}}(t)\mathbf{1}_{|\mathcal{S}|}\}^{-1} \in \mathbb{R}^n$ be a scaling factor in which the inversion is entrywisely applied, so that $\operatorname{diag}(\mathbf{s}(t))\mathbf{X}_{\mathcal{S}}(t)$ gives the subcompositions formed by the components in $\mathcal{S}$. Then, we have

$$\log(\mathbf{X}(t))\beta(t) = \{\log(\mathbf{X}_{\mathcal{S}}(t)) + \log(\mathbf{s}(t))\mathbf{1}_{|\mathcal{S}|}^{\mathrm{T}}\}\beta_{\mathcal{S}}(t)$$
$$= \log(\operatorname{diag}(\mathbf{s}(t))\mathbf{X}_{\mathcal{S}}(t))\beta_{\mathcal{S}}(t).$$

In particular, when there are only two nonzero components, for example, $\beta_1(t) \neq 0$, $\beta_2(t) \neq 0$ and $\beta_j(t) = 0$ for $j = 3, \ldots, p$, it is necessarily true that $\beta_1(t) = -\beta_2(t)$ due to the zero-sum constraint. This is neither an unpleasant artifact nor a limitation of the proposed method. This special case can be understood from the above property of subcomposition coherence: the analysis becomes the same as using the subcompositions formed from the first two components of $\mathbf{X}(t)$; consequently, the two possible log-ratios are exactly opposite to each other, so do their corresponding coefficient curves. Therefore, this feature is consistent with the data structure, as in two-part compositional data; either part carries exactly the same information.

## 4. Computation.

4.1. *Solving and tuning constrained group lasso.* The problem in (9) is convex, and we solve it by an augmented Lagrangian algorithm (Boyd et al. (2011)). To save space, details are provided in Section 1 the of Supplementary Material (Sun et al. (2020)).

A general way to select the tuning parameters, that is, the basis dimension $k$ and the group penalty level $\lambda$, is the $K$-fold cross-validation (Stone (1974)) which is based on the predictive performance of the models. However, it is well known that the best model for prediction may not coincide with that for variable selection, and, in fact, the former often leads to overselection. This phenomenon under our model is revealed in Section 5, where it is shown that consistent component selection shall be based on the zero pattern of a thresholded estimator. Following Fan and Tang (2013) and Lin et al. (2014), we thus also experiment with minimizing a generalized information criterion (GIC) for model selection, which favors more sparse models,

$$\mathrm{GIC}(\lambda, k) = \log(\widehat{\sigma}^2(\lambda, k)) + (s(\lambda, k) - 1)k \log(\max\{pk + 1 + p_c, n\})\frac{\log(\log n)}{n},$$

where $\widehat{\sigma}^2(\lambda, k)$ is the mean squared error define as $\|\mathbf{y} - \widehat{\beta}_0(\lambda, k)\mathbf{1}_n - \mathbf{Z}_c\widehat{\beta}_c(\lambda, k) - \mathbf{Z}\widehat{\beta}(\lambda, k)\|^2/n$ with $\widehat{\beta}_0(\lambda, k)$, $\widehat{\beta}_c(\lambda, k)$ and $\widehat{\beta}(\lambda, k)$ being the regularized estimators of regression coefficients and $s(\lambda, k)$ is the number of nonzero coefficient groups in $\widehat{\beta}(\lambda, k)$.

4.2. *On discrete time observations.* So far we have treated the integrated design matrix $\mathbf{Z}$, defined in (7), as given. In practical situations, however, the functional compositional predictors are most often not observed continuously but at discrete points, so $\mathbf{Z}$ can not be computed exactly. It is preferable that the induced uncertainty is considered in statistical modeling. In functional regression with a scalar response, Ramsay and Silverman (2005) discussed using truncated basis expansions for both the functional predictor and the functional coefficient curve to convert the infinite dimensional problem to finite dimensional, where truncation can be viewed as a type of regularization. Integrals were approximated by finite Riemann sums with discrete observations. The subsequent methodological development in functional regression has mainly followed along this general strategy with various choices of basis functions and associated regularization approaches (Morris (2015)). For example, a functional predictor could be expanded by its eigenbasis via a functional principal component analysis, and the coefficient function could be expanded either by the same eigenbasis or by other basis, such as wavelet or spline.

Due to the nature of the compositional data, ideally the functional compositions shall be expanded by a multivariate basis that preserves the simplex structure under truncation or other types of regularization which, however, to the best of our knowledge is not yet available. In essence, a multivariate functional principal component analysis for compositional data, or a joint modeling approach of both the functional compositions and the regression, is needed which is beyond the scope of the current work.

For the preterm infant study we take a pragmatic way of lifting the discrete-time data to continuous time. In this study stool sample of each baby was collected daily whenever available; this resulted in a good coverage rate with, on average, 12.2 daily samples for each infant over a 24-day study period. Also, biologists believe that the gut microbiome compositions change continuously over time. As such, we simply apply linear interpolation to obtain continuous time compositional curves. It can be readily seen that the linear interpolation approach amounts to compute $\mathbf{Z}$, defined in (7), using the trapezoid rule.

Specifically, suppose for each $i = 1, \ldots, n$, we observe $\mathbf{x}_i(t) = [x_{i1}(t), \ldots, x_{ip}(t)]^{\mathrm{T}}$ at discrete time points $t_{i,v} \in \mathbb{T} = [T_1, T_2]$, for $v = 1, \ldots, m_i$. That is, different subjects may be observed at different sets of time points in $\mathbb{T}$. Correspondingly, we have

$$\mathbf{z}_i(t) = [z_{i1}(t), \ldots, z_{ip}(t)]^{\mathrm{T}}, \quad t = t_{i,1}, \ldots, t_{i,m_i}, i = 1, \ldots, n.$$

Recall that $\mathbf{Z} = \int_{t \in \mathbb{T}} \mathbf{Z}(t)(\mathbf{I}_p \otimes \mathbf{\Phi}(t)^{\mathrm{T}}) \, dt \in \mathbb{R}^{n \times (pk)}$. Let $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p] \in \mathbb{R}^{n \times (pk)}$ with $\mathbf{Z}_j = [z_{ijl}]_{n \times k} \in \mathbb{R}^{n \times k}$ for $j = 1, \ldots, p$. Adopting linear interpolation, the entries of $\mathbf{Z}$ are computed using the trapezoid rule as follows:

$$
\begin{aligned}
(10) \quad z_{ijl} = {}& \sum_{v=2}^{m_i} \big(\phi_l(t_{i,v-1}) z_{ij}(t_{i,v-1}) + \phi_l(t_{i,v}) z_{ij}(t_{i,v})\big) \frac{t_{i,v} - t_{i,v-1}}{2} \\
& + \phi_l(t_{i,1}) z_{ij}(t_{i,1})(t_{i,1} - T_0) + \phi_l(t_{i,m_i}) z_{ij}(t_{i,m_i})(T_1 - t_{i,m_i}),
\end{aligned}
$$

for $l = 1, \ldots, k$. In what follows, unless otherwise noted, the integrals in the case of discrete data are computed using the above trapezoid rule.

## 5. Theoretical perspectives.

Here, we attempt to provide some theoretical perspectives of two questions of practical concerns: (1) whether it is indeed beneficial to use the linearly constrained formulation rather than a naive baseline formulation, which chooses an arbitrary reference component to perform the log-ratio transformation of the compositional predictors and then proceeds with an unconstrained group lasso regression, and (2) whether the proposed method can accurately identify the relevant compositional predictors.

