Supporting Information for "Multivariate Log-Contrast Regression with Sub-Compositional Predictors: Testing the Association Between Preterm Infants' Gut Microbiome and Neurobehavioral Outcomes" by Xiaokang Liu, Xiaomei Cong, Gen Li, Kendra Maas, and Kun Chen

September 24, 2021

Web Appendix A Computation Details

In this section, computational algorithms are introduced to obtain the scaled iRRR estimator and the score matrix. For readers' convenience, we first reproduce the scaled composite nuclear norm penalization approach (3) and the score matrix estimation framework (8) of the main paper. They are given by

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

and

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

respectively. Since the intercept and the control variables can be treated as a group with penalty zero, instead of solving (1), we only need to focus on

$$
\begin{aligned}
(\hat{\mathbf{B}}^n, \hat{\sigma}) &= \arg\min_{\mathbf{B},\sigma} \mathbf{L}_{\mathbf{w}}(\mathbf{B}, \sigma) \\
&= \arg\min_{\mathbf{B},\sigma} \left\{ \frac{1}{2nq\sigma} \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \frac{\sigma}{2} + \lambda \sum_{k=1}^{K} w_k \|\mathbf{B}_k\|_* \right\}. \qquad (3)
\end{aligned}
$$

As for the two algorithms to solve (3), one is derived as a block-wise coordinate descent algorithm and another is built on the alternating direction method of multipliers (Boyd et al., 2011, ADMM). Both methods have good performance in our simulation. As for the score matrix estimation with (2), an ADMM algorithm is proposed.

A.1 Scaled iRRR Estimation

With a given σ, we have

$$
\sigma \mathbf{L}_{\mathbf{w}}(\mathbf{B}, \sigma) = \mathbf{L}_{\mathbf{w}^*}(\mathbf{B}) + \frac{\sigma^2}{2}
$$

where $\mathbf{w}^* = \sigma \mathbf{w} = (\sigma w_1, \ldots, \sigma w_K)^{\mathrm{T}}$ and $\mathbf{L}_{\mathbf{w}^*}(\mathbf{B})$ is the objective function in the original iRRR estimation framework (Li et al., 2019)

$$
\hat{\mathbf{B}}^n(\mathbf{w}) = \arg\min_{\mathbf{B} \in \mathbb{R}^{p \times q}} \mathbf{L}_{\mathbf{w}}(\mathbf{B}) = \arg\min_{\mathbf{B} \in \mathbb{R}^{p \times q}} \left\{ \frac{1}{2nq} \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda \sum_{k=1}^{K} w_k \|\mathbf{B}_k\|_* \right\}. \qquad (4)
$$

The notation $\hat{\mathbf{B}}^n(\mathbf{w})$ emphasizes the dependence of the estimator on the weight $\mathbf{w}$. Therefore, a block-wise coordinate descent algorithm can be applied to solve (3). Suppose at the k-th iteration, we have $\hat{\sigma}^{(k)}$ and $\hat{\mathbf{B}}^{n(k)}(\mathbf{w}^{(k)})$. Then at the $(k+1)$-th iteration, the updating

procedure is summarized as

$$\hat{\sigma}^{(k+1)} \leftarrow \|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{n(k)}(\mathbf{w}^{(k)})\|_F/\sqrt{nq}, \tag{5}$$

$$\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}\hat{\sigma}^{(k+1)}, \tag{6}$$

$$\hat{\mathbf{B}}^{n(k+1)}(\mathbf{w}^{(k+1)}) \leftarrow \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} \mathbf{L}_{\mathbf{w}^{(k+1)}}(\mathbf{B}). \tag{7}$$

We stop iteration when $\hat{\sigma}$ gets converged. Due to the joint convexity of (3), the estimates produced from the above iterative algorithm converge to the minimizer of (3), with $\hat{\mathbf{B}}^n = \hat{\mathbf{B}}^n(\hat{\sigma}\mathbf{w})$. In step (7), the optimization problem is solved by using an ADMM based algorithm described in Li et al. (2019).

