Received: 29 December 2020 | Revised: 25 September 2021 | Accepted: 15 November 2021

DOI: 10.1002/sim.9273

RESEARCH ARTICLE

Statistics in Medicine WILEY

Multivariate log-contrast regression with sub-compositional predictors: Testing the association between preterm infants' gut microbiome and neurobehavioral outcomes

Xiaokang Liu[1] | **Xiaomei Cong**[2] | **Gen Li**[3] | **Kendra Maas**[4] | **Kun Chen**[5]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]School of Nursing, University of Connecticut, Storrs, Connecticut, USA

[3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

[4]Microbial Analysis, Resources, and Services, University of Connecticut, Storrs, Connecticut, USA

[5]Department of Statistics, University of Connecticut, Storrs, Connecticut, USA

Correspondence
Xiaokang Liu, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, USA.
Email: xiaokang.liu@pennmedicine.upenn.edu

Funding information
National Institutes of Health National Institute of Nursing Research, Grant/Award Numbers: R01NR016928, K23NR014674; National Science Foundation, Grant/Award Numbers: DMS-1613295, IIS-1718798

To link a clinical outcome with compositional predictors in microbiome analysis, the linear log-contrast model is a popular choice, and the inference procedure for assessing the significance of each covariate is also available. However, with the existence of multiple potentially interrelated outcomes and the information of the taxonomic hierarchy of bacteria, a multivariate analysis method that considers the group structure of compositional covariates and an accompanying group inference method are still lacking. Motivated by a study for identifying the microbes in the gut microbiome of preterm infants that impact their later neurobehavioral outcomes, we formulate a constrained integrative multi-view regression. The neurobehavioral scores form multivariate responses, the log-transformed sub-compositional microbiome data form multi-view feature matrices, and a set of linear constraints on their corresponding sub-coefficient matrices ensures the sub-compositional nature. We assume all the sub-coefficient matrices are possible of low-rank to enable joint selection and inference of sub-compositions/views. We propose a scaled composite nuclear norm penalization approach for model estimation and develop a hypothesis testing procedure through de-biasing to assess the significance of different views. Simulation studies confirm the effectiveness of the proposed procedure. We apply the method to the preterm infant study, and the identified microbes are mostly consistent with existing studies and biological understandings.

KEYWORDS
compositional data, group inference, integrative multivariate analysis, multi-view learning, nuclear norm penalization

1 | INTRODUCTION

In recent years, there has been a dramatic increase in survival among preterm infants from 15% to over 90%[1,2] due to the advancement in neonatal care. However, studies showed that stressful early life experience, as exemplified by the accumulated stress and insults that the preterm infants encounter during their stay in neonatal intensive care units (NICU), could cause long-term adverse consequences for their neurodevelopmental and health outcomes. For example, Mwaniki et al[3] reported that close to 40% of NICU survivors had at least one neurodevelopmental deficit that may be attributed to stress/pain at NICU, caused by maternal separations, painful procedures, among others. As such, understanding the linkage between the stress/pain and the onset of the altered neuro-immune progress holds the key

to reduce health consequences of prematurity. This is permitted by the functional association between the central nervous system and gastrointestinal tract.[4] With the regulation of this "gut-brain axis," accumulated stress imprints on the gut microbiome compositions,[5,6] and thus the link between neonatal insults and neurological disorders can be approached through examining the association between preterm infants' gut microbiome compositions and their later neurodevelopment measurements.

To investigate the aforementioned problem, a preterm infant study was conducted in a NICU in the United States. Stable preterm infants were recruited, and fecal samples were collected during the infant's first month of postnatal age on a daily basis when available. From each fecal sample, bacterial DNA was isolated and extracted, and gut microbiome data were then obtained through DNA sequencing and data processing. Gender, delivery type, birth weight, feeding type, gestational age, and postnatal age were recorded for each infant. Infant neurobehavioral outcomes were measured when the infant reached 36 to 38 weeks of gestational age, using the NICU Network Neurobehavioral Scale (NNNS). More details on the study and data are provided in Section 5.

With the collected data, assessment of which microbes are associated with the neurobehavioral development of the preterm infants can be conducted through a statistical regression analysis, with the NNNS scores being the outcomes and the gut microbiome compositions as the predictors. There are several unique challenges in this problem. First, the NNNS consists of 13 sub-scales on various aspects of neurobehavioral development. As such, an overall assessment about whether the neurobehavioral development is impacted at all by the gut microbiome calls for a multivariate estimation and testing procedure that can utilize all the sub-scale scores simultaneously. Indeed, our preliminary analysis shows that these sub-scale scores are distinct yet interrelated. A multivariate procedure could result in more accurate estimation and more powerful tests than its univariate counterparts. Moreover, the candidate predictors constructed from the microbiome data are structurally very rich and complex: they are high-dimensional, compositional, and hierarchical. A compositional observation is a multivariate vector with elements being proportions, which are non-negative and satisfy the constraint that their summation is unity. In our problem, the data on bacterial taxa are presented as groups of sub-compositions in conformity with the taxonomic hierarchy of bacteria, that is, each taxon is represented by a group of compositions at a lower taxonomic rank. These unique features call for a tailored dimension reduction approach that allows high-dimensional inference to be made at the group level for testing each taxon component.

Compositional data analysis is of great importance in a broad range of scientific fields, including microbiology, ecology, and geology. The simplex and non-Euclidean structure of the data impedes the application of many classical statistical methods. Much foundational work on the treatment of compositional data was done by Aitchison.[7,8] In the regression realm, a foundational work is the linear log-contrast model;[9] in its symmetric form, the response is regressed on the logarithmic transformed compositional predictors and a zero-sum constraint is imposed on the coefficient vector to keep the simplex geometry. Compositional data on microbiome are often high-dimensional, as it is common that a sample could produce hundreds of operational taxonomic units. We refer to Li's work[10] for a recent comprehensive review on microbiome compositional data analysis. In particular, various sparsity-inducing penalized estimation methods were proposed to enable the selection of a smaller set of predictive compositions.[11-14] Shi et al[15] extended the sparse regression model to perform high-dimensional sub-compositional analysis, in which the predictors form several compositional groups according to the taxonomic hierarchy of the microbes; a de-biased estimation procedure was adopted to perform statistical inference. Another kind of regression methods conduct sufficient dimension reduction or low-rank estimation.[16,17]

Lots of efforts have also been devoted to designing powerful testing methods in microbiome association studies. Koh et al[18] proposed an optimal microbiome-based association test (OMiAT) that also analyzes taxa based on their lower-level sub-taxa via group association test, and it deals with a single outcome variable and obtains *P*-values through the permutation method. To boost test power when multiple outcomes are available, the idea of leveraging correlation among outcomes has been investigated in microbiome association studies.[19-22] For example, Zhan et al[23] proposed MMiRKAT to test the effect of the microbiome on outcomes based on kernel machine regression where the association is evaluated by a variance-component score test. Zhan et al[24] used a kernel RV coefficient test to measure the global association between the microbiome and a set of phenotypes. However, these methods mainly focused on testing the overall effect of the microbiome on outcomes. To the best of our knowledge, a method that exploits the multivariate nature of the problem to carry out group inference to identify predictive taxa from several non-informative taxa is still lacking.

To assess the association between neurobehavioral outcomes of the preterm infants and their gut microbiome compositions during NICU stay, we propose a multivariate log-contrast regression with grouped sub-compositional

predictors. Motivated by Lin et al[11] and Li et al,[25] we formulate the problem as a constrained integrative multi-view regression, in which the neurobehavioral outcomes form the response matrix, the log-transformed sub-compositional data form the multi-view feature matrices, and a set of linear constraints on their corresponding coefficient matrices ensure the obedience of the simplex geometry of the compositions. The linear constraints are then conveniently absorbed through parameter transformation. The sub-compositions within each group may be strongly correlated, each individual variable may only have a weak influence and it is likely that only a few of the taxa are predictive. Thus, we assume that the sub-coefficient matrices are possible of low ranks. This assumption induces a parsimonious and highly interpretable model for dealing with high-dimensional grouped sub-compositions, that is, the outcomes are associated with the microbes through different sets of latent sub-compositional factors from different bacterial taxa, and a taxon becomes irrelevant to the outcomes when its corresponding sub-coefficient matrix is of zero rank. We develop a scaled composite nuclear norm penalization approach for model estimation and a high-dimensional hypothesis testing procedure through a de-biasing technique. We stress that the proposed approach is generally applicable for a wide range of multivariate multi-view regression problems, and to the best of our knowledge, our work is among the first to develop statistical inference methods for testing high-dimensional low-rank coefficient matrices.

The rest of this article is organized as follows. Section 2.1 introduces the multivariate log-contrast model, where the implication of the integrative low-rank structure on analyzing sub-compositional predictors is elaborated. Section 2.2 develops a scaled composite nuclear norm penalization approach for estimating both the mean structure and the variance. Computational algorithms and theoretical guarantees on the resulting estimators are then presented. Section 3 develops the inference procedure and its related theoretical results. Simulation studies to demonstrate the proposed inference procedure are presented in Section 4. Section 5 details the application of the method in the preterm infant study. A few concluding remarks and future research directions are provided in Section 6.