We first describe the setup. Our analysis is under the setting when the basis expansion in (6) holds and the integrated design matrix $\mathbf{Z}$ is available. The results are nonasymptotic, where both the number of functional predictors $p$ and the degrees of freedom of the basis functions $k$ are allowed to grow with the sample size $n$. For any $\beta = [\beta_1^{\mathrm{T}}, \ldots, \beta_p^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{pk}$, define $\beta_{\bar{r}} \in \mathbb{R}^{(p-1)k}$ as a subvector of $\beta$ by removing its $r$th component $\beta_r$, for each $r = 1, \ldots, p$. Let $\mathcal{J} \subset \{1, \ldots, p\}$ be an index set, and denote $\beta_{\mathcal{J}}$ be a subvector of $\beta$ consisting of $\beta_j$, $j \in \mathcal{J}$. Denote $\mathcal{J}^c$ as the complement of $\mathcal{J}$. Recall that $\mathbf{X}(t) = [\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)]^{\mathrm{T}} \in \mathbb{R}^{n \times p}$, $\mathbf{Z}(t) = [z_{ij}(t)] \in \mathbb{R}^{n \times p}$ with $z_{ij}(t) = \log(x_{ij}(t))$, and $\widetilde{\mathbf{Z}}_{\bar{r}}(t) = [\widetilde{z}_{ij}(t)] \in \mathbb{R}^{n \times (p-1)}$ with $\widetilde{z}_{ij}(t) = \log(x_{ij}(t)/x_{ir}(t))$ for each $r = 1, \ldots, p$. Moreover, due to (6), we define $\widetilde{\mathbf{Z}}_{\bar{r}} = \int_{t \in \mathbb{T}} \widetilde{\mathbf{Z}}_{\bar{r}}(t)(\mathbf{I}_p \otimes \mathbf{\Phi}(t)^{\mathrm{T}}) \, dt \in \mathbb{R}^{n \times (p-1)k}$ and $\mathbf{Z} = \int_{t \in \mathbb{T}} \mathbf{Z}(t)(\mathbf{I}_p \otimes \mathbf{\Phi}(t)^{\mathrm{T}}) \, dt \in \mathbb{R}^{n \times (pk)}$ as in (7). Write $\widetilde{\mathbf{Z}}_{\bar{r}} = [\widetilde{\mathbf{Z}}_{\bar{r},1}, \ldots, \widetilde{\mathbf{Z}}_{\bar{r},r-1}, \widetilde{\mathbf{Z}}_{\bar{r},r+1}, \ldots, \widetilde{\mathbf{Z}}_{\bar{r},p}]$ with each $\widetilde{\mathbf{Z}}_{\bar{r},j} \in \mathbb{R}^{n \times k}$. Write $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$ with each $\mathbf{Z}_j \in \mathbb{R}^{n \times k}$. Let $\Psi_{\bar{r},j} = \widetilde{\mathbf{Z}}_{\bar{r},j}^{\mathrm{T}} \widetilde{\mathbf{Z}}_{\bar{r},j}/n$, for $r = 1, \ldots, p$, $j = 1, \ldots, p$ and $j \neq r$. It boils down to analyze the constrained linear model with grouped predictors in

(8). For simplicity, we omit the intercept and the control variables and write the model as

$$\mathbf{y} = \mathbf{Z}\beta^* + \mathbf{e} \quad \text{s.t.} \sum_{j=1}^{p} \beta_j^* = \mathbf{0},$$

where $\beta^* = [\beta_1^{*T}, \ldots, \beta_p^{*T}]^T \in \mathbb{R}^{pk}$. Recall that $\mathcal{S} = \{j; \beta_j^*(t) \neq \mathbf{0}, j = 1, \ldots, p.\} = \{j; \beta_j^* \neq \mathbf{0}, j = 1, \ldots, p.\}$ and $s^* = |\mathcal{S}| \ll p$.

The proposed constrained group lasso estimator is

$$(11) \qquad \widehat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{Z}\beta\|^2 + \lambda \sum_{j=1}^{p} \|\beta_j\| \right\} \quad \text{s.t.} \sum_{j=1}^{p} \beta_j = \mathbf{0}.$$

This estimator satisfies that $\widehat{\beta}_r = -\sum_{j \neq r}^{p} \widehat{\beta}_j$. Therefore, it holds true that for any $r = 1, \ldots, p$,

$$\widehat{\beta}_{\bar{r}} = \arg\min_{\beta_{\bar{r}}} \left\{ \frac{1}{2n} \|\mathbf{y} - \widetilde{\mathbf{Z}}_{\bar{r}}\beta_{\bar{r}}\|^2 + \lambda \sum_{j \neq r}^{p} \|\beta_j\| + \lambda \left\| \sum_{j \neq r}^{p} \beta_j \right\| \right\}.$$

On the other hand, as to the baseline method, when the $r$th component is choosing as the reference level, the estimator is given by

$$(12) \qquad \widetilde{\beta}_{\bar{r}} = \arg\min_{\beta_{\bar{r}}} \left\{ \frac{1}{2n} \|\mathbf{y} - \widetilde{\mathbf{Z}}_{\bar{r}}\beta_{\bar{r}}\|^2 + \lambda \sum_{j \neq r}^{p} \|\beta_j\| \right\}.$$

Our analysis follows and extends the work by Lounici et al. (2011) on group lasso to the case of constrained group lasso in (11) arising from functional compositional data analysis. All the proofs are provided in Section 2 of the Supplementary Material (Sun et al. (2020)).

ASSUMPTION 1.    The error terms $e_1, \ldots, e_n$ are independently and identically distributed as N(0, 1) random variables.

ASSUMPTION 2 (Restricted Eigenvalue Condition (RE)).    There exists $\kappa > 0$, such that

$$\min \left\{ \frac{\|\mathbf{Z}\Delta\|}{\sqrt{n}\|\Delta_{\mathcal{J}}\|} : |\mathcal{J}| \leq s^*, \Delta \in \mathbb{R}^{pk} \neq \mathbf{0}, \sum_{j=1}^{p} \Delta_j = \mathbf{0}, \right.$$

$$\left. \sum_{j \in \mathcal{J}^c} \|\Delta_j\| + \min_j \|\Delta_j\| \leq 3 \sum_{j \in \mathcal{J}} \|\Delta_j\| \right\} \geq \kappa.$$

THEOREM 1 (Error Bounds).    *Suppose Assumptions 1–2 hold. Choose*

$$\lambda \geq \min_r \max_{j \neq r} \frac{2\sigma}{\sqrt{n}} \sqrt{\operatorname{tr}(\Psi_{\bar{r},j}) + 2\sigma_{\max}(\Psi_{\bar{r},j})(2q\log(p-1) + \sqrt{kq\log(p-1)})}.$$

*Then, with probability at least $1 - 2(p-1)^{1-q}$, the constrained group lasso estimator $\widehat{\beta}$ in (11) satisfies that*

$$\frac{1}{n}\|\mathbf{Z}(\widehat{\beta} - \beta^*)\|^2 \leq \frac{16\lambda^2 s^*}{\kappa^2},$$

$$\sum_{j=1}^{p} \|\widehat{\beta}_j - \beta_j^*\| + \min_j \|\widehat{\beta}_j - \beta_j^*\| \leq \frac{16\lambda s^*}{\kappa^2}.$$

It is interesting to compare with the baseline approach in (12) for which, once a baseline $r$ is chosen, its theoretical property mimics that of the regular group lasso model with $p -$ 1 groups (Lounici et al. (2011)). Due to the linear constraints, the restricted set of $\Delta$ in Assumption 2 for which the minimum is taken is smaller than that of the regular group lasso estimator. As such, the condition for the constrained model becomes weaker in general. Also, in Theorem 1 the choice of $\lambda$, which directly impacts the final estimation error rate, is taken as a minimal value over $r$, the choice of the baseline. Therefore, in view of the RE condition and the choice of $\lambda$, our results reveal that the proposed method is capable of achieving the best possible performance of the baseline method under a possibly weaker condition.