An alternative is to directly apply ADMM to solve (3). Let $\mathbf{A}_k$ $(k = 1, \ldots, K)$ be the surrogate variables of $\mathbf{B}_k$ with the same dimension, we optimize

$$\min_{\mathbf{A}_k,\mathbf{B}_k,\sigma} \frac{1}{2nq\sigma}\left\|\mathbf{Y} - \sum_{k=1}^{K}\mathbf{X}_k\mathbf{B}_k\right\|_F^2 + \frac{\sigma}{2} + \lambda\sum_{k=1}^{K} w_k\|\mathbf{A}_k\|_*$$
$$\text{s.t. } \mathbf{A}_k = \mathbf{B}_k,\ k = 1, \ldots, K.$$

Let $\mathbf{\Lambda} = (\mathbf{\Lambda}_1^{\mathrm{T}}, \ldots, \mathbf{\Lambda}_K^{\mathrm{T}})^{\mathrm{T}}$ be the Lagrange parameter with each $\mathbf{\Lambda}_k \in \mathbb{R}^{p_k\times q}$, then the augmented Lagrangian objective function is

$$\begin{aligned} D(\mathbf{Y}; \mathbf{A}, \mathbf{B}, \sigma, \mathbf{\Lambda}) = {} & \frac{1}{2nq\sigma}\left\|\mathbf{Y} - \sum_{k=1}^{K}\mathbf{X}_k\mathbf{B}_k\right\|_F^2 + \frac{\sigma}{2} + \lambda\sum_{k=1}^{K} w_k\|\mathbf{A}_k\|_* \\ & + \sum_{k=1}^{K}\langle\mathbf{\Lambda}_k, \mathbf{A}_k - \mathbf{B}_k\rangle_F + \frac{\rho}{2}\sum_{k=1}^{K}\|\mathbf{A}_k - \mathbf{B}_k\|_F^2, \end{aligned}$$

where ρ is a pre-specified constant to control the step size. Let $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$, $\tilde{\sigma}$ and $\tilde{\mathbf{\Lambda}}$ be the estimates from the last iteration, then in the primal step we first update $(\mathbf{B}, \sigma)$ with the given $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{\Lambda}}$, secondly estimate $\mathbf{A}$ based on the updated $(\mathbf{B}, \sigma)$ and $\tilde{\mathbf{\Lambda}}$, and finally conduct

the dual step. Specifically, to update $(\mathbf{B}, \sigma)$ we first estimate $\mathbf{B}$ with

$$\hat{\mathbf{B}} = \left(\frac{1}{nq\tilde{\sigma}}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \rho\mathbf{I}_p\right)^{-1}\left(\frac{1}{nq\tilde{\sigma}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} + \tilde{\mathbf{\Lambda}} + \rho\tilde{\mathbf{A}}\right), \tag{8}$$

and then update σ with

$$\hat{\sigma} = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\|_F}{\sqrt{nq}}. \tag{9}$$

Here we remark that we only update $(\mathbf{B}, \sigma)$ once in each iteration, and it works well in simulation. As for the estimation of $\mathbf{A}$, note that the objective function is separable with respect to each $\mathbf{A}_k$ given $(\hat{\mathbf{B}}, \hat{\sigma}, \tilde{\mathbf{\Lambda}})$, i.e.,

$$D(\mathbf{Y}, \mathbf{A}_k, \hat{\mathbf{B}}, \hat{\sigma}, \tilde{\mathbf{\Lambda}}) = \lambda w_k\|\mathbf{A}_k\|_* + \left\langle \tilde{\mathbf{\Lambda}}_k, \mathbf{A}_k \right\rangle_F + \frac{\rho}{2}\|\mathbf{A}_k\|_F^2 - \rho\left\langle \mathbf{A}_k, \hat{\mathbf{B}}_k \right\rangle_F.$$