2 | MULTIVARIATE LOG-CONTRAST MODEL WITH SUB-COMPOSITIONAL PREDICTORS

2.1 | Model

Our work was motivated by the need of identifying gut microbiome taxa during the early postnatal period of preterm infants that may impact their later neurobehavioral outcomes. Microbiome data commonly manifest themselves as compositions. Concretely, a p dimensional compositional vector represents the relative abundances of p different taxa in a sample, and its entries are non-negative and sum up to one. Therefore, the data are multivariate in nature and reside in a simplex that does not admit the familiar Euclidean geometry. In regression analysis with compositional covariates, the log-ratio transformations are commonly adopted to lift the compositions from the simplex to the Euclidean space, which assumes the data lie in a strictly positive simplex $\mathbf{z}_i \in \mathbb{S}^{p-1} = \{[z_{i1}, \ldots, z_{ip}]^{\mathrm{T}} \in \mathbb{R}^p;\ z_{ij} > 0, \sum_{j=1}^{p} z_{ij} = 1\}$. In practice, preprocessing steps such as replacing zero counts with some small numbers (eg, the maximum rounding error) are applied. We adopt this pragmatic log-contrast regression approach in our work.

Another important feature of microbiome data is the presence of the evolutionary history charted through a taxonomic hierarchy. The major taxonomic ranks are domain, kingdom, phylum, class, order, family, genus, and species, from the highest to the lowest. Such a structure provides crucial information about the relationship between different microbes and proves useful in various analyses.[13,15] In practice, selecting the taxonomic rank or ranks at which to perform the statistical analysis depends on both the scientific problem of interest itself and the tradeoff between data quality and data resolution: the lower the rank, the higher the resolution of the taxonomic categories, but the sparser the data for each category. A good compromise is achieved by the sub-compositional regression analysis,[15] in which the effect of a taxon on the outcome at the rank of primary interest is investigated through its more information-rich sub-compositions at a lower taxonomic rank.

In the preterm infant study, the microbiome data can be presented as sub-compositional data of different bacterial taxa at the order level, each consists of a group of compositions at the genus level. To formulate, suppose we have K taxa, and within the kth taxon there are p_k many taxa that are of a lower rank. Let $z_{k,i,j}$ be the subcomposition of the jth genus under the kth order for the ith observation, $\mathbf{z}_{k,i} = [z_{k,i,1}, \ldots, z_{k,i,p_k}]^{\mathrm{T}} \in \mathbb{R}^{p_k}$ be the compositional vector of the kth order for the ith observation, and $\mathbf{Z}_k = [\mathbf{z}_{k,1}, \ldots, \mathbf{z}_{k,n}]^{\mathrm{T}} \in \mathbb{R}^{n \times p_k}$ be the data matrix of the kth order. As such, the integrated sub-compositional design matrix, that is, $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_K] \in \mathbb{R}^{n \times p}$ with $p = \sum_{k=1}^{K} p_k$, naturally admits a grouped

or multi-view structure, and it satisfies that $\mathbf{z}_{k,i} \in \mathbb{S}^{p_k-1},\ k = 1, \ldots, K; i = 1, \ldots, n$. Let $\tilde{\mathbf{Z}}_k = \log(\mathbf{Z}_k)$ and $\tilde{\mathbf{Z}} = \log(\mathbf{Z})$ be the corresponding log-transformed sub-compositional data, where $\log(\cdot)$ is applied entrywisely. Also let $\mathbf{Z}_0 \in \mathbb{R}^{n\times p_0}$ be the data matrix of control variables.

In this work, we concern multivariate outcomes, for example, the 13 sub-scale NNNS scores in the preterm infant study. Let $\mathbf{Y} \in \mathbb{R}^{n\times q}$ be the response matrix consisting of data collected from the same n subjects on q outcome variables. We now propose the multivariate log-contrast model with grouped sub-compositional predictors,

$$\mathbf{Y} = \mathbf{1}_n \boldsymbol{\mu}^{*\mathrm{T}} + \mathbf{Z}_0\mathbf{C}_0^* + \sum_{k=1}^{K} \tilde{\mathbf{Z}}_k\mathbf{C}_k^* + \mathbf{E}, \quad \text{s.t. } \mathbf{1}_{p_k}^{\mathrm{T}}\mathbf{C}_k^* = \mathbf{0},\ k = 1, \ldots, K, \tag{1}$$

where $\mu^* \in \mathbb{R}^q$ is the intercept vector, $\mathbf{C}_k^*$ is the kth coefficient sub-matrix for each $k = 0, \ldots, K$, and $\mathbf{E}$ is the random error matrix with independent entries whose mean is zero and standard deviation is σ. The linear constraints are to ensure the simplex geometry of the model.

To express the model in an unconstrained form, one can write the model in terms of log-ratio transformed compositional predictors, but the choice of the baseline taxa may lead to inconsistency in model estimation when regularization is adopted.[11] Another way is through a linear transformation of the parameters. Let's rewrite the linear constrains to be $\mathbf{L}^{\mathrm{T}}\mathbf{C} = \mathbf{0}$, with $\mathbf{C} = (\mathbf{C}_1^{\mathrm{T}}, \ldots, \mathbf{C}_K^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{L} = \mathrm{diag}\{\mathbf{1}_{p_1}, \ldots, \mathbf{1}_{p_K}\}$ and write the set of solutions to $\mathbf{L}^{\mathrm{T}}\mathbf{C} = \mathbf{0}$ as $\{(\mathbf{I}_p - \mathbf{P}_{\mathbb{C}(\mathbf{L})})\mathbf{B} : \mathbf{B} \in \mathbb{R}^{p\times q}\}$ where $\mathbf{P}_{\mathbb{C}(\mathbf{L})}$ is the orthogonal projection matrix of the column space of $\mathbf{L}$. Define $\mathbf{X} = \tilde{\mathbf{Z}}(\mathbf{I}_p - \mathbf{P}_{\mathbb{C}(\mathbf{L})}) = (\mathbf{X}_1, \ldots, \mathbf{X}_K)$, we obtain an unrestricted model

$$\mathbf{Y} = \mathbf{1}_n \mu^{*\mathrm{T}} + \mathbf{Z}_0\mathbf{C}_0^* + \sum_{k=1}^{K} \mathbf{X}_k\mathbf{B}_k^* + \mathbf{E}, \tag{2}$$

where $\mathbf{X}$ is the projected design matrix and $\mathbf{B}^* = (\mathbf{B}_1^{*\mathrm{T}}, \ldots, \mathbf{B}_K^{*\mathrm{T}})^{\mathrm{T}}$ is the corresponding coefficient matrix. From the specific form of $\mathbf{L}$, the linear constraints are imposed on each coefficient sub-matrix separately and thus the transformed design matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)$ still keeps the original grouping structure, so assessing the effect of the kth taxon can be done through testing $\mathrm{H}_0 : \mathbf{X}_k\mathbf{B}_k^* = \mathbf{0}$. Here we note that $\mathbf{B}_k^*$ is not estimable as $\mathbf{X}_k$ is not of full column rank, so we do not test $\mathrm{H}_0 : \mathbf{B}_k^* = \mathbf{0}$. In fact, the transformation on each $\mathbf{Z}_k$ is equivalent to doing a centered log-ratio transformation[8] to the sub-compositions. Henceforth, we focus on model (2).

For leveraging the association among outcomes to facilitate dimension reduction and model interpretation, we assume that each $\mathbf{B}_k^*$ is possible of low rank. That is, model (2) exhibits a taxon-specific low-rank structure. Specifically, suppose the rank of each coefficient sub-matrix is $\mathrm{rank}(\mathbf{B}_k^*) = r_k^* \le \min(p_k, q)$, for $k = 1, \ldots, K$. This structure reduces the number of unknown parameters in $\mathbf{B}^*$ to $\sum_{k=1}^{K}(p_k + q - r_k^*)r_k^*$. When r_k^*'s are small, it could be much smaller than pq, the number of unknown parameters when performing q separate univariate regressions. We can then write $\mathbf{B}_k^* = \mathbf{J}_k\mathbf{R}_k^{\mathrm{T}}$ as its full-rank decomposition, where $\mathbf{J}_k \in \mathbb{R}^{p_k\times r_k^*}$ and $\mathbf{R}_k \in \mathbb{R}^{q\times r_k^*}$ are both of full column rank. Thus $\mathbf{X}_k^* = \mathbf{X}_k\mathbf{J}_k = \tilde{\mathbf{Z}}_k(\mathbf{I}_{p_k} - \mathbf{1}_{p_k}\mathbf{1}_{p_k}^{\mathrm{T}}/p_k)\mathbf{J}_k$ provides a few latent factors of the original log-transformed data and maintains the compositional structure since it still holds that $\mathbf{1}_{p_k}^{\mathrm{T}}(\mathbf{I}_{p_k} - \mathbf{1}_{p_k}\mathbf{1}_{p_k}^{\mathrm{T}}/p_k)\mathbf{J}_k = \mathbf{0}$. These latent factors share the same structure as the principal components constructed in Aitchison's work,[26] where a log linear contrast form of principal component analysis (PCA) for compositional data was proposed to extract informative compositional proportions; see, also, Aitchison and Egozcue.[27] However, PCA is unsupervised and utilizes no information from the response, and a naive PCA of all compositional data ignores the sub-compositional structure that embodies the taxonomic hierarchy. Here, the taxon-specific multi-view low-rank structure differs in two aspects, as illustrated in Figure 1. First, $\mathbf{X}_k^*$'s are jointly predictive of $\mathbf{Y}$ since their estimation is under the supervision of $\mathbf{Y}$. Second, the dimension reduction is conducted in a taxon-specific fashion to make use of the structural information and facilitate model interpretation.