ASSUMPTION 3 ($\beta$-min Condition).    Choose the same $\lambda$ as in Theorem 1. Assume that

$$\min_{j \in \mathcal{S}} \|\beta_j^*\| > \frac{16\lambda s^*}{\kappa^2}.$$

COROLLARY 2 (Selection Consistency).    *Suppose Assumptions 1–3 hold. Let*

$$\widehat{\mathcal{S}} = \left\{ j : \|\widehat{\beta}_j\| > \frac{8\lambda s^*}{\kappa^2} \right\}.$$

*Then, with probability at least $1 - 2(p-1)^{1-q}$, we have that $\widehat{\mathcal{S}} = \mathcal{S}$.*

Corollary 2 reveals the "overselection" phenomenon due to convex penalization; see, for example, Wei and Huang (2010). That is, the proposed constrained group lasso estimator in general does not miss important variable groups/components, albeit overselecting some irrelevant ones. As such, a thresholding operation is preferred in order to recover the correct sparsity pattern exactly. However, the theoretical threshold is not available in practice, as it involves unknown quantities such as $\sigma^2$ and $\kappa$. Nevertheless, the results provide guarantee that using the original estimator can avoid false negatives at the expense of some false positives which is acceptable in many applications.

**6. Simulation.**    We conduct simulation studies to compare the performance of our proposed sparse functional log-contrast regression via constrained group lasso (CGL) in (9), the baseline approach in the form of (12) via group lasso (BGL) in which the reference level is chosen randomly, the naive approach of group lasso (GL) in which the zero-sum constraints are ignored in (9), cross-sectional method (I) of taking average of observations along time (Average) and cross sectional method (II) of considering the snapshot of the most significant time point (Snapshot).

The compositional data are generated as follows. We first generate $M$ time points within the interval $[0, 1]$, that is, $0 = t_1 < \cdots < t_M = 1$. For inducing dependence between time points, we consider an autoregressive correlation structure, $\Sigma_T = [\rho_T^{|\mu-\nu|}]_{M \times M}$, where $1 \leq \mu, \nu \leq M$; for inducing dependence between compositions, we consider a compound symmetry correlation structure, $\Sigma_X = [\rho_X^{I(j=j')}]_{p \times p}$, where $1 \leq j, j' \leq p$ and $I(\cdot)$ is the indicator function. The "nonnormalized" data for each subject $i$, $i = 1, \ldots, n$, are then generated from multivariate normal distribution as $\mathbf{w}_i = [\mathbf{w}_i(t_1)^{\mathrm{T}}, \ldots, \mathbf{w}_i(t_M)^{\mathrm{T}}]^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \sigma_X^2(\Sigma_T \otimes \Sigma_X))$, where each $\mathbf{w}_i(t_\nu) \in \mathbb{R}^p$ for $\nu = 1, \ldots, M$. Finally, the compositional data are obtained as $x_{ij}(t_\nu) = \exp(w_{ij}(t_\nu)) / \sum_{j=1}^p \exp(w_{ij}(t_\nu))$, for $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $\nu = 1, \ldots, M$. The regression curves $\beta^*(t)$ are generated as $\mathbf{B}^*\mathbf{\Phi}(t)$, where $\Psi(t)$ is from a set of cubic spline basis computed using the bs function in the R package splines with $t \in \{t_1, \ldots, t_M\}$ and degrees of freedom set to 5. The first three rows of $\mathbf{B}^*$ are set as $[1, 0, 1, 0, -0.5]$, $[0, 0, -1, 0, 1]$ and $[-1, 0, 0, 0, -0.5]$, respectively, and the rest are

set to zero. The intercept is set to be $\beta_0^* = 1$, and, for simplicity, we do not consider additional control. The error terms are generated as independent $N(0, \sigma^2)$ random variables where $\sigma^2$ is set to control the signal to noise ratio (SNR). Finally, the response $\mathbf{y}$ is generated by model (5), where the integral is computed as in (10). We have experimented with $(n, p) \in \{(50, 30), (100, 30), (100, 100), (100, 200)\}$ and parameter settings $M = 20$, $\sigma_X^2 = 9$, $\rho_T \in \{0, 0.6\}$, $\rho_X \in \{0, 0.6\}$ and SNR $= \{2, 4\}$. The simulation is repeated 100 times under each setting.

The prediction error (Pred) is measured by $\|\mathbf{y}_{te} - \mathbf{Z}_{te}\widehat{\beta}\|^2 / n_{te}$, computed from an independently generated test sample $(\mathbf{y}_{te}; \mathbf{X}_{te}(t), t \in \{t_1, \ldots, t_M\})$ of size $n_{te} = 500$. The estimation error (Est) is measured by $\sum_{j=1}^{p} (\int_{[0,1]} |\widehat{\beta}_j(t) - \beta_j^*(t)|^2 dt)^{1/2} / p$. For variable selection of the compositional components, we report the false positive rate (FPR) and the false negative rate (FNR), based on the sparsity patterns of $\widehat{\beta}(t)$ and $\beta^*(t)$. We have experimented with both 10-fold cross-validation (CV) and GIC to select from a grid of $(k, \lambda)$ values. As shown in Corollary 2, a thresholding of the estimator is preferred for the purpose of variable selection, although the ideal threshold is not available in practice. Here, with the same spirit and based on empirical evidence, we define the selected index set $\widehat{S}$ based on the relative magnitudes of the $p$ estimated coefficient curves,

$$\widehat{S} = \left\{ j; \left( \int_{[0,1]} \widehat{\beta}_j^2(t) \, dt \right)^{1/2} \Big/ \left\{ \sum_{j=1}^{p} \left( \int_{[0,1]} \widehat{\beta}_j^2(t) \, dt \right)^{1/2} \right\} \geq 1/p, j = 1, \ldots, p \right\}.$$

That is, we only count the components whose relative "energy" exceeds the average $1/p$ as selected.

The simulation results for $(n, p) = (50, 30)$ and $(n, p) = (100, 200)$ with SNR $= 4$ are reported in Tables 1–2. The two naive methods, Average and Snapshot, perform much worse in prediction than other methods; they tend to miss important variables, as seen from their high FNR values. To save space, we do not show the results of Average and Snapshot with GIC tuning, and we omit the results on the estimation errors as they deliver very similar message as the prediction errors. In general, CGL shows better predictive and selection performance than both GL and BGL, and in some cases the improvement can even be substantial. We have also tried the unpenalized least squares estimator which fails miserably in prediction and, hence, is omitted. The BGL method performs the worst among the three. The two tuning methods, CV and GIC, show quite different behaviors: the former generally yields larger false positive rates and much smaller false negative rates than the latter. Indeed, this is consistent with the theoretical results in Section 5 that the proposed convex regularized estimation approach has a screening property when tuned based on optimizing predictive performance, that is, with high probability the method does not miss any important variables but, consequently, it has a tendency of overselection. Nevertheless, the CV-tuned estimators rarely miss important components and perform much better in prediction compared to their GIC-tuned counterparts. Therefore, CV may be preferable in practice when one cares more about prediction and can afford some false alarms for the capture of all the relevant signals.

Figure 2 shows boxplots of prediction errors from CV tuning for various simulation settings. We do not include Average and Snapshot methods as they perform much worse; more detailed results of SNR $= 2$ is reported in Section 4 of the Supplementary Material (Sun et al. (2020)). The performance of all methods deteriorates when the SNR becomes smaller, the between-component correlation becomes smaller or the between-time correlation becomes stronger. Small between-component correlation causes the presence of a few dominating compositional components due to the unit-sum constraints, while large between-time correlation makes the functional compositions smooth over time and, consequently, makes it hard to distinguish the relevant components from the others.