Minimizing the above objective function with respect to $\mathbf{A}_k$ is equivalent to solving

$$\min_{\mathbf{A}_k} \frac{1}{2}\|\mathbf{A}_k - (\hat{\mathbf{B}}_k - \tilde{\mathbf{\Lambda}}_k/\rho)\|_F^2 + \frac{\lambda w_k}{\rho}\|\mathbf{A}_k\|_*,$$

which has an explicit solution (Cai et al., 2010)

$$\hat{\mathbf{A}}_k = \mathbf{U}_k\mathcal{S}(\mathbf{D}_k, w_k\lambda/\rho)\mathbf{V}_k^{\mathrm{T}}, \tag{10}$$

where $\mathbf{U}_k$, $\mathbf{V}_k$ and $\mathbf{D}_k$ come from the singular value decomposition $(\hat{\mathbf{B}}_k - \tilde{\mathbf{\Lambda}}_k/\rho) = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^{\mathrm{T}}$, and $\mathcal{S}(\mathbf{D}_k, w_k\lambda/\rho) = (\mathbf{D}_k - w_k\lambda/\rho)_+$ is the soft-thresholding operator to all the diagonal elements of $\mathbf{D}_k$. Finally, based on the updated $\hat{\mathbf{A}}_k$ and $\hat{\mathbf{B}}_k$, the dual step is

$$\hat{\mathbf{\Lambda}}_k = \tilde{\mathbf{\Lambda}}_k + \rho\left(\hat{\mathbf{A}}_k - \hat{\mathbf{B}}_k\right), \ k = 1, \ldots, K. \tag{11}$$

For establishing the stopping rule, the primal residual and dual residual are defined as

$$r_{primal} = \|\hat{\mathbf{A}} - \hat{\mathbf{B}}\|_F,$$
$$r_{dual} = \rho\|\hat{\mathbf{A}} - \tilde{\mathbf{A}}\|_F. \tag{12}$$

Once both residuals fall below a pre-specified tolerance level, we stop the iteration. In practice, we can gradually increase the step size ρ to accelerate the algorithm (He et al., 2000). The procedure is summarized in Algorithm 1.

Algorithm 1 The ADMM algorithm to solve (3).

Parameter: λ, ρ.
Initialize $\mathbf{A}$, $\mathbf{B}$, σ and the Lagrange multiplier $\mathbf{\Lambda}$;
while The stopping criterion is not satisfied **do**

- Primal step:
  - Update $\mathbf{B}_k,\ k = 1, \ldots, K$ by (8);
  - Update σ by (9);
  - Update $\mathbf{A}_k,\ k = 1, \ldots, K$ by (10);
- Dual step:
  - Update $\mathbf{\Lambda}$ by (11);
- Calculate the primal and dual residuals defined in (12);
- (Optional) Increase ρ by a small amount, e.g., $\rho \leftarrow 1.01\rho$.

end while

A.2 Score Matrix Estimation

In order to obtain the score matrix, we need to solve (2) which can be formulated as

$$\min_{\mathbf{A}_k, \mathbf{B}_k} \frac{1}{2n}\|\mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_k\mathbf{B}_k\|_F^2 + \lambda \sum_{k=1}^{K} w_k\|\mathbf{X}_k\mathbf{A}_k\|_* \tag{13}$$
$$\text{s.t. } \mathbf{X}_k\mathbf{A}_k = \mathbf{X}_k\mathbf{B}_k,\ k = 1, \ldots, K.$$

Different from the original iRRR estimation framework (4), in (13) we penalize the nuclear norm of the group effect $\mathbf{X}_k\mathbf{B}_k$ directly. The ADMM algorithm proposed in Li et al. (2019) can be applied here with a small modification, i.e., based on $\hat{\mathbf{B}}$ and $\tilde{\mathbf{\Lambda}}$ we need to update $\mathbf{X}_k\mathbf{A}_k$ but not $\mathbf{A}_k$ from

$$\min_{\mathbf{A}_k\mathbf{X}_k} \frac{1}{2}\|\mathbf{X}_k\mathbf{A}_k - (\mathbf{X}_k\hat{\mathbf{B}}_k - \tilde{\mathbf{\Lambda}}_k/\rho)\|_F^2 + \frac{\lambda w_k}{\rho}\|\mathbf{X}_k\mathbf{A}_k\|_*.$$