Here we remark that the correlation among outcomes we mainly considered here comes from the similar functional relationship between each response and the covariates, which could make the effective dimension of $\mathbf{Y}$ much smaller than q. While performing q separate univariate regressions completely ignores the multivariate nature of the problem, the multivariate regression setting and the imposed rank restrictions on $\mathbf{B}_k^*$'s can exploit the approximate low-rank structure of $\mathbf{Y}$ and induce a nice interpretation of the model. The independence assumption on entries of $\mathbf{E}$ does not contradict the existence of this association. Although allowing the rows in $\mathbf{E}$ to have a nonidentity covariance matrix $\boldsymbol{\Sigma}$ is a more general setting when considering multiple responses, it poses a great challenge to derive the subsequent asymptotic inference and beyond the scope of the current study.

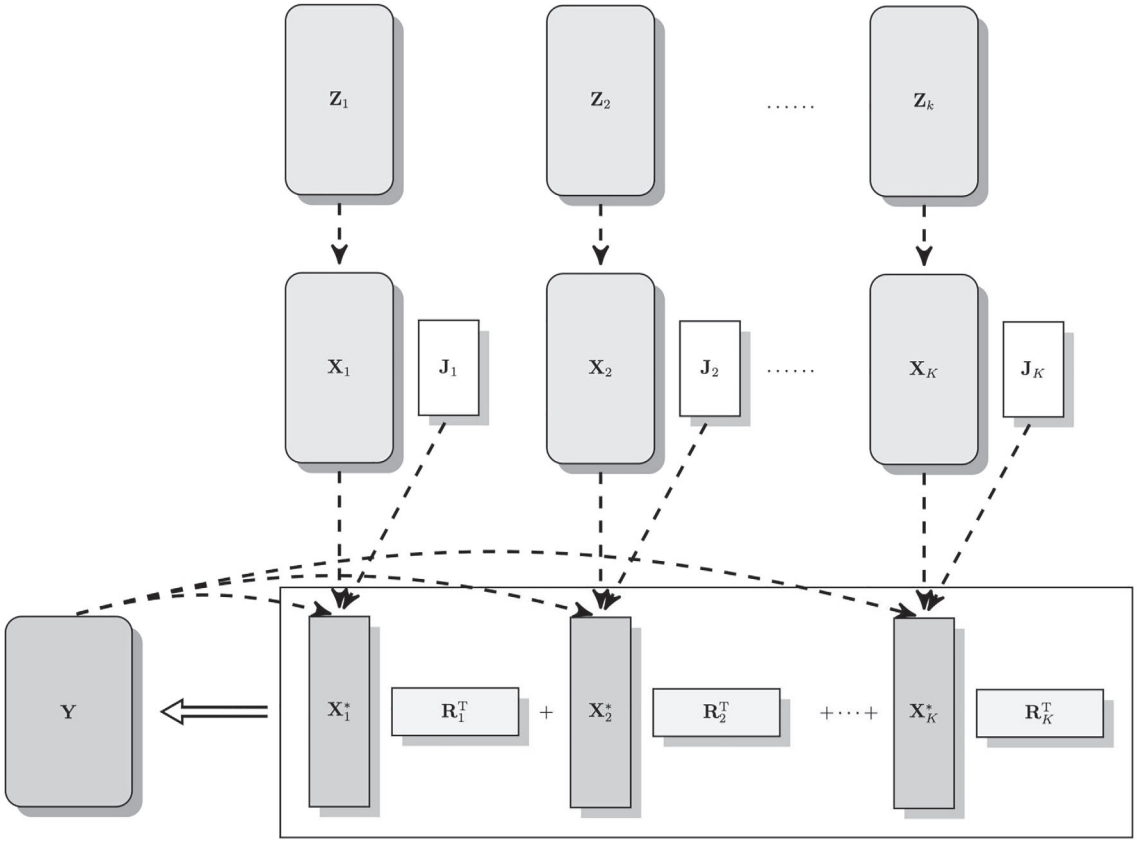

FIGURE 1 Diagram of the taxon-specific low-rank multivariate log-contrast model with grouped sub-compositional predictors. Latent taxon-specific features $\mathbf{X}_k^*$ are learned from each log-transformed sub-compositions under the compositional constraints and the supervision of $\mathbf{Y}$

2.2 | Estimation via scaled composite nuclear norm penalization

Model (2) can be recognized as an integrative reduced-rank regression (iRRR) model proposed by Li et al,[25] in which a composite nuclear norm penalization approach was developed for estimating the regression coefficients. However, due to the need for enabling statistical inference, the estimation of the error variance and the adaptive estimation of the coefficient matrix are both crucial. Therefore, following the scaled lasso[28] framework, we develop a scaled composite nuclear norm penalization approach,

$$(\hat{\mu}, \hat{\mathbf{C}}_0, \hat{\mathbf{B}}^n, \hat{\sigma}) = \arg \min_{\mu, \mathbf{C}_0, \mathbf{B}, \sigma} \left\{ \frac{1}{2nq\sigma} \|\mathbf{Y} - \mathbf{1}_n \mu^{\mathrm{T}} - \mathbf{Z}_0 \mathbf{C}_0 - \mathbf{X}\mathbf{B}\|_F^2 + \frac{\sigma}{2} + \lambda \sum_{k=1}^{K} w_k \|\mathbf{B}_k\|_* \right\}, \tag{3}$$

where for each $k = 1, \ldots, K$, $\|\mathbf{B}_k\|_*$ denotes the nuclear norm of matrix $\mathbf{B}_k$, and λ is a tuning parameter to control the amount of regularization (no penalization on $\mathbf{C}_0$). We choose the weights as $w_k = d_1(\mathbf{X}_k)\{\sqrt{p_k q} + \sqrt{2\log(K/\epsilon)}\}/(nq)$ for some $0 < \epsilon < 1$ to achieve desired statistical performance (see Theorem 1), where $d_j()$ denotes the jth largest singular value of an enclosed matrix. The application of the composite nuclear norm penalty nicely bridges low-rank models and group sparse models. Specifically, the nuclear norm of a matrix is the ℓ_1 norm of its singular values and the composite nuclear norm penalty promotes sparsity of singular values of each coefficient sub-matrix, which enforces each sub-matrix to be of low-rank and could even make the sub-matrix to be entirely zero to achieve group selection. We have developed efficient algorithms to solve (3), which are presented in Web Appendix A. The resulting estimator is termed as the scaled iRRR estimator.

We investigate the theoretical properties of the scaled iRRR estimator. For simplicity, we present the analysis with the model without the intercept and the control variables, that is, $\mathbf{Y} = \mathbf{X}\mathbf{B}^* + \mathbf{E} = \sum_{k=1}^{K} \mathbf{X}_k \mathbf{B}_k^* + \mathbf{E}$, since the results can be easily extended to the general model (2) with a fixed number of controls. A restricted strong convexity (RSC) condition[29,30]

is exploited to ensure the convexity of the loss function on a restricted parameter space. Specifically, the design matrix $\mathbf{X}$ satisfies the RSC condition over a restricted set $C(r_1, \ldots, r_K; \eta, \delta) \in \mathbb{R}^{p\times q}$ if there exists a constant $\kappa(\mathbf{X}) > 0$ such that $\frac{1}{2n}\|\mathbf{X}\Delta\|_F^2 \geq \kappa(\mathbf{X})\|\Delta\|_F^2$ for all $\Delta \in C(r_1, \ldots, r_K; \eta, \delta)$. Here r_k is the rank imposed on each coefficient sub-matrix and satisfies $1 \leq r_k \leq \min(p_k, q)$, η is a positive constant and δ is a tolerance parameter from RSC condition. For details about the restricted set, refer to Li et al.[25] Next, we give the main theoretical result of the scaled iRRR estimator.