TABLE 1
*Simulation results for* $(n, p) = (50, 30)$ *and* SNR $= 4$. *Reported are the average values over* 100 *simulation runs with the standard deviations in parentheses. For better presentation, the values of Pred are multiplied by* 10

| $(\rho_X, \rho_T)$ | Criterion | Method | Pred | FPR (%) | FNR (%) |
|---|---|---|---|---|---|
| (0, 0) | CV | BGL | 0.39 (0.01) | 28.85 (1.28) | 0.00 (0.00) |
| | | GL | 0.39 (0.01) | 27.48 (1.35) | 0.00 (0.00) |
| | | CGL | 0.34 (0.01) | 29.22 (1.43) | 0.00 (0.00) |
| | | Average | 2.03 (0.03) | 12.00 (1.37) | 66.00 (3.45) |
| | | Snapshot | 2.11 (0.05) | 16.74 (1.77) | 53.00 (3.45) |
| | GIC | BGL | 1.46 (0.06) | 4.04 (0.19) | 48.00 (3.33) |
| | | GL | 1.44 (0.05) | 0.19 (0.08) | 52.67 (2.69) |
| | | CGL | 1.24 (0.05) | 1.63 (0.24) | 20.00 (2.37) |
| (0, 0.6) | CV | BGL | 1.27 (0.04) | 30.70 (1.48) | 0.33 (0.33) |
| | | GL | 1.21 (0.03) | 29.04 (1.40) | 0.00 (0.00) |
| | | CGL | 1.13 (0.03) | 29.67 (1.43) | 0.00 (0.00) |
| | | Average | 5.58 (0.14) | 19.67 (1.96) | 34.00 (3.45) |
| | | Snapshot | 5.31 (0.10) | 23.44 (1.60) | 22.67 (1.83) |
| | GIC | BGL | 4.61 (0.16) | 3.74 (0.16) | 52.67 (2.60) |
| | | GL | 3.93 (0.12) | 0.11 (0.06) | 51.67 (2.39) |
| | | CGL | 3.91 (0.17) | 1.52 (0.24) | 23.67 (2.19) |
| (0.6, 0) | CV | BGL | 0.15 (0.01) | 29.26 (1.35) | 0.00 (0.00) |
| | | GL | 0.16 (0.00) | 29.93 (1.42) | 0.00 (0.00) |
| | | CGL | 0.14 (0.00) | 29.07 (1.22) | 0.00 (0.00) |
| | | Average | 0.80 (0.01) | 14.63 (1.70) | 57.33 (3.52) |
| | | Snapshot | 0.85 (0.02) | 16.70 (1.73) | 57.00 (3.29) |
| | GIC | BGL | 0.65 (0.02) | 3.81 (0.19) | 56.33 (2.67) |
| | | GL | 0.62 (0.02) | 0.19 (0.08) | 59.33 (2.25) |
| | | CGL | 0.54 (0.02) | 1.63 (0.22) | 22.67 (2.22) |
| (0.6, 0.6) | CV | BGL | 0.53 (0.02) | 33.52 (1.38) | 0.33 (0.33) |
| | | GL | 0.49 (0.02) | 30.22 (1.31) | 0.00 (0.00) |
| | | CGL | 0.45 (0.01) | 30.37 (1.44) | 0.00 (0.00) |
| | | Average | 2.02 (0.04) | 22.81 (1.89) | 26.33 (2.81) |
| | | Snapshot | 2.10 (0.03) | 22.85 (1.60) | 25.67 (1.76) |
| | GIC | BGL | 1.85 (0.06) | 3.81 (0.15) | 53.67 (2.59) |
| | | GL | 1.69 (0.05) | 0.11 (0.06) | 57.67 (2.00) |
| | | CGL | 1.52 (0.06) | 1.74 (0.23) | 25.00 (2.24) |

As suggested by a referee, we also perform a simulation example to better mimic the real data from the preterm infant study, in which the observed compositional data matrix from the preterm infant study is directly used as the design matrix to generate data, the sample size is the same as the real data ($n = 34$, $p = 22$) and the SNR is set as comparable to the real data (SNR $= 1.5$). The nonzero coefficients of the taxa are either set the same as those used in the other simulation examples or those estimated from CGL on the real data. The simulation is repeated 100 times. The average scaled prediction error is 3.10 (0.03), 3.10 (0.03) and 2.99 (0.02) in the first setting and 1.88 (0.02), 1.37 (0.01) and 1.33 (0.01) in the second setting, for BGL, GL and CGL tuned by CV, respectively. This result further confirms that CGL can lead to slightly improved model performance while maintaining the simplex geometry for ease of model interpretation.

**7. Linking microbiome trajectories to neurobehavioral outcomes.** Recall that our main objective is to identify the microbiome markers that are predictive of later infant neu-

TABLE 2
*Simulation results for $(n, p) = (100, 200)$ and* SNR $= 4$. *The layout is the same as in Table* 1

| $(\rho_X, \rho_T)$ | Criterion | Method | Pred | FPR (%) | FNR (%) |
|---|---|---|---|---|---|
| (0, 0) | CV | BGL | 0.31 (0.01) | 15.28 (0.48) | 0.00 (0.00) |
| | | GL | 0.31 (0.01) | 15.27 (0.48) | 0.00 (0.00) |
| | | CGL | 0.29 (0.00) | 15.57 (0.51) | 0.00 (0.00) |
| | | Average | 1.98 (0.03) | 3.04 (0.41) | 73.33 (3.11) |
| | | Snapshot | 1.99 (0.03) | 4.82 (0.64) | 59.33 (3.20) |
| | GIC | BGL | 1.45 (0.05) | 0.51 (0.01) | 44.00 (3.07) |
| | | GL | 1.33 (0.05) | 0.01 (0.01) | 46.33 (2.88) |
| | | CGL | 1.13 (0.05) | 0.19 (0.03) | 11.67 (1.73) |
| (0, 0.6) | CV | BGL | 1.02 (0.02) | 16.26 (0.51) | 0.00 (0.00) |
| | | GL | 0.97 (0.02) | 15.62 (0.52) | 0.00 (0.00) |
| | | CGL | 0.94 (0.02) | 16.32 (0.50) | 0.00 (0.00) |
| | | Average | 5.41 (0.10) | 7.17 (0.70) | 27.00 (2.71) |
| | | Snapshot | 5.14 (0.10) | 6.57 (0.56) | 27.67 (1.26) |
| | GIC | BGL | 4.15 (0.16) | 0.51 (0.01) | 43.00 (3.01) |
| | | GL | 3.44 (0.12) | 0.01 (0.01) | 42.67 (2.92) |
| | | CGL | 3.57 (0.15) | 0.10 (0.02) | 16.00 (1.92) |
| (0.6, 0) | CV | BGL | 0.12 (0.00) | 14.78 (0.49) | 0.00 (0.00) |
| | | GL | 0.12 (0.00) | 15.44 (0.61) | 0.00 (0.00) |
| | | CGL | 0.12 (0.00) | 15.07 (0.55) | 0.00 (0.00) |
| | | Average | 0.80 (0.01) | 5.14 (0.68) | 58.33 (3.80) |
| | | Snapshot | 0.81 (0.01) | 4.30 (0.49) | 61.67 (2.82) |
| | GIC | BGL | 0.55 (0.02) | 0.53 (0.01) | 39.00 (3.39) |
| | | GL | 0.47 (0.02) | 0.02 (0.01) | 36.33 (3.22) |
| | | CGL | 0.41 (0.02) | 0.15 (0.03) | 9.67 (1.79) |
| (0.6, 0.6) | CV | BGL | 0.41 (0.01) | 16.21 (0.50) | 0.00 (0.00) |
| | | GL | 0.40 (0.01) | 15.30 (0.55) | 0.00 (0.00) |
| | | CGL | 0.39 (0.01) | 15.59 (0.50) | 0.00 (0.00) |
| | | Average | 2.03 (0.03) | 7.35 (0.63) | 16.67 (2.25) |
| | | Snapshot | 2.09 (0.04) | 7.62 (0.67) | 27.33 (1.29) |
| | GIC | BGL | 1.76 (0.06) | 0.52 (0.01) | 48.33 (2.93) |
| | | GL | 1.46 (0.05) | 0.01 (0.01) | 47.33 (2.73) |
| | | CGL | 1.40 (0.06) | 0.15 (0.03) | 17.00 (1.98) |