By conducting singular value decomposition to $\mathbf{X}_k\hat{\mathbf{B}}_k - \tilde{\mathbf{\Lambda}}_k/\rho$ and applying the soft thresholding operator to its singular values with the threshold value $\lambda w_k/\rho$ we can update $\mathbf{X}_k\mathbf{A}_k$. Accordingly, we have

$$\begin{aligned}
\hat{\mathbf{\Lambda}}_k &= \tilde{\mathbf{\Lambda}}_k + \rho(\mathbf{X}_k\hat{\mathbf{A}}_k - \mathbf{X}_k\hat{\mathbf{B}}_k),\\
r_{primal} &= \|\mathbf{X}\hat{\mathbf{A}} - \mathbf{X}\hat{\mathbf{B}}\|_F,\\
r_{dual} &= \rho\sum_{k=1}^{K}\|\mathbf{X}_k^{\mathrm{T}}(\mathbf{X}\hat{\mathbf{A}} - \mathbf{X}\tilde{\mathbf{A}})\|_F.
\end{aligned}$$

Once both residuals fall below a pre-specified tolerance level we stop the algorithm.

Web Appendix B Proof of Theorem 1

Proof of Theorem 1. We follow the proof in Mitra and Zhang (2016). With $\eta > 0$, first define

$$\mu(\mathbf{w},\eta) = \frac{8(1+\eta)(2+\eta)}{\eta^2}\frac{\sum_{k=1}^{K} qr_k\lambda^2 w_k^2}{\kappa(\mathbf{X})},\ \tau_+ = \frac{2+\eta}{1+\eta}\mu(\mathbf{w},\eta),\ \tau_- = \frac{2}{1+\eta}\mu(\mathbf{w},\eta), \quad (14)$$

and an event

$$\mathcal{E} = \cap_{k=1}^{K}\mathcal{A}_k = \cap_{k=1}^{K}\left\{\frac{d_1(\mathbf{X}_k^{\mathrm{T}}\mathbf{E})}{nq\sigma^*/\sqrt{1+\tau_-}} \le \frac{\lambda w_k}{1+\eta}\right\}.$$

Let $\boldsymbol{\Delta} = \hat{\mathbf{B}}^n(t\mathbf{w}) - \mathbf{B}^*$, $\boldsymbol{\Delta}_k = \hat{\mathbf{B}}_k^n(t\mathbf{w}) - \mathbf{B}_k^*$, $\hat{\sigma}^2(t\mathbf{w}) = \|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\|_F^2/(nq)$ and $t \geq \sigma^*/\sqrt{1+\tau_-}$, then

$$\begin{aligned}
\sigma^{*2} - \hat{\sigma}^2(t\mathbf{w}) &= \frac{\|\mathbf{Y} - \mathbf{X}\mathbf{B}^*\|_F^2}{nq} - \frac{\|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\|_F^2}{nq} \\
&= \frac{\langle \mathbf{X}\boldsymbol{\Delta}, 2\mathbf{E} - \mathbf{X}\boldsymbol{\Delta}\rangle_F}{nq} \\
&= \frac{\langle \mathbf{X}\boldsymbol{\Delta}, \mathbf{Y} + \mathbf{E} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\rangle_F}{nq} \\
&= \frac{\langle \mathbf{X}\boldsymbol{\Delta}, \mathbf{E}\rangle_F}{nq} + \frac{\langle \mathbf{X}\boldsymbol{\Delta}, \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\rangle_F}{nq}. \qquad (15)
\end{aligned}$$