Theorem 1. *Assume that* $vec(\mathbf{E}) \sim \mathcal{N}_{nq}(\mathbf{0}, \sigma^2\mathbf{I}_{nq})$. *Let* $(\hat{\mathbf{B}}^n, \hat{\sigma})$ *be a solution of optimization problem (3),* $\mathbf{B}^*$ *be the true coefficient matrix, and* $\sigma^* = \|\mathbf{Y} - \mathbf{X}\mathbf{B}^*\|_F/\sqrt{nq}$ *be the oracle noise level. Suppose* $\mathbf{X}$ *satisfies the RSC condition with* $\kappa(\mathbf{X}) > 0$ *over* $C(r_1, \ldots, r_K; \eta, \delta)$. *When* $w_k = d_1(\mathbf{X}_k)w_{*,k}/\sqrt{nq}$ *with* $w_{*,k} = \sqrt{p_k/n} + \sqrt{2\log(K/\epsilon)/(nq)}$ *and* $0 < \epsilon < 1$, *if we let* $\lambda = (1+\theta)(1+\eta)/\sqrt{[1 - 16(1+\eta)^2(2+\eta)\sum_{k=1}^K qr_kw_k^2/\{\eta^2\kappa(\mathbf{X})\}]_+}$ *for any* $\theta > 0$, *then with probability at least* $1-\epsilon$, *we have* $\sum_{k=1}^K \lambda w_k\|\hat{\mathbf{B}}_k^n - \mathbf{B}_k^*\|_* \leqslant \frac{\sigma^* q\sum_{k=1}^K r_k\lambda^2 w_k^2}{\sqrt{1-\tau_+}\kappa(\mathbf{X})}$, $\|\hat{\mathbf{B}}^n - \mathbf{B}^*\|_F^2 \leqslant \frac{\sigma^{*2}q^2\sum_{k=1}^K r_k\lambda^2 w_k^2}{(1-\tau_+)\kappa^2(\mathbf{X})}$, *and*

$$\frac{1}{\sqrt{nq}}\sum_{k=1}^K \frac{\lambda w_{*,k}}{\sigma}\|\mathbf{X}_k\hat{\mathbf{B}}_k^n - \mathbf{X}_k\mathbf{B}_k^*\|_F = O_p\left(\sum_{k=1}^K qr_k\lambda^2 w_k^2\right) \tag{4}$$

when each $\mathbf{B}_k^*$ *is exactly of rank* r_k *and* $\tau_+ = 8(2+\eta)^2\sum_{k=1}^K qr_k\lambda^2 w_k^2/\{\eta^2\kappa(\mathbf{X})\}$. *In addition, if* $\sqrt{nq}\sum_{k=1}^K qr_k\lambda^2 w_k^2/\kappa(\mathbf{X}) \to 0$, *we have*

$$\sqrt{nq}\left(\frac{\hat{\sigma}}{\sigma} - 1\right) \to \mathcal{N}\left(0, \frac{1}{2}\right). \tag{5}$$

The proof is relegated to Web Appendix B. Theorem 1 provides the error rates of the scaled iRRR estimator $\hat{\mathbf{B}}^n$, and establishes the consistency and the asymptotic distribution of $\hat{\sigma}$. The incorporation of noise level estimation leads to the major difference between Theorem 1 here and Theorem 2 in Li et al.[25] In particular, the specific forms of w_k's are derived from different probability inequalities in proofs of two theorems. Moreover, Theorem 1 is able to recover the error rates of both the scaled group lasso estimator[31] and scaled lasso estimator.[28] With the assumption that $\lambda d_1(\mathbf{X}_k)/\sqrt{n} \asymp 1$ and plug in the exact form of w_k's we have

$$q\sum_{k=1}^K r_k\lambda^2 w_k^2 \asymp \frac{\sum_{k=1}^K r_k\{p_kq + 2\log(K/\epsilon)\}}{nq}. \tag{6}$$

By letting $q = 1$, (6) reduces to the rate of the scaled group lasso estimator in mixed ℓ_2 loss under a strong group sparsity condition,[32] which is of the order $\{s + g\log(K/\epsilon)\}/n$ with g the number of predictive groups and s the number of entries contained in these groups. If further we let $K = p$ and $p_k = 1$ for all k, then the rate becomes $s\sqrt{\log(p/\epsilon)/n}$ with s the cardinality of the active set, which is the rate for the scaled lasso in ℓ_1 loss.

3 | HYPOTHESIS TESTING FOR SUB-COMPOSITIONAL INFERENCE

We concern the problem of testing $\mathrm{H}_0 : \mathbf{X}_k\mathbf{B}_k^* = \mathbf{0}$ vs $\mathrm{H}_1 : \mathbf{X}_k\mathbf{B}_k^* \neq \mathbf{0}$ under model (2), from which the test result indicates the significance level of the predictive power of the kth group of covariates on the responses when controlling the effects from other covariates. In the preterm infant gut microbiome study, the application of the proposed test can facilitate the identification of potential biomarkers, ie, bacterial taxa that relate to later neurological disorders with any given level of confidence. One feature of the problem is that we need to test whether a group of covariates is predictive at all to multiple responses, while most available methods focus on inference with a single response or on inference for all covariates as a whole. Statistical inference for regularized estimators is undergoing exciting development in recent years. Our approach is built upon the scaled iRRR and the work by Mitra and Zhang[31] and Zhang and Zhang[33] on the low-dimensional projection estimator (LDPE). See Web Appendix C for a brief overview of high-dimensional inference procedures and the LDPE approach in particular. The details of our proposed inference procedure are provided in Web Appendix D. In what follows, we summarize the main steps of implementing the proposed method.

Let $\mathbf{S}_k \in \mathbb{R}^{n\times p_k}$ be the score matrix of $\mathbf{X}_k$, a critical tool used in LDPE to correct the bias caused by regularization and only depends on $\mathbf{X}$. Write $\mathbf{Q}_k$ and $\mathbf{P}_{0,k}$ be the orthogonal projection matrices onto the column spaces of $\mathbf{X}_k$ ($\mathbb{C}(\mathbf{X}_k)$) and $\mathbf{S}_k$ ($\mathbb{C}(\mathbf{S}_k)$), respectively, and let $\mathbf{P}_k$ be the projection matrix of $\mathbb{C}(\mathbf{P}_{0,k}\mathbf{Q}_k)$. If $\text{rank}(\mathbf{S}_k^{\mathrm{T}}\mathbf{X}_k) = \text{rank}(\mathbf{X}_k)$, which guarantees the effectiveness of the de-biasing procedure, then with $r_k' = \text{rank}(\mathbf{P}_k) = \text{rank}(\mathbf{X}_k)$ and the assumption on the error matrix that $vec(\mathbf{E}) \sim \mathcal{N}_{nq}(\mathbf{0}, \sigma^2\mathbf{I}_{nq})$, we have a test statistic

$$T_k = \frac{1}{\hat{\sigma}^2}\left\|\mathbf{P}_k\left(\mathbf{Y} - \sum_{j\neq k}\mathbf{X}_j\hat{\mathbf{B}}_j^n\right)\right\|_F^2 \overset{H_0}{\sim} \chi^2_{r_k' q} \tag{7}$$

asymptotically, where $\hat{\mathbf{B}}_j^n$ and $\hat{\sigma}$ are the scaled iRRR estimator and $\mathbf{P}_k$ can be estimated from

$$\hat{\mathbf{\Gamma}}_{-k} = \arg\min_{\mathbf{\Gamma}_{j,j\neq k}}\left\{\frac{1}{2n}\|\mathbf{X}_k - \sum_{j\neq k}\mathbf{X}_j\mathbf{\Gamma}_j\|_F^2 + \sum_{j\neq k}\frac{\xi w_j''}{\sqrt{n}}\|\mathbf{X}_j\mathbf{\Gamma}_j\|_*\right\} \tag{8}$$

with w_j'' some prespecified weights and ξ a tuning parameter. We estimate the score matrix through $\mathbf{S}_k = \mathbf{X}_k - \mathbf{X}_{-k}\hat{\mathbf{\Gamma}}_{-k}$ and $\mathbf{P}_k = \mathbf{S}_k(\mathbf{S}_k^{\mathrm{T}}\mathbf{S}_k)^{-1}\mathbf{S}_k^{\mathrm{T}}$. The algorithm to solve (8) is provided in Web Appendix A. For the selection of ξ, in practice we only have to find a ξ to make sure $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k)) < 1$, which implies the key condition $\text{rank}(\mathbf{S}_k^{\mathrm{T}}\mathbf{X}_k) = \text{rank}(\mathbf{X}_k)$ in de-biasing and testing.[31] The validity of the test is guaranteed by the following result.

Theorem 2. *Let $(\hat{\mathbf{B}}^n, \hat{\sigma})$ be from solving (3), $\mathbf{P}_k$ from (8) with $w_j'' = w_{*j}$ and $w_{*j} = \sqrt{p_j/n} + \sqrt{2\log(K/\epsilon)/(nq)}$, $0 < \epsilon < 1$. The proposed asymptotic hypothesis testing procedure is valid if*

$$\frac{r_k'}{n} \to 0,\ \sum_{j=1}^{K}\frac{r_j\{p_j q + 2\log(K/\epsilon)\}}{\sqrt{nq}}\left\{\xi d_{min}(\mathbf{S}_k/\sqrt{n})^{-1}\right\} \to 0, \tag{9}$$

where $d_{min}(\cdot)$ is the smallest singular value of an enclosed matrix.

The proof is in Web Appendix E. Theorem 2 implies that, once the sample size is large enough compared to the test size r_k' and the model complexity, the bias can be ignored and the asymptotic test can produce reliable inference results. As such, with a prefixed significance level α, we reject the null hypothesis if $T_k > \chi^2_{\alpha,r_k' q}$, the αth upper quantile of the $\chi^2_{r_k' q}$ distribution. Again, due to the application of the LDPE and the generality of the scaled iRRR framework, the derived test can be specialized to solve lasso and group lasso estimator inference problems.