rodevelopment as measured by NNNS. This predictive association, if proven true, can provide supporting evidence to the claim that the stressful early life experience of preterm infants is imprinting gut microbiome by the regulation of the brain-gut axis. We tackle the problem with the functional log-contrast regression model in (5), in which the composite NSTRESS score serves as the response variable, the gut microbiome observed during the early postnatal period serves as the functional compositional predictors and the infant characteristics listed in Table 3 below serve as the time-invariate control variables. We apply the proposed CGL approach for model estimation and compositional component selection. The cubic spline basis is used, and the tuning of the degrees of freedom $k$ as well as the sparsity parameter $\lambda$ is done using cross-validation.

After controlling for the effects of several infant characteristics, our approach is able to identify four bacteria categories at the order level that are associated with the neurobehavioral outcome of infant. (GIC tuning selects exactly the same four taxa.) Before we discuss the selected microbiome markers, let's first focus on the effects of the control variables. Table 3 shows the estimated coefficients of the control variables along with some descriptive
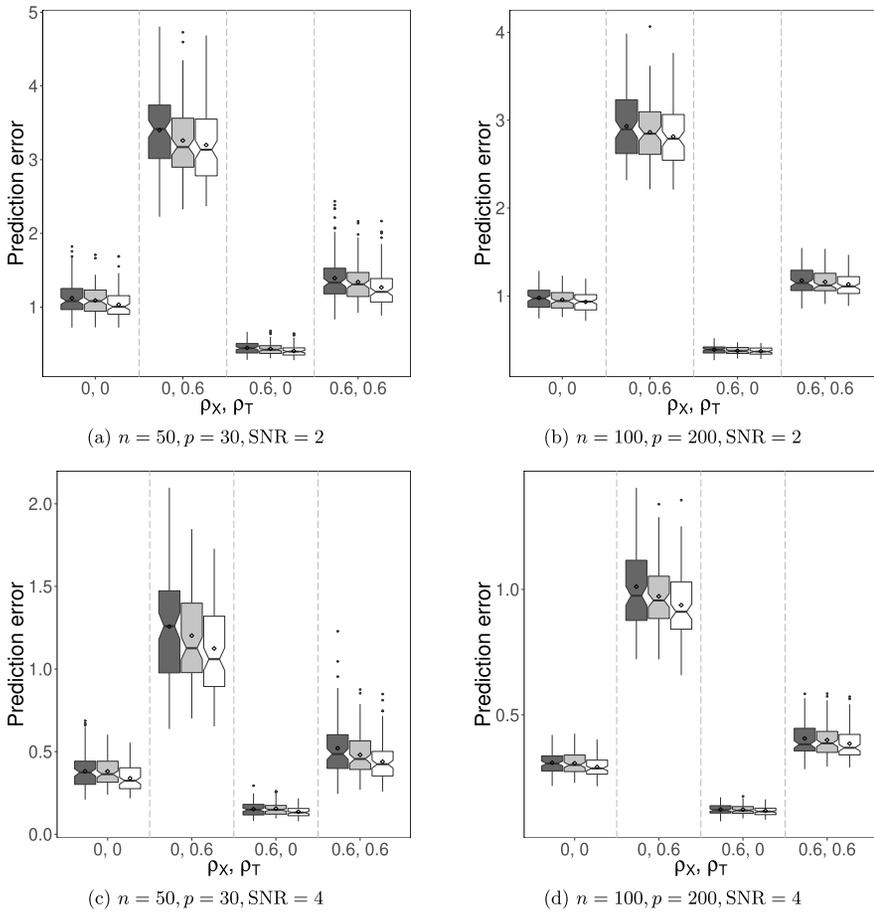
FIG. 2. *Boxplots of prediction errors for various simulation settings. The dark grey, light grey and white colors correspond to three different estimation methods BGL, GL and CGL, respectively.*

statistics. It is seen that the neurobehavioral outcome is better (i.e., NSTRESS is small) for infants with larger birth weight, smaller SNAPE-II score and more mother's breast milk for feeding. Regarding the delivery of infant, vaginal delivery and the absence of premature rupture of membranes are associated with better neurobehavioral development. These interesting

TABLE 3
*Descriptive statistics of infant characteristics and their estimated coefficients from fitting the sparse functional log-contrast regression. Values of estimated coefficient are multiplied by* 100

| Numerical variable | Mean (sd) | Estimated coefficient |
|---|---|---|
| Birth weight (in gram) | 1451.7 (479.3) | −0.003 |
| SNAPE-II | 9.3 (10.6) | 0.107 |
| %MBM | 61.8 (29.9) | −9.272 |
| | | |
| Binary variable | Mean | |
| Gender (female = 1) | 50.0% | −0.064 |
| PROM (yes = 1) | 44.1% | 2.761 |
| Delivery type (vaginal = 1) | 35.3% | −5.105 |