We first deal with the first term on the right hand side of (15)

$$\begin{aligned}
\frac{|\langle \mathbf{X}\boldsymbol{\Delta}, \mathbf{E}\rangle_F|}{nq} &\leq \sum_{k=1}^K \frac{d_1(\mathbf{X}_k^{\mathrm{T}}\mathbf{E})\|\boldsymbol{\Delta}_k\|_*}{nq} \\
&\leq \frac{\lambda t}{1+\eta}\sum_{k=1}^K w_k\|\boldsymbol{\Delta}_k\|_*. \qquad (16)
\end{aligned}$$

The last inequality is built on the event $\mathcal{E}$ with $t \geq \sigma^*/\sqrt{1+\tau_-}$. Next we deal with the second term on the right hand side of (15)

$$\frac{|\langle \mathbf{X}\boldsymbol{\Delta}, \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\rangle_F|}{nq} \leq \frac{1}{nq}\sum_{k=1}^K \|\boldsymbol{\Delta}_k\|_* d_1(\mathbf{X}_k^{\mathrm{T}}\{\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\}).$$

In order to bound $d_1(\mathbf{X}_k^{\mathrm{T}}\{\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\})$, recall that $\hat{\mathbf{B}}^n(t\mathbf{w})$ is a minimizer of $\mathbf{L}_{t\mathbf{w}}(\mathbf{B})$ if and only if there exists a diagonal matrix $\mathbf{J}_k$ with $d_1(\mathbf{J}_k) \leq 1$ such that

$$\mathbf{X}_k^{\mathrm{T}}\{\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\} = \lambda t n q w_k \mathbf{U}_k\mathbf{J}_k\mathbf{V}_k^{\mathrm{T}},\ k = 1, \ldots, K,$$

where $\mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^{\mathrm{T}}$ is the singular value decomposition of $\hat{\mathbf{B}}_k^n(t\mathbf{w})$ (Watson, 1992). Thus, for each k we have

$$d_1(\mathbf{X}_k^{\mathrm{T}}\{\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}(t\mathbf{w})\}) \leq \lambda t n q w_k$$

and

$$\frac{|\langle \mathbf{X}\boldsymbol{\Delta}, \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\rangle_F|}{nq} \leq t\lambda \sum_{k=1}^{K} w_k \|\boldsymbol{\Delta}_k\|_*. \tag{17}$$

Then, from

$$\frac{\langle \mathbf{X}\boldsymbol{\Delta}, 2\mathbf{E} - \mathbf{X}\boldsymbol{\Delta}\rangle_F}{nq} \leq \frac{\langle \mathbf{X}\boldsymbol{\Delta}, 2\mathbf{E}\rangle_F}{nq}$$

and inequality (16) we have

$$\sigma^{*2} - \hat{\sigma}^2(t\mathbf{w}) \leq \frac{|\langle \mathbf{X}\boldsymbol{\Delta}, 2\mathbf{E}\rangle_F|}{nq} \leq \frac{2t\lambda}{1+\eta} \sum_{k=1}^{K} w_k \|\boldsymbol{\Delta}_k\|_*, \tag{18}$$

and from (15), (16) and (17) we have

$$\begin{aligned} \sigma^{*2} - \hat{\sigma}^2(t\mathbf{w}) &\geq -\frac{|\langle \mathbf{X}\boldsymbol{\Delta}, \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}(t\mathbf{w})\rangle_F|}{nq} - \frac{|\langle \mathbf{X}\boldsymbol{\Delta}, \mathbf{E}\rangle_F|}{nq} \\ &\geq -\frac{t\lambda(2+\eta)}{1+\eta} \sum_{k=1}^{K} w_k \|\boldsymbol{\Delta}_k\|_*. \end{aligned} \tag{19}$$

Therefore, we have

$$-\frac{t\lambda(2+\eta)}{1+\eta} \sum_{k=1}^{K} w_k \|\boldsymbol{\Delta}_k\|_* \leq \sigma^{*2} - \hat{\sigma}^2(t\mathbf{w}) \leq \frac{2t\lambda}{1+\eta} \sum_{k=1}^{K} w_k \|\boldsymbol{\Delta}_k\|_*. \tag{20}$$