4 | SIMULATION

We conduct simulation studies to investigate the performance of the proposed method in making group inference. To show the power gained by multivariate testing, we also apply scaled group lasso testing procedure[31] to each response and exploit a union test[34] principle to combine the results. Specifically, $\mathbf{X}_k$ is significantly associated with $\mathbf{Y}$ if it is significantly associated with at least one of the q responses in $\mathbf{Y}$. Two procedures are employed to control the familywise type I error, one is the Bonferroni correction and another is the harmonic mean P-value (HMP) test.[35] Three simulation scenarios are considered: (1) the predictors in $\mathbf{X}$ are generated from multivariate normal distributions; (2) we mimic the structure of the preterm infant data to generate compositional predictors which are then processed to produce $\mathbf{X}$, and (3) we directly use the observed compositional data from the preterm infant study through resampling with replacement. The latter two are to investigate the behaviors of the proposed method with realistic microbiome data, see Web Appendix F for details and results.

4.1 | Simulation with normally distributed predictors

We work on two model settings with different dimensionality and complexity:

1. $n = 500, q = 5, p = 50, K = 5, p_i = 10,\ i = 1,\dots,5$, and $r_1^* = 2,\ r_i^* = 0,\ i = 2,\dots,5$.
2. $n = 200, q = 10, p = 200, K = 20, p_i = 10,\ i = 1,\dots,10$, and $r_1^* = 1, r_i^* = 0,\ i = 2,\dots,20$.

The design matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K) \in \mathbb{R}^{n\times p}$, true coefficient matrix $\mathbf{B}^* = (\mathbf{B}_1^{*\mathrm{T}}, \ldots, \mathbf{B}_K^{*\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p\times q}$ and the corresponding response matrix $\mathbf{Y} \in \mathbb{R}^{n\times q}$ are generated as below:

1. Generate $\mathbf{B}_k^* \in \mathbb{R}^{p_k\times q}$ of rank r_k^*, $k = 1, \ldots, K$, through full-rank decomposition $\mathbf{B}_k^* = \mathbf{J}_k\mathbf{R}_k^{\mathrm{T}}$ where $\mathbf{J}_k \in \mathbb{R}^{p_k\times r_k^*}$ and $\mathbf{R}_k \in \mathbb{R}^{q\times r_k^*}$, and each entry of both $\mathbf{J}_k$ and $\mathbf{R}_k$ is generated from $\mathcal{N}(0, 1)$. Then we scale the coefficient matrix to make its largest entry to be 1.
2. Each row of $\mathbf{X}$ is generated independently from a multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Two covariance structures are considered, (1) within-group autoregressive, that is, $\mathbf{\Sigma}$ is block diagonal with diagonal blocks $\mathbf{\Sigma}_k = (\rho^{|i-j|}) \in \mathbb{R}^{p_k\times p_k}$, and (2) among-group autoregressive, that is, $\mathbf{\Sigma} = (\rho^{|i-j|})$. The correlation strength ρ_x is in $\{0, 0.5\}$.
3. The entries of $\mathbf{E}$ are drawn from $\mathcal{N}(0, \sigma^2)$ and the response matrix $\mathbf{Y}$ is obtained from $\mathbf{Y} = \mathbf{X}\mathbf{B}^* + \mathbf{E}$, where σ^2 is set to control the signal to noise ratio (SNR), defined as the ratio between the standard deviation of the linear predictor $\sum_{k=1}^K \mathbf{X}_k\mathbf{B}_k^*$ and the standard deviation of the random error. We consider SNR $\in \{0.1, 0.2, 0.4\}$.

In each replication, we generate $(\mathbf{X}, \mathbf{Y})$ and conduct group-wise tests with significance level 0.05. We use a Bayesian information criterion (BIC)[36] to select the tuning parameter in the scaled iRRR and the scaled group lasso regression. As for the estimation of the score matrix, we use $\xi = 1$ in (8) which is verified to be adequate for satisfying $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k)) < 1$ in all the settings. Under each setting, the simulation is repeated 300 times. We compute the mean and standard deviation of $\hat{\sigma}/\sigma - 1$ and $|\hat{\sigma}/\sigma - 1|$, respectively, to measure the performance of the noise level estimation. For assessing the inference procedure, we compute both the false positive rate (FP), the proportion of time the test for an irrelevant group is rejected, and the true positive rate (TP), the proportion of time the test for a relevant group is rejected.

We first examine the asymptotic distributions of both $\sqrt{2nq}(\hat{\sigma}/\sigma - 1)$ and the pivotal statistic $\|\mathbf{P}_k\mathbf{E} - \mathrm{Rem}_k\|_F^2/\hat{\sigma}^2$ (refer to Web Appendix D) using Setting 1, and we fix a randomly generated $\mathbf{X}$ with the within-group correlation setup and generate $\mathbf{E}$ in each replication. Figure 2 displays the normal Q-Q plots of $\sqrt{2nq}(\hat{\sigma}/\sigma - 1)$ under different SNR and ρ_x settings. In each plot, the majority of the points approximately lie on a straight line that is coincident with or parallel to the diagonal line. The parallel discrepancy above the diagonal line is caused by the fact that BIC is in favor of a sparser model which leads to the overestimation of the noise level. Figure 3 displays the χ^2 Q-Q plots to verify the asymptotic distribution of the pivotal statistic. Indeed it approximately follows a χ^2 distribution with degree of freedom $r_k'q$. See Web Figures 1 and 2 for the Q-Q plots with $\rho_x = 0$.

Table 1 reports the detailed results on hypothesis testing under different settings with $\mathbf{X}$ being generated from the within-group correlation setup. Correlation patterns among covariates appear to have little effect on these results. See Web Table 4 for the results with the among-group correlation setup. In general, the magnitude of SNR and the model dimensionality have great influence on the TP, which measures the power of the test, while the type I error rate, that is, FP, is not sensitive to the change of these two factors and only oscillates slightly around 0.05. Specifically, in Setting 1 where $n > p$, the power of the test is moderate when SNR is very low. When the SNR becomes stronger, the power of the test in these two settings dramatically increase to be close to 1. Setting 2 is a high-dimensional situation; when SNR is low,

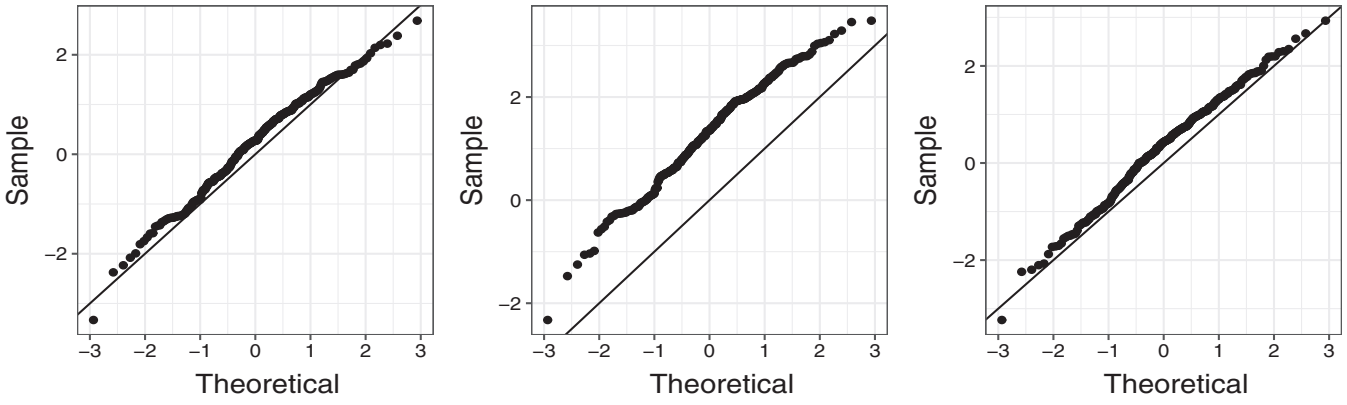

FIGURE 2 Simulation results for Setting 1 with $\rho_x = 0.5$ from 300 simulation runs: From left to right are the Q-Q plots of $\sqrt{2nq}(\hat{\sigma}/\sigma - 1)$ vs $\mathcal{N}(0, 1)$ with SNR = 0.1, 0.2, and 0.4, respectively

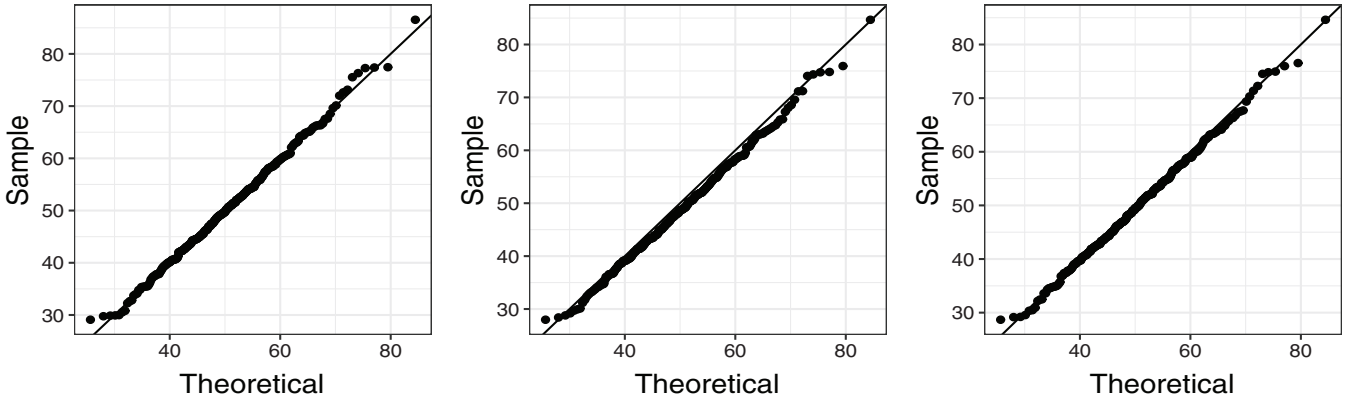

FIGURE 3 Simulation results for Setting 1 with $\rho_x = 0.5$ from 300 simulation runs: from left to right are the Q-Q plots of $\|\mathbf{P}_3\mathbf{E} - \text{Rem}_3\|_F^2/\hat{\sigma}^2$ vs $\chi^2_{r_3^* q}$ with SNR = 0.1, 0.2, and 0.4, respectively