(a) Category 1

(b) Category 9

(c) Category 10
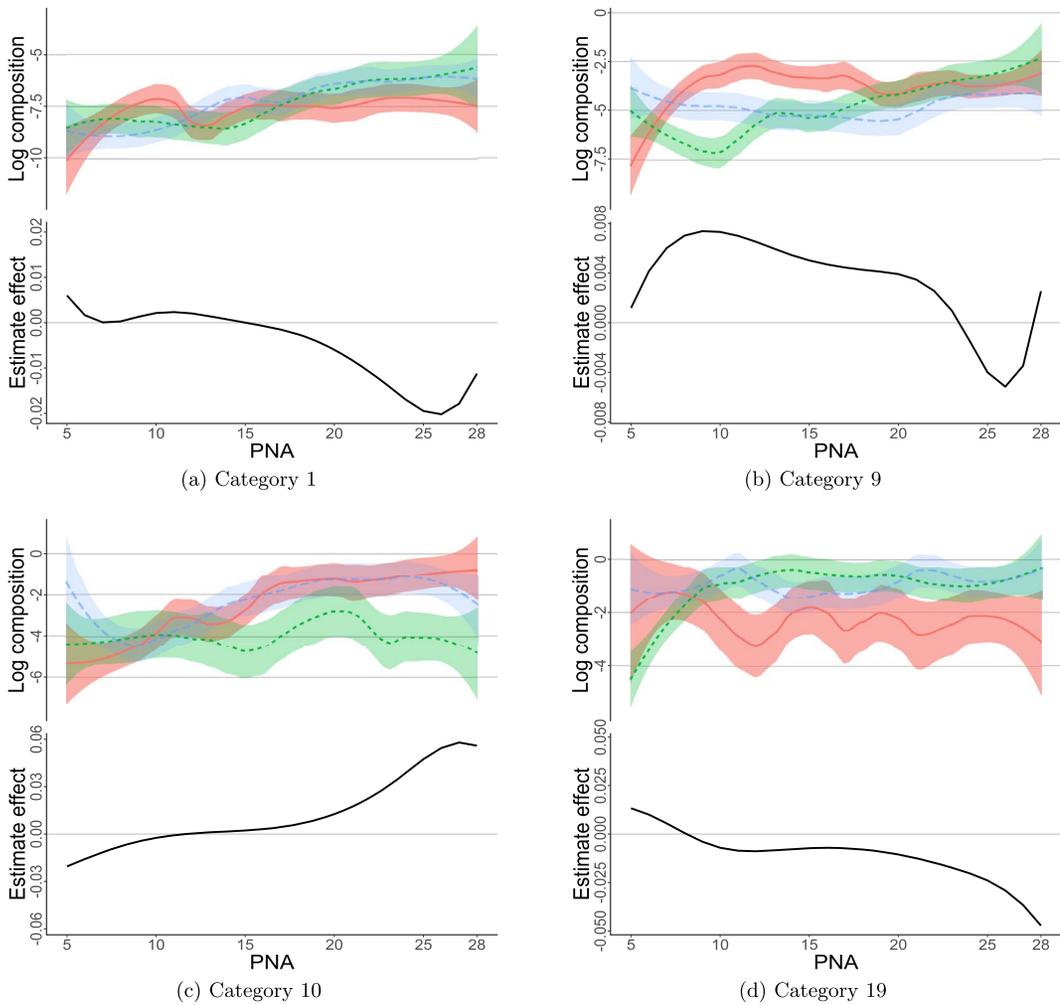
(d) Category 19

FIG. 3.    *Estimated effects of the four selected bacteria categories at the order level over infant's postnatal age (PNA) of five to 28 days. In each subgraph the upper panel shows how this category changes over time for three clusters of infants. For each category the clusters are based on the percentiles of its partial residuals, obtained by subtracting the estimated effects of the control variables and other selected bacteria categories from the observed NSTRESS scores. The curve with its 90% confidence band is shown in red for the high group, in blue for the medium group and in green for the low group.*

and intuitive results are consistent with existing literature (Neu and Rushing (2011), Feldman and Eidelman (2003)). The analysis also shows that female infants tend to perform slightly better than male, after accounting for other effects.

The estimated functional effects of the four selected bacteria categories are shown in the four panels of Figure 3, respectively. In each panel the lower part shows the estimated functional effects of a category over time (between five and 28 days of postnatal age), and the upper part attempts to show, directly from raw data, how this category changes over time for infants with high, medium or low "adjusted" NSTRESS score, obtained by subtracting the estimated effects of the control variables and other selected bacteria categories from the observed NSTRESS scores. Specifically, we construct smoothed curves of log-compositions of each selected category for three clusters of infants (using locally weighted scatterplot smoothing). For each category the clusters are based on the percentiles of its "adjusted" NSTRESS score. The curve with its 90% confidence band is shown in red for the high group, that is, infants with the upper one-third of the adjusted scores, in blue for the medium group, that
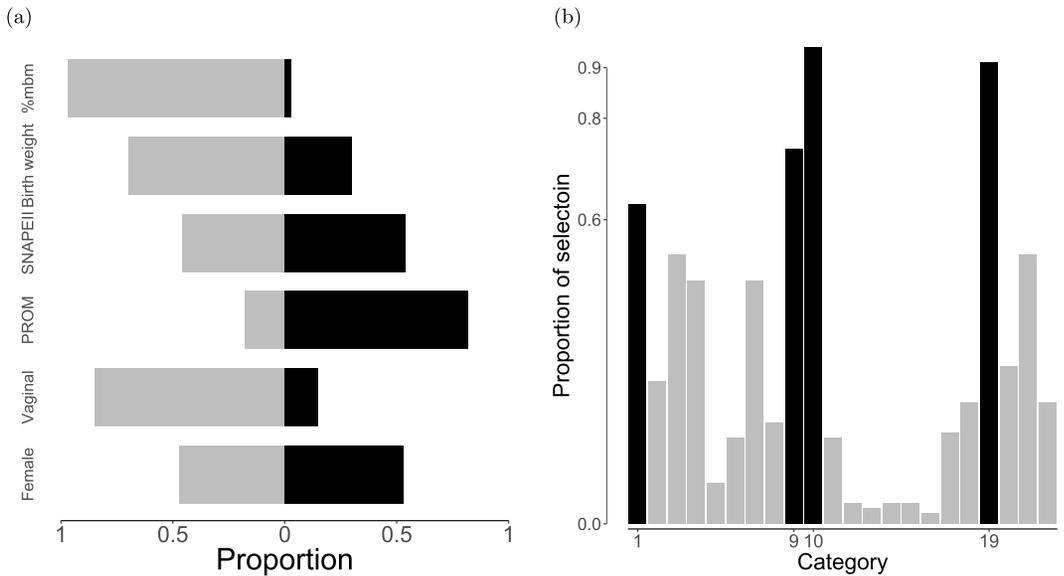
FIG. 4.   *Selection results from* 100 *bootstrap samples.* (*a*) *Proportions of the signs of the estimated coefficients of the control variables. Proportions of positive signs are shown as black blocks to the right, and those of negative signs are shown as light gray blocks to the left.* (*b*) *Proportions of selecting the* 22 *bacteria categories at the order level. The bars of the four selected categories from fitting the original data are colored in black.*

is, infants with the middle one-third of the adjusted scores and in green for the low group, that is, infants with the lower one-third of the adjusted scores. As an example, for category 1 the red curve increases in the beginning to be above the other two curves and then becomes mostly below them in the later stage. This suggests that the time-varying effect of category 1 on the NSTRESS score is first positive and then negative, which is clearly reflected by the estimated functional effects. Similarly, for the other three selected categories the patterns of the estimated effects agree well with those of the observed data. This verifies visually that our proposed model and the estimation approach yield sensible results.

To access the stability of the results, we have generated 100 bootstrap samples and used the same cross-validation procedure to select the best models. The results are show in Figure 4. The signs of the coefficients of the control variables are quite stable, except for the gender and SNAPE-II variables; this shows that these two variables may not have much effect on the outcome when conditioning on other terms in the model. For each control variable the sign with the higher proportion among its 100 bootstrap estimates agrees with that of the estimate from fitting the original data, except for the gender. Furthermore, the top four categories with the highest proportions of being selected in bootstrap coincide with the categories selected from fitting the original data. Categories 10 and 19 are selected about 90% of the times, while 9 and 1 are selected more than 70% and 60% of the times, respectively.

Category 10 consists of Clostridiales, which are an order of bacteria belonging to the phylum Firmicutes. Studies showed that infants fed with mother's milk had significantly higher abundance in Clostridiales (Cong (2016)). Clostridiales are generally regarded as hallmarks of a healthy gut; it can be a sign of infection when their subtypes such as Eubacteria die off in the large intestine. Our results show that controlling for other effects in the model, the effect of Clostridiales on the stress score switches from negative to positive during the postnatal days from five to 28. Category 9 consists of Lactobacillales, or lactic acid bacteria (LAB), another order of bacteria belonging to the phylum Firmicutes. These bacteria are usually found in decomposing plants and milk products; they are considered beneficial and produce organic acids, such as lactic acid from carbohydrates. Our analysis shows that controlling for the other effects in the model, higher LAB proportions are associated with higher

stress scores for a period of time during the early postnatal days. Both Clostridiales and LAB belong to phylum Firmicutes, which make up the largest portion of the human gut microbiome, and the abundance of Firmicutes has been shown to be associated with inflammation and obesity (Boulangé et al. (2016), Clarke et al. (2012)). Category 19 consists of Enterobacteriales, an order of gram-negative bacteria. They are responsible for various infections, such as bacteremia, lower respiratory tract infections, skin infections, etc. Category 1 consists of other unclassified bacteria. The functional regression analysis presented here may lead to a better understanding of how the trajectories of gut microbiome during early postnatal stage impact neurobehavioral outcomes of infants through the gut-brain axis.