Next, we derive the rate of $\sum_{k=1}^{K} w_k \|\boldsymbol{\Delta}_k\|_*$ by analyzing

$$\frac{\hat{\mathbf{B}}^n(t\mathbf{w})}{t} = \arg\min_{\mathbf{B} \in \mathbb{R}^{p \times q}} \left\{ \frac{1}{2nq} \|\mathbf{Y}/t - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \sum_{k=1}^{K} w_k \|\mathbf{B}_k\|_* \right\}. \tag{21}$$

Since $t \geq \sigma^*/\sqrt{1+\tau_-}$, event $\mathcal{E}$ leads to

$$\cap_{k=1}^{K} \left\{ \frac{d_1(\mathbf{X}_k^{\mathrm{T}}\mathbf{E}/t)}{nq} \leq \frac{\lambda w_k}{1+\eta} \right\},$$

which facilitates the application of Theorem 2 in Li et al. (2019) to (21) and we get

$$t^{-1}\lambda\sum_{k=1}^{K} w_k\|\mathbf{\Delta}_k\|_* = \lambda\sum_{k=1}^{K} w_k\|\hat{\mathbf{B}}_k^n(t\mathbf{w})/t - \mathbf{B}_k^*/t\|_* \leq \frac{8(1+\eta)(2+\eta)}{\eta^2}\frac{\sum_{k=1}^{K} r_k q\lambda^2 w_k^2}{\kappa(\mathbf{X})} = \mu(\mathbf{w},\eta).$$

It follows that

$$t\lambda\sum_{k=1}^{K} w_k\|\mathbf{\Delta}_k\|_* \leq t^2\mu(\mathbf{w},\eta), \tag{22}$$

which together with (20) leads to

$$-\frac{2+\eta}{1+\eta}t^2\mu(\mathbf{w},\eta) \leq \sigma^{*2} - \hat{\sigma}^2(t\mathbf{w}) \leq \frac{2}{1+\eta}t^2\mu(\mathbf{w},\eta).$$

Recall the definition of τ_+ and τ_- in (14), we have

$$-\tau_+ t^2 \leq \sigma^{*2} - \hat{\sigma}^2(t\mathbf{w}) \leq \tau_- t^2. \tag{23}$$

The second inequality in (23) with $t = \sigma^*/\sqrt{1+\tau_-}$ leads to $t^2 - \hat{\sigma}^2(t\mathbf{w}) \leq t^2 - \sigma^{*2} + \tau_- t^2 = 0$, which indicates $\hat{\sigma}(t\mathbf{w}) \geq t = \sigma^*/\sqrt{1+\tau_-}$. Assume $\sigma^*/\sqrt{1-\tau_+} \geq \sigma^*/\sqrt{1+\tau_-}$, then the first inequality of (23) with $t = \sigma^*/\sqrt{1-\tau_+}$ implies $t^2 - \hat{\sigma}^2(t\mathbf{w}) \geq t^2 - \sigma^{*2} - \tau_+ t^2 = 0$, i.e., $\hat{\sigma}(t\mathbf{w}) \leq t = \sigma^*/\sqrt{1-\tau_+}$. Due to the joint convexity of the scaled iRRR framework (3), we have $\hat{\sigma}(t\mathbf{w}) = \|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n(t\mathbf{w})\|_F/\sqrt{nq}$ converges to $\hat{\sigma}$ which is the minimizer of (3) and consequently we have

$$\frac{\sigma^*}{\sqrt{1+\tau_-}} \leq \hat{\sigma} \leq \frac{\sigma^*}{\sqrt{1-\tau_+}}$$

and

$$\left|\frac{\hat{\sigma}}{\sigma^*} - 1\right| = o_p(\mu(\mathbf{w},\eta)).$$

If we have $\sqrt{nq}\mu(\mathbf{w},\eta)\to 0$, then $\left|\frac{\hat{\sigma}}{\sigma^*}-1\right| = o_p((nq)^{-1/2})$. Moreover, if $vec(\mathbf{E})\sim\mathcal{N}_{nq}(0,\sigma^2\mathbf{I}_{nq})$ we have $\sigma^*/\sigma\sim\chi_{nq}/\sqrt{nq}$. Then, by central limit theorem we get