TABLE 1 Simulation results with $\mathbf{X}$ being generated from multivariate normal distribution with the within-group correlation setup

			Multivariate			Univariate (Bonf)			Univariate (HMP)		
Design (SNR, ρ_x)	$\hat{\sigma}/\sigma - 1$	$\lvert\hat{\sigma}/\sigma - 1\rvert$	**G1**	**G2**	**G3**	**G1**	**G2**	**G3**	**G1**	**G2**	**G3**
Setting 1											
(0.1, 0.0)	0.34 (1.44)	1.17 (0.90)	0.65	0.05	0.05	0.51	0.03	0.03	0.54	0.03	0.03
(0.1, 0.5)	0.34 (1.44)	1.17 (0.90)	0.65	0.05	0.05	0.52	0.03	0.03	0.56	0.03	0.03
(0.2, 0.0)	1.82 (1.45)	1.97 (1.23)	1.00	0.04	0.04	1.00	0.03	0.03	1.00	0.03	0.03
(0.2, 0.5)	1.81 (1.45)	1.97 (1.23)	1.00	0.03	0.04	1.00	0.03	0.03	1.00	0.03	0.03
(0.4, 0.0)	0.48 (1.51)	1.23 (1.00)	1.00	0.04	0.05	1.00	0.02	0.03	1.00	0.02	0.03
(0.4, 0.5)	1.11 (1.60)	1.52 (1.21)	1.00	0.04	0.04	1.00	0.03	0.03	1.00	0.03	0.03
Setting 2											
(0.1, 0.0)	0.18 (1.61)	1.30 (0.96)	0.17	0.04	0.05	0.10	0.04	0.05	0.11	0.05	0.05
(0.1, 0.5)	0.18 (1.61)	1.30 (0.95)	0.20	0.04	0.06	0.12	0.04	0.05	0.13	0.03	0.05
(0.2, 0.0)	1.65 (1.63)	1.89 (1.34)	0.69	0.03	0.02	0.55	0.03	0.04	0.59	0.04	0.04
(0.2, 0.5)	1.65 (1.62)	1.88 (1.34)	0.83	0.03	0.05	0.65	0.03	0.05	0.70	0.03	0.04
(0.4, 0.0)	0.43 (1.65)	1.35 (1.03)	1.00	0.04	0.05	1.00	0.02	0.03	1.00	0.03	0.03
(0.4, 0.5)	1.55 (1.72)	1.88 (1.34)	1.00	0.03	0.05	1.00	0.02	0.04	1.00	0.02	0.04

Note: The performance of noise level estimation is displayed in terms of the mean (×100) and standard error (×100, in parenthesis) of $\hat{\sigma}/\sigma - 1$ and $|\hat{\sigma}/\sigma - 1|$, respectively. In both settings, we have $r_1^* \neq 0$ and $r_2^* = r_3^* = 0$. Each group is denoted as "G" followed by its group number. For the two univariate methods, we use "Bonf" to represent Bonferroni adjustment and use "HMP" to represent the harmonic mean *P*-value test.

the power of the test is generally lower than in Setting 1. With a higher signal strength, that is, SNR = 0.4, the power of the test achieves 1. Moreover, in these two settings, multivariate analysis consistently performs better than the univariate analysis in terms of the power of the test.

4.2 | Simulation with generated compositional data

We conduct simulations based on generated compositional data, and details of data generation are provided in Web Appendix F. The results are shown in Table 2. The test has relatively low power for both the multivariate and univariate methods when the signal is weak, ie, when the largest entry in $\mathbf{B}^*$ is set to be 0.2 and

TABLE 2 Simulation results based on the generated compositional data across 300 replications

			Multivariate				Univariate (Bonf)				Univariate (HMP)			
Design (SNR, ρ_X)	$\hat{\sigma}/\sigma - 1$	$\lvert\hat{\sigma}/\sigma - 1\rvert$	**G1**	**G2**	**G6**	**G7**	**G1**	**G2**	**G6**	**G7**	**G1**	**G2**	**G6**	**G7**
(0.2, 0.2)	0.04 (1.24)	0.99 (0.74)	0.13	0.06	0.14	0.02	0.11	0.04	0.09	0.04	0.12	0.04	0.08	0.04
(0.2, 0.5)	−0.05 (1.24)	0.99 (0.74)	0.10	0.05	0.10	0.04	0.09	0.04	0.07	0.04	0.09	0.05	0.07	0.04
(0.4, 0.2)	0.76 (1.25)	1.18 (0.87)	0.66	0.04	0.72	0.02	0.48	0.04	0.61	0.04	0.56	0.03	0.65	0.04
(0.4, 0.5)	0.43 (1.25)	1.06 (0.79)	0.41	0.05	0.46	0.03	0.27	0.04	0.39	0.03	0.29	0.04	0.41	0.04
(0.8, 0.2)	3.08 (1.45)	3.09 (1.42)	1.00	0.02	1.00	0.01	1.00	0.03	1.00	0.03	1.00	0.02	1.00	0.03
(0.8, 0.5)	2.34 (1.25)	2.38 (1.17)	1.00	0.02	1.00	0.02	0.99	0.03	0.98	0.03	1.00	0.03	0.99	0.03

Note: The performance of noise level estimation is displayed in terms of the mean (×100) and standard error (×100, in parenthesis) of $\hat{\sigma}/\sigma - 1$ and $|\hat{\sigma}/\sigma - 1|$, respectively. In the simulation setting, we have $r_1^* = r_6^* = 1$ and $r_2^* = r_7^* = 0$. Each group is denoted as "G" followed by its group number. For the two univariate methods, we use "Bonf" to represent Bonferroni adjustment and use "HMP" to represent the harmonic mean *P*-value test.

SNR = 0.2. The power of the test gradually increases when the SNR becomes larger, and the false positive rate is well controlled around 0.05 except for the case with severely overestimated σ when SNR = 0.8. Moreover, the test power is hampered for groups containing highly abundant taxa. This phenomenon could be related to $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ that is directly affected by the abundance level. Specifically, in this simulation, for the groups that have unbalanced taxa distributions (group 1-5), their $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ is close to 1, and for the groups whose components take comparable proportions (group 6-10), the corresponding $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ is much smaller (see Web Table 3). As we discussed before, $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ measures the uniqueness of the information carried by the kth group, thus a smaller value indicates a higher inference accuracy. In addition, there is an improvement in the power brought by the multivariate analysis in this setting.

5 | ASSESSING THE ASSOCIATION BETWEEN PRETERM INFANTS' GUT MICROBIOME AND NEUROBEHAVIORAL OUTCOMES

5.1 | Data description

The study was conducted at a Level IV NICU in the northeast region of the United States. Fecal samples were collected in a daily manner when available during the first month of the postnatal age of infants, from which bacteria DNA were isolated and extracted.[13,37,38] The V4 region of the 16S rRNA genes was sequenced and analyzed using the Illumina platform and QIIME,[38] and microbiome data were obtained. There were $n = 38$ infants under study.

In practice, the selection of the taxonomic ranks at which to perform the statistical analysis depends on both the scientific problem itself and trade-off between data quality and resolution: the lower the taxonomic rank, the higher the resolution of the taxonomic units, but the sparser or the less reliable the data in each unit. To achieve a compromise, here we perform a sub-compositional analysis: we assess the effects of the order-level gut microbes through compositions at the genus level, a lower taxonomic rank. The microbes were categorized into 62 genera ($\sum_{k=1}^{K} p_k = 62$), which can be grouped into $K = 11$ predictor sets based on their orders. The original orders only containing a single genus were put together as the "Other" group. The preterm infant data were longitudinal, with on average 11.4 daily observations per infant through the 30 day postnatal period. In this study, we concern average microbiome compositions in three stages, that is, stage 1 (postnatal age of 0-10 days, $n = 33$), stage 2 (postnatal age of 11-20 days, $n = 38$), and stage 3 (postnatal age of 21-30 days, $n = 29$), in order to enhance data stability and capture the potential time-varying effects of the gut microbiome on the later neurodevelopmental responses. We also performed an analysis on the average compositions of the entire time period. Figure 4 displays the average abundance of the orders for each infant at different stages. Before calculating the compositions, we replaced the zero counts by 0.5, the maximum rounding error, to avoid singularity.[8] Several control variables characterizing demographical and clinical information of infants were included ($p_0 = 6$), including gender (binary, female = 1), delivery type (binary, vaginal = 1), premature rupture of membranes (PROM, yes = 1), score for Neonatal Acute Physiology-Perinatal