We also repeat the analysis on a lower level taxon, that is, the genus level. Five out of $p = 62$ genera are selected, and their estimated functional effects are shown in the five panels of Figure 5. The tendency of the estimated effects adequately reflects those of the observed data, and the results are consistent with previous study on the order level. In particular, the five selected genera all belong to the four selected order categories; see Table 4. Genus 38 comprises genus Veillonella, belonging to the order Clostridiales. Veillonella have been implicated as pathogens; they are often associated with oral, central nervous system and various soft tissue infections. Our results show that controlling for the other effect in the model, the effect of Veillonella on the stress score works similarly to that of Clostridiales, switching from negative to positive. Genus 20 consists of Enterococcus which is an large genus of bacteria belonging to the order LAB. In humans, E. faecalis and E. faecium are the most abundant species of this genus found in fecal content, comprising up to 1% of the adult intestinal microbiota. Although utilization of Enterococci as probiotics has been under controversial discussion, enterococcal strains such as E. faecium SF68 and E. faecalis Symbio-flor have been marketed as probiotics for decades without incidence and with very few reported adverse events (Franz et al. (2011)). On the other aspect, Enterococci is also important nosocomial pathogens that cause bacteraemia, endocarditis and other infections. Same as LAB, controlling for the other effect in the model, our results show that higher Enterococcus proportions are associated with higher stress scores for a period of time during the early postnatal days. Genus 55 is Shigella, belonging to the order Enterobacteriales. Shigella is considered as pathogen causing shigellosis. The main sign of shigella infection is diarrhea which often is bloody. However, shigellosis rarely affects infants during the first month of life. Even in highly endemic areas, neonatal shigellosis is exceedingly uncommon (Haltalin (1967)). Our analysis shows that controlling for the other effects in the model, the effect of Shigella changes from positive to negative during early postnatal days. Genus 48 consists of other unclassified genera of bacteria that belong to the order Enterobacteriales. Genus 1 consists of other unclassified bacteria.

Lastly, we compare the three functional estimation methods BGL, GL and CGL in this real data analysis. Recall that using full data, CGL selected orders 1, 9, 10 and 19. In contrast, GL selected orders 9, 10 and 19 but missed order 1. The selected taxa using BGL varied drastically with different choices of the baseline taxon; for example, it selected order 10 when order 1 was used as the baseline, it selected 9, 10, 14, 15 and 17 when order 12 was used as the baseline and it failed to select any taxon when order 2 was used as the baseline. To objectively compare the three methods, we have also performed an out-of-sample random splitting procedure. Specifically, we randomly divide the data into a training sample of 90% subjects to fit the models and a testing sample of 10% patients to evaluate the out-of-sample performance of the models. This procedure is repeated 100 times, and the results are averaged. For the order level analysis, the average scaled prediction errors are 0.213 (0.478), 0.136 (0.123) and 0.127 (0.095), for BGL, GL and CGL, respectively. The results confirm that CGL is preferred in practice due to its improved predictive performance and its validity in maintaining the log-contrast model interpretation and avoiding arbitrary selection of the baseline. Indeed, here the baseline method, BGL, performs much worse than the other two methods and exhibits a much larger variation in selection.
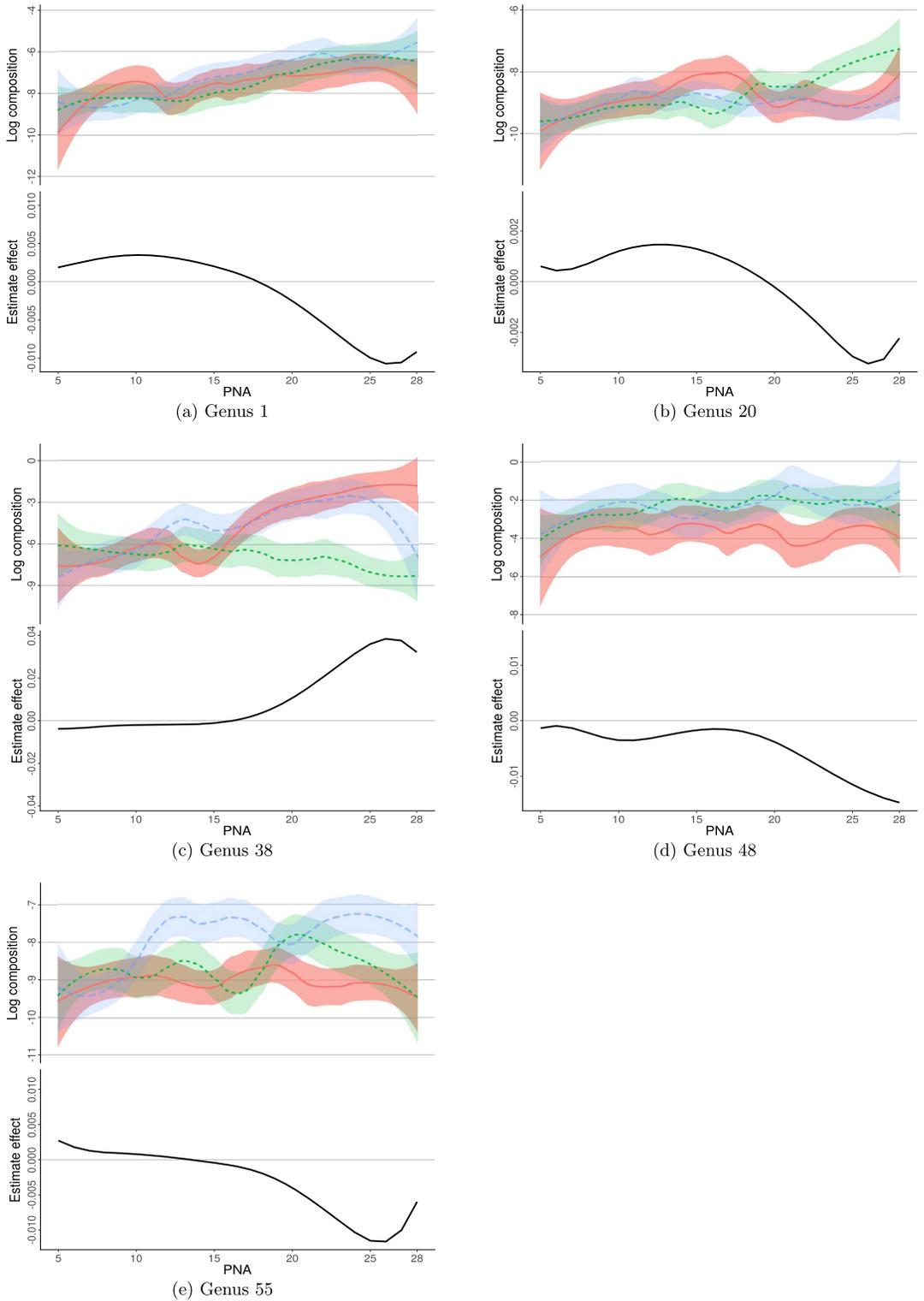
FIG. 5. *Estimated effects of the five selected bacteria categories at the genus level over infant's postnatal age* (*PNA*) *of five to* 28 *days. The layout is the same as in Figure* 3.