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

Consequently, we can prove

$$\sqrt{nq}\left(\frac{\hat{\sigma}}{\sigma}-1\right)\to\mathcal{N}\left(0,\frac{1}{2}\right). \tag{24}$$

Next we derive the estimation error bound for $\hat{\mathbf{B}}^n(\hat{\sigma}\mathbf{w})$ (i.e., $\hat{\mathbf{B}}^n$) under the framework (21). Since $\hat{\sigma}\geq\sigma^*/\sqrt{1+\tau_-}$, the estimation error bounds are

$$\begin{aligned}\|\hat{\mathbf{B}}^n-\mathbf{B}^*\|_F^2 &\preceq \frac{\sigma^{*2}\sum_{k=1}^K r_k q^2\lambda^2 w_k^2}{(1-\tau_+)\kappa^2(\mathbf{X})},\\ \sum_{k=1}^K \lambda w_k\|\hat{\mathbf{B}}_k^n-\mathbf{B}_k^*\|_* &\preceq \frac{\sigma^*\sum_{k=1}^K r_k q\lambda^2 w_k^2}{\sqrt{1-\tau_+}\kappa(\mathbf{X})}\end{aligned}$$

by applying Theorem 2 in Li et al. (2019) and the fact that $\hat{\sigma}\leq\sigma^*/\sqrt{1-\tau_+}$. Finally, we prove $\mathbf{P}(\mathcal{E})>1-\epsilon$ with some $0<\epsilon<1$. Note that, follow the same reasoning as the proof of Theorem 6 in Mitra and Zhang (2016) with the assumption $vec(\mathbf{E})\sim\mathcal{N}_{nq}(0,\sigma^2\mathbf{I}_{nq})$, it can be verified that if we let $w_k = d_1(\mathbf{X}_k)\left\{\sqrt{p_k/n}+\sqrt{2\log(K/\epsilon)/(nq)}\right\}/\sqrt{nq}$ with a properly selected λ then we have $\mathbf{P}(\mathcal{E})>1-\epsilon$. This completes the proof.

□

Web Appendix C A Brief Overview of High-Dimensional Inference Procedures and the LDPE Approach

Researches on statistical inference for regularized estimators emerged in recent years as the prevailing of high-dimensional statistics. Regularized estimation methods are commonly used in high dimensional linear regression problems, e.g., lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), and group lasso (Yuan and Lin, 2006). However, due to the regularization, the resulting estimator is often biased and not in an explicit form, making its sampling distribution complicated and even intractable. In order to account for uncertainty in estimation and assess the selected model, several methods have been proposed for assigning p-values and constructing confidence intervals for a single or a group of coefficients. See, e.g., Knight and Fu (2000); Wasserman and Roeder (2009); Meinshausen et al. (2009); Chatterjee and Lahiri (2013); Ning and Liu (2017); Shi et al. (2019). One popular class of method utilizes the projection and bias-correction technique, where a de-biasing procedure is first applied to the regularized estimator and then the asymptotic distribution is derived for the resulting estimator. For example, Bühlmann (2013) applied bias correction to a Ridge estimator and derived an inference procedure, Zhang and Zhang (2014) and Javanmard and Montanari (2014) considered the lasso estimator. Shi et al. (2016) generalized the procedure of Javanmard and Montanari (2014) to make inference for lasso estimator obtained under multiple linear constraints on coefficients. To facilitate chi-square type hypothesis testing for a possibly large group of coefficients without inflating the required sample size due to group size, Mitra and Zhang (2016) generalized the idea of the low-dimensional projecting estimator (Zhang and Zhang, 2014, LDPE) to correct the bias of a scaled group lasso estimator. Although the above methods are effective under various model settings, to the best of our knowledge, so far there is not much work focus on inference in high-dimensional multivariate regression, especially for rank restricted models, which motivates

the derivation of the inference method considered in this paper.