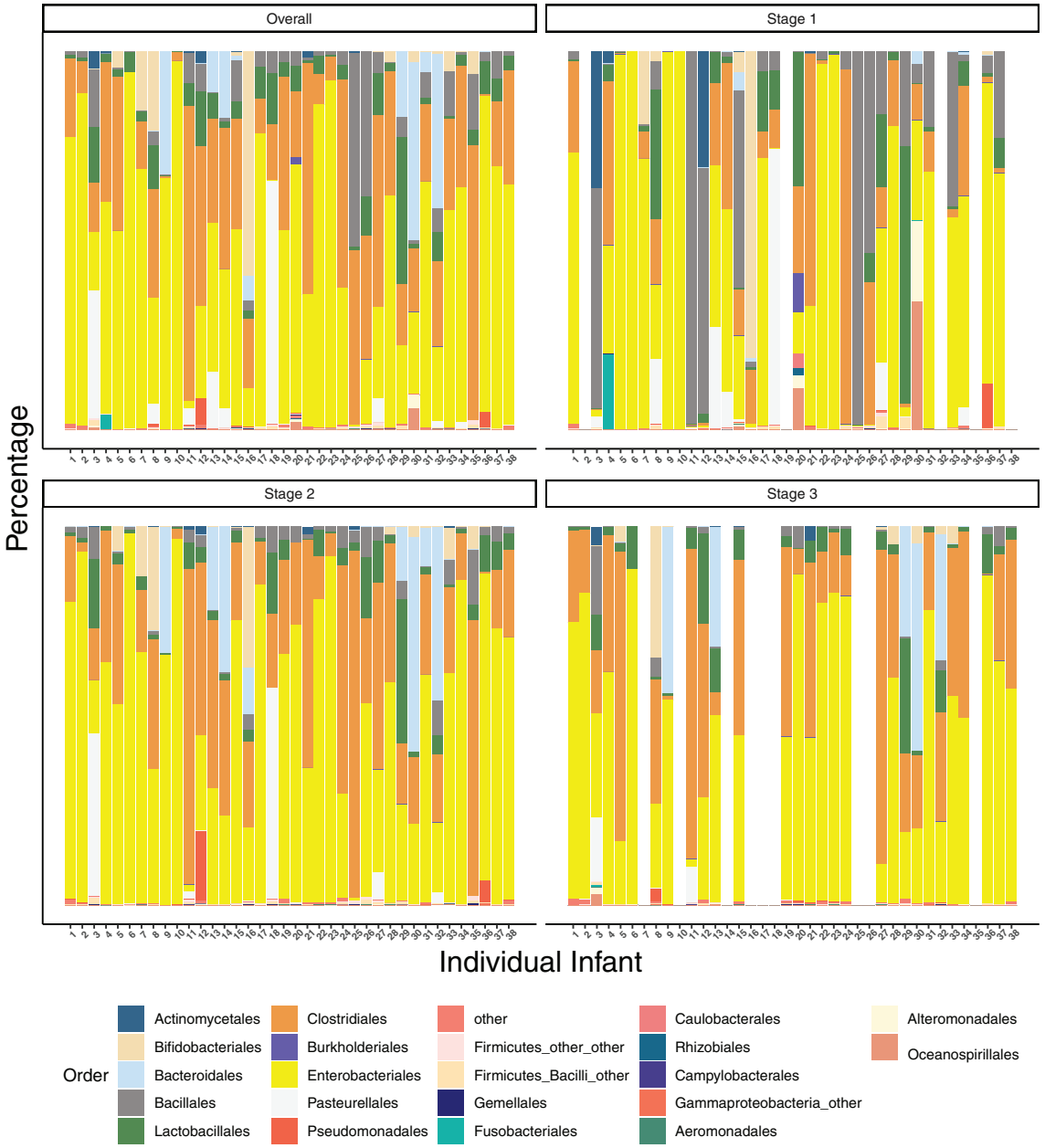

FIGURE 4 The average abundance profiles of the 22 orders for each infant at three stages: stage 1 (postnatal age of 0-10 days; $n = 33$), stage 2 (postnatal age of 11-20 days; $n = 38$), and stage 3 (postnatal age of 21-30 days; $n = 29$). The profiles of the infants with no observation are shown as in white color

Extension-II (SNAPPE-II), birth weight (in gram), and the mean percentage of feeding by mother's breast milk (MBM, in percentage).

The infants' neurobehavioral outcomes were measured when the infant reached 36 to 38 weeks of gestational age using NNNS. NNNS is a comprehensive assessment of both neurologic integrity and behavioral function for infants. It consists of 13 sub-scale scores including habituation, attention, handling, quality of movement, regulation, nonoptimal reflexes, asymmetric reflexes, stress/abstinence, arousal, hypertonicity, hypotonicity, excitability, and lethargy. These scores were obtained by summarizing several examination results within each sub-category in the form of the sum or mean, and all of them can be regarded as continuous measurements with a higher score on each scale implying a higher level of the construct.[39] We discarded sub-scales hypertonicity, hypotonicity, and asymmetric reflexes since their scores are mostly zero (over 65%, which severely destroyed the normal assumption on outcomes) and focused on the other 10 standardized sub-scale scores ($q = 10$).

5.2 | Results

The results are shown in Table 3. To control the false discovery rate (FDR) when multiple tests are conducted, we mark the identified orders based on the corrected *P*-values with Benjamini-Hochberg adjustment.[40] First, we observe that the predictive effects of the microbiome on the neurobehavioral development measurements appear to be dynamic, that is, in different time periods, the identified taxa are not the same. This reflects the fact that the gut microbiome compositions in early postnatal period are highly variable, due to their sensitivity to illnesses, changes in diet and environment.[41,42] Specifically, by controlling the FDR under 10%, the identified orders from all analyses are Actinomycetales, Clostridiales, Burkholderiales, and the aggregated group "Others." If we set 0.05 as the significance level without doing multiple testing adjustment, there is one more significant order, Lactobacillales. This dataset is also analyzed by Sun et al[13] through a sparse log-contrast functional regression method to identify predictive gut bacterial orders to the stress/abstinence sub-scale, and their selected orders based on penalized estimation include Lactobacillales, Clostridiales, Enterobacteriales, and the group "Others," which are very consistent with our results. Here we stress that our work is quite different from Sun et al[13]: our analysis assesses the multivariate association between the orders and the multiple neurodevelopment measurements using a valid statistical inference procedure through sub-compositional analysis, while Sun et al[13] emphasized estimating the dynamic effects of the orders to the stress score alone by fitting a regularized functional regression with the order-level data.

We have also conducted univariate analysis for each sub-scale, that is, we fit the proposed model with each sub-scale score as the univariate response and make inference. For each time period, this procedure produces a large number of tests. By controlling the FDR under 10%, only in stage 2 we can identify the order Burkholderiales to be predictive to excitability, stress/abstinence and handling (refer to Web Figure 3). This is not surprising as this dataset has a limited sample size and a weak signal strength. To gain more power in practice, the proposed multivariate test can be used first to verify the existence of any association between neurodevelopment and gut taxa, and univariate tests can then be conducted as post-hoc analysis to further inspect the pairwise associations. As such, we only conduct the univariate tests related to the orders identified from our multivariate analysis. With the FDR controlled at 10%, Burkholderiales is found to be significant in stage 2, and Actinomycetales is identified to be significant in stage 3 (see Web Table 5).

Most of the identified orders are known to be of various biological functions to human beings. Both Lactobacillales and Clostridiales belong to the phylum Firmicutes, which are found to be abundant for infants fed with mother's breast milk.[43] Lactobacillales are usually found in decomposing plants and milk products, and they commonly exist in food and are found to contribute to the healthy microbiota of animal and human mucosal surfaces. Clostridiales are commonly found in the gut microbiome and some Clostridiales-associated bacterial genera in the gut are

TABLE 3 Raw *P*-values from the sub-compositional analysis applied to the preterm infant data

	Overall	Stage 1	Stage 2	Stage 3
Actinomycetales	0.48	0.58	0.59	**0.00***
Bifidobacteriales	0.90	0.40	0.83	0.87
Bacteroidales	0.57	0.44	0.46	0.49
Bacillales	0.68	0.27	0.21	0.38
Lactobacillales	**0.03**	0.34	0.09	0.68
Clostridiales	**0.01***	**0.02**	**0.02**	**0.04**
Burkholderiales	0.46	0.45	**0.00***	0.20
Enterobacteriales	0.08	0.20	0.72	0.36
Pasteurellales	0.09	0.27	0.89	0.60
Pseudomonadales	0.55	0.78	0.56	0.57
Others	**0.00***	0.52	0.12	0.21

Note: Without multiple adjustment, the identified orders under significance level 0.05 are marked in bold. With BH adjustment to control the FDR under 10%, the identified orders are marked with an asterisk.

correlated with brain connectivity and health function.[44] The order Burkholderiales includes pathogens that are related to inflammatory bowel disease, especially for children's ulcerative colitis.[45] The genus Actinomyces from the order Actinomycetales is observed in this study. As a commensal bacteria that colonizes the oral cavity, gastrointestinal or genitourinal tract, Actinomyces normally cause no disease. However, invasive disease may occur when mucosal wall undergoes destruction.[46] Moreover, certain species in Actinomyces is known to possess the metabolic potential to breakdown and recycle organic compounds, for example, glucose and starch.[47] As for the effects of the control variables on the neurodevelopment of preterm infants, the estimated coefficients from the overall model and its related discussion are in Web Appendix G.