TABLE 4
*Comparison of selection of microbiome markers between order level and genus level*

| Order level | Genus level |
| --- | --- |
| 1: Others | 1: Others |
| 9: Lactobacillales<br>Produce organic acids such as lactic acid from carbohydrates. | 20: Enterococcus<br>It's used as probiotics in humans;<br>It's considered as pathogens that cause bacteraemia, endocarditis and other infections. |
| 10: Clostridiales<br>It's generally regarded as hallmarks of a healthy gut;<br>It's a sign of infection when their subtypes such as Eubacteria die off in the large intestine. | 38: Veillonella<br>It's implicated as pathogens;<br>It's associated with oral infections and various soft tissue infections. |
| 19: Enterobacteriales<br>It's responsible for various infections such as bacteremia, lower respiratory tract infections, skin infections, etc. | 55: Shigella<br>It's considered as pathogen causing shigellosis;<br>Shigellosis is exceedingly uncommon for infants during the first month of life<br>48: Others |

**8. Discussion.** We have attempted a functional log-contrast regression approach to identify trajectories of gut microbiome components during early postnatal stage that are associated with later neurobehavioral outcomes of preterm infants. There are several directions for future research to address the limitations of the current work.

Our results on order and genus levels give a general idea of how the microbial communities effect health outcomes; to fully decipher their roles, further analysis on species level or even operational taxonomic unit (OTU) is needed. However, the problem of choosing microbiome resolution (or taxonomic rank) in data analysis is both biologicial and statistical. In this work our choices are made based on both the interest of the biologists and by taking into consideration of the limited sample size of $n = 34$; while we focus on the order level with $p = 22$, the additional analysis on the genus level, which has a much larger number of predictors, provides a sanity check. We suggest the practitioners to do the same. For the compositional data on microbiomes, there is clearly a trade-off between data quality and data resolution— the lower the rank, the higher the data resolution (there are more taxa) but the sparser the data (there are more zeros and each composition is converted from a smaller count). More generally, as microbiomes admit a taxonomic hierarchy, a fundamental statistical problem is how to search for the taxonomic rank or ranks that which best predicts an outcome of interest.

The data analysis can benefit from extending the model to consider potential interactions between the control variables and the gut microbiome, as it is possible, for example, that the effects of certain microbiome markers differ for male and female infants. Our method is based on constrained and regularized estimation, with which it is not straightforward to produce valid confidence intervals. It is thus pressing to tackle this postselection inference problem (Shi, Zhang and Li (2016), Taylor and Tibshirani (2015)) in the context of functional regression. Extensions to binary outcome or mixture model setup are interesting and could be widely applicable; indeed, it is of interest to see whether there exists a subgroup structure among the preterm infants. To take into account the uncertainty due to discrete observations, it is urgent to develop smoothing or dimension reduction methods, such as multivariate functional principal component analysis for compositional data observed discretely over time. A joint modeling approach of both the regression and the functional compositions themselves may also be fruitful.

## SUPPLEMENTARY MATERIAL

**Supplement to "Log-contrast regression with functional compositional predictors: Linking preterm infant's gut microbiome trajectories to neurobehavioral outcome"** (DOI: 10.1214/20-AOAS1357SUPP; .pdf). We provide details of the computational algorithm in Section 4, proofs of Theorem 1 and Corollary 2, more details on the microbiome data in the preterm infant study, and additional simulation results.

## REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. With discussion. MR0676206

AITCHISON, J. (2003). *The Statistical Analysis of Compositional Data*. Blackburn Press, NJ. MR0865647 https://doi.org/10.1007/978-94-009-4109-0

AITCHISON, J. and BACON-SHONE, J. (1984). Log-contrast models for experiments with mixtures. *Biometrika* **71** 323–330.

AITCHISON, J. and EGOZCUE, J. J. (2005). Compositional data analysis: Where are we and where should we be heading? *Math. Geol.* **37** 829–850. MR2183640 https://doi.org/10.1007/s11004-005-7383-7

BOMAR, L., MALTZ, M., COLSTON, S. and GRAF, J. (2011). Directed culturing of microorganisms using metatranscriptomics. *mBio* **2** e00012-11. https://doi.org/10.1128/mBio.00012-11

BOULANGÉ, C. L., NEVES, A. L., CHILLOUX, J., NICHOLSON, J. K. and DUMAS, M.-E. (2016). Impact of the gut microbiota on inflammation, obesity, and metabolic disease. *Gen. Med.* **8** 42. https://doi.org/10.1186/s13073-016-0303-2

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Foundations and Trends in Machine Learning* **3**, 1. https://doi.org/10.1561/2200000016

CLARKE, S. F., MURPHY, E. F., NILAWEERA, K., ROSS, P. R., SHANAHAN, F., O'TOOLE, P. W. and COTTER, P. D. (2012). The gut microbiota and its relationship to diet and obesity: New insights. *Gut Microbes* **3** 186–202. https://doi.org/10.4161/gmic.20168

CONG, X. (2016). Gut microbiome developmental patterns in early life of preterm infants: Impacts of feeding and gender. *PLoS ONE* **11** 1–19.

CONG, X., JUDGE, M., XU, W. and DIALLO, A. (2017). Influence of feeding type on gut microbiome development in hospitalized preterm infants. *Nursing Research* **66** 123–133.

DINAN, T. G. and CRYAN, J. F. (2012). Regulation of the stress response by the gut microbiota: Implications for psychoneuroendocrinology. *Psychoneuroendocrinology* **37** 1369–1378.

FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. MR3065478 https://doi.org/10.1111/rssb.12001

FANAROFF, A. A., HACK, M. and MC, W. (2003). The NICHD neonatal research network: Changes in practice and outcomes during the first 15 years. *Semin. Perinatol.* **27** 281–287.

FELDMAN, R. and EIDELMAN, A. I. (2003). Direct and indirect effects of breast milk on the neurobehavioral and cognitive development of premature infants. *Dev. Psychobiol.* **43** 109–119.

FRANZ, C. M., HUCH, M., ABRIOUEL, H., HOLZAPFEL, W. H. and GÁLVEZ, A. V. (2011). Enterococci as probiotics and their implications in food safety. *Int. J. Food Microbiol.* **151 2** 125–40.

HALTALIN, K. C. (1967). Neonatal shigellosis: Report of 16 cases and review of the literature. *JAMA Pediatrics* **114** 603–611.

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* **2** 73–94.

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797. MR3286917 https://doi.org/10.1093/biomet/asu031

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865 https://doi.org/10.1214/11-AOS896

MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359. https://doi.org/10.1146/annurev-statistics-010814-020413

MWANIKI, M. K., ATIENO, M., LAWN, J. E. and NEWTON, C. R. (2012). Long-term neurodevelopmental outcomes after intrauterine and neonatal insults: A systematic review. *Lancet* **379** 445–452.

NEU, J. and RUSHING, J. M. (2011). Cesarean versus vaginal delivery: Long-term infant outcomes and the hygiene hypothesis. *Clin. Perinatol.* **38 2** 321–31.

R CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2168993

SHI, P., ZHANG, A. and LI, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **10** 1019–1040. MR3528370 https://doi.org/10.1214/16-AOAS928

STOLL, B. J., HANSEN, N. I. and BELL, E. F. (2010). Neonatal outcomes of extremely preterm infants from the NICHD neonatal research network. *Pediatrics* **126** 443–456.

STONE, M. (1974). Cross-validation and multinomial prediction. *Biometrika* **61** 509–515. MR0415896 https://doi.org/10.1093/biomet/61.3.509

SUN, Z. and CHEN, K. (2020). Compack: Regression with compositional covariates. R package version 0.1.0.

SUN, Z., XU, W., CONG, X., LI, G. and CHEN, K. (2020). Supplement to "Log-contrast regression with functional compositional predictors: Linking preterm infants' gut microbiome trajectories to neurobehavioral outcome." https://doi.org/10.1214/20-AOAS1357SUPP

TAYLOR, J. and TIBSHIRANI, R. J. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA* **112** 7629–7634. MR3371123 https://doi.org/10.1073/pnas.1507583112

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

WEI, F. and HUANG, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli* **16** 1369–1384. MR2759183 https://doi.org/10.3150/10-BEJ252

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x