It is worthwhile to dive deeper into the LDPE approach proposed by Zhang and Zhang (2014), as our proposed method will be built upon it. The illustration proceeds under a multiple linear regression model $\mathbf{y} = \mathbf{x}_1\beta_1 + \ldots + \mathbf{x}_p\beta_p + \epsilon$, where $\mathbf{y} \in \mathbb{R}^n$ is the response vector and $\mathbf{x}_j \in \mathbb{R}^n$ is a vector consisting of observations of the j-th predictor. Suppose we are interested in the effect of predictor $\mathbf{x}_j$ $(1 \leq j \leq p)$ on the response. The initial estimator $\hat{\beta}_j^n$ can be obtained by lasso method. As we mentioned before, lasso estimator is biased due to the regularization on coefficients. The effect of $\mathbf{x}_j$ on response cannot be fully represented by $\hat{\beta}_j^n$, hence a properly selected score vector is used to recover the part of information that is lost in regularization. The resulting LDPE $\hat{\beta}_j$ can be written as

$$\hat{\beta}_j = \hat{\beta}_j^n + \frac{\mathbf{z}_j^{\mathrm{T}}(\mathbf{y} - \sum_{l=1}^p \mathbf{x}_l\hat{\beta}_l^n)}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j}, \tag{25}$$

where the score vector $\mathbf{z}_j$ has the same dimension as $\mathbf{x}_j$ and only depends on the design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. The score vector $\mathbf{z}_j$ serves as a tool to extract the information that is only related to $\mathbf{x}_j$ from the residual, then to correct the bias this part of effect is added back to $\hat{\beta}_j^n$ after standardization.

The classical scenario with $n > p$ can help us understand the mechanism of the above procedure better. When $n > p$, $\mathbf{z}_j$ can be set as $\mathbf{x}_j^{\perp}$, the projection of $\mathbf{x}_j$ onto the orthogonal complement of the column space of $\mathbf{X}_{-j}$ (the design matrix with the j-th column deleted). This choice of $\mathbf{z}_j$ can be regarded as the information only carried by the j-th predictor and satisfies $\mathbf{z}_j^{\mathrm{T}}\mathbf{X}_{-j} = \mathbf{0}$. Then whatever the initial estimator is, the resulting de-biased estimator is the least square estimator which is unbiased. However, in order to satisfy $\mathbf{z}_j^{\mathrm{T}}\mathbf{X}_{-j} = \mathbf{0}$ when $p < n$, $\mathbf{z}_j$ needs to be a zero vector which consequently makes (25) ineffective. Therefore, in order to apply the de-biasing procedure (25) in the high-dimensional scenario, we have to relax the requirement $\mathbf{z}_j^{\mathrm{T}}\mathbf{X}_{-j} = \mathbf{0}$, i.e., to approximate $\mathbf{x}_j^{\perp}$ to control the bias caused by $\mathbf{z}_j^{\mathrm{T}}\mathbf{X}_{-j} \neq \mathbf{0}$ to be under a tolerable level. For example, the score vector

is obtained by applying lasso to the regression of $\mathbf{x}_j$ on $\mathbf{X}_{-j}$ in Zhang and Zhang (2014), while in Javanmard and Montanari (2014) the score vector is estimated from an optimization program that minimizes the variance of the de-biased estimator while control its bias.

Web Appendix D Derivation of The Inference Procedure

In this section, we introduce the main steps of establishing our proposed method follow Mitra and Zhang (2016), which include (1) the exploitation of LDPE to correct the bias of the scaled iRRR estimator, (2) the construction of a χ^2-type test statistic based on the de-biased estimator, and (3) the estimation of the required score matrix and the derivation of the theoretical guarantee of the reliability