To summarize, the identified bacterial taxa are mostly consistent with existing studies and biological understandings. Therefore, our approach provides rigorous supporting evidence that stressful early life experiences imprint gut microbiome through the regulation of the gut-brain axis and impact later neurodevelopment.

6 | DISCUSSION

We propose a multivariate multi-view log-contrast model to facilitate sub-composition selection, which together with an asymptotic hypothesis testing procedure successfully identifies several neurodevelopment related bacteria taxa in a preterm infant study. There are many directions for future research. BIC is used to select the tuning parameter for the scaled iRRR in our simulation and application, however, as it poses a larger penalty on the model size the BIC may lead to overestimation of the noise level. We have also experimented with cross validation and found that in most situations the over-selection of cross validation will cause underestimation of the noise level, which often leads to inflation of the false positive rate in the subsequent inference. It is necessary to explore other approaches to tuning to obtain a more accurate estimate of the noise level. Another pressing issue is to comprehensively investigate the robustness of the method to the violation of the homoscedasticity, independence, and normality of the error terms, since the theoretical guarantees of the scaled iRRR and the inference procedure are built on these strong assumptions. The extension to general covariance structure is appealing but challenging. Combining our approach with a covariate-adjusted (inverse) covariance estimation method would yield even greater performance gains. However, it would require more efforts to derive the inference procedure, as the direct generalization of the existing inference methods (eg, LDPE) will not work anymore.

ACKNOWLEDGEMENTS

The authors thank the editor, associate editor, and referees for helpful comments and suggestions. This work was supported by National Science Foundation under grants DMS-1613295 and IIS-1718798. This work was supported by National Institutes of Health National Institute of Nursing Research under grants R01NR016928 and K23NR014674.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the authors.

ORCID

Xiaokang Liu https://orcid.org/0000-0001-6920-5598

REFERENCES

1. Fanaroff AA, Hack M, Walsh MC. The NICHD neonatal research network: changes in practice and outcomes during the first 15 years. *Semin Perinatol*. 2003;27(4):281-287.
2. Stoll BJ, Hansen NI, Bell EF, et al. Neonatal outcomes of extremely preterm infants from the NICHD neonatal research network. *Pediatrics*. 2010;126(3):443-456.
3. Mwaniki MK, Atieno M, Lawn JE, Newton CR. Long-term neurodevelopmental outcomes after intrauterine and neonatal insults: a systematic review. *Lancet*. 2012;379(9814):445-452.
4. Carabotti M, Scirocco A, Maselli MA, Severi C. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Ann Gastroenterol*. 2015;28(2):203-209.

5. Dinan TG, Cryan JF. Regulation of the stress response by the gut microbiota: implications for psychoneuroendocrinology. *Psychoneuroendocrinology*. 2012;37(9):1369-1378.
6. Cong X, Xu W, Romisher R, et al. Microbiome: gut microbiome and infant health: brain-gut-microbiota axis and host genetic factors. *Yale J Biol Med*. 2016;89(3):299-308.
7. Aitchison J. The statistical analysis of compositional data. *J R Stat Soc Ser B Stat Methodol*. 1982;44(2):139-160.
8. Aitchison J. *The Statistical Analysis of Compositional Data*. Caldwell, NJ: Blackburn Press; 2003.
9. Aitchison J, Bacon-shone J. Log contrast models for experiments with mixtures. *Biometrika*. 1984;71(2):323-330.
10. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu Rev Stat Appl*. 2015;2(1):73-94.
11. Lin W, Shi P, Feng R, Li H. Variable selection in regression with compositional covariates. *Biometrika*. 2014;101(4):785-797.
12. Wang T, Zhao H. Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann Appl Stat*. 2017;11(2):771-791.
13. Sun Z, Xu W, Cong X, Li G, Chen K. Log-contrast regression with functional compositional predictors: linking preterm infants' gut microbiome trajectories to neurobehavioral outcome. *Ann Appl Stat*. 2020;14(3):1535-1556.
14. Combettes PL, Müller CL. Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *Stat Biosci*. 2021;13:217-242.
15. Shi P, Zhang A, Li H. Regression analysis for microbiome compositional data. *Ann Appl Stat*. 2016;10(2):1019-1040.
16. Tomassi D, Forzani L, Duarte S, Pfeiffer RM. Sufficient dimension reduction for compositional data. *Biostatistics*. 2021;22(4):687-705.
17. Wang H, Wang Z, Wang S. Sliced inverse regression method for multivariate compositional data modeling. *Stat Pap*. 2021;62:361-393.
18. Koh H, Blaser MJ, Li H. A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome*. 2017;5(1):45.
19. Maity A, Sullivan PF, Tzeng JI. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet Epidemiol*. 2012;36(7):686-695.
20. Maity A, Zhao J, Sullivan PF, Tzeng JY. Inference on phenotype-specific effects of genes using multivariate kernel machine regression. *Genet Epidemiol*. 2018;42(1):64-79.
21. Zhan X, Zhao N, Plantinga A, et al. Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics*. 2017;206(4):1779-1790.
22. Davenport CA, Maity A, Sullivan PF, Tzeng JY. A powerful test for snp effects on multivariate binary outcomes using kernel machine regression. *Stat Biosci*. 2018;10(1):117-138.
23. Zhan X, Tong X, Zhao N, Maity A, Wu MC, Chen J. A small-sample multivariate kernel machine test for microbiome association studies. *Genet Epidemiol*. 2017;41(3):210-220.
24. Zhan X, Plantinga A, Zhao N, Wu MC. A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics*. 2017;73(4):1453-1463.
25. Li G, Liu X, Chen K. Integrative multi-view regression: bridging group-sparse and low-rank models. *Biometrics*. 2019;75(2):593-602.
26. Aitchison J. Principal component analysis of compositional data. *Biometrika*. 1983;70(1):57-65.
27. Aitchison J, Egozcue JJ. Compositional data analysis: where are we and where should we be heading? *Math Geol*. 2005;37:829-850.
28. Sun T, Zhang CH. Scaled sparse linear regression. *Biometrika*. 2012;99(4):879-898.
29. Negahban S, Wainwright MJ. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann Stat*. 2011;39(2):1069-1097.
30. Negahban S, Ravikumar P, Wainwright MJ, Yu B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Stat Sci*. 2012;27(4):538-557.
31. Mitra R, Zhang CH. The benefit of group sparsity in group inference with de-biased scaled group Lasso. *Electron J Stat*. 2016;10(2):1829-1873.
32. Huang J, Zhang T. The benefit of group sparsity. *Ann Stat*. 2010;38(4):1978-2004.
33. Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *J R Stat Soc Ser B Stat Methodol*. 2014;76(1):217-242.
34. Roy SN. On a heuristic method of test construction and its use in multivariate analysis. *Ann Math Stat*. 1953;24(2):220-238.
35. Wilson DJ. The harmonic mean p-value for combining dependent tests. *Proc Natl Acad Sci*. 2019;116(4):1195-1200.
36. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
37. Bomar L, Maltz M, Colston S, Graf J. Directed culturing of microorganisms using metatranscriptomics. *MBio*. 2011;2(2):e00012-e00011.
38. Cong X, Judge M, Xu W, et al. Influence of infant feeding type on gut microbiome development in hospitalized preterm infants. *Nurs Res*. 2017;66(2):123-133.
39. Lester BM, Tronick EZ. The neonatal intensive care unit network neurobehavioral scale procedures. *Pediatrics*. 2004;113(Suppl 2):641-667.
40. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol*. 1995;57(1):289-300.
41. Nuriel-Ohayon M, Neuman H, Koren O. Microbial changes during pregnancy, birth, and infancy. *Front Microbiol*. 2016;7:1031.
42. Koenig JE, Spor A, Scalfone N, et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci*. 2011;108(Suppl 1):4578-4585.
43. Cong X, Xu W, Janton S, et al. Gut microbiome developmental patterns in early life of preterm infants: impacts of feeding and gender. *PLoS One*. 2016;11(4):e0152751.

44. Labus J, Hsiao E, Tap J, et al. Clostridia from the gut microbiome are associated with brain functional connectivity and evoked symptoms in IBS. *Gastroenterologia*. 2017;152(5):S40.
45. Rudi K, Ricanek P, Tannæs T, Brackmann S, Perminow G, Vatn MH. Analysing tap-water from households of patients with inflammatory bowel disease in Norway. In: Hoorfar J, ed. *Case Studies in Food Safety and Authenticity*. Cambridge, England: Woodhead Publishing; 2012:130-137.
46. Gillespie SH. *Medical Microbiology Illustrated*. Oxford, UK: Butterworth-Heinemann; 1994.
47. Hanning I, Diaz-Sanchez S. The functionality of the gastrointestinal microbiome in non-human animals. *Microbiome*. 2015;3:51.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu X, Cong X, Li G, Maas K, Chen K. Multivariate log-contrast regression with sub-compositional predictors: Testing the association between preterm infants' gut microbiome and neurobehavioral outcomes. *Statistics in Medicine*. 2022;41(3):580-594. doi: 10.1002/sim.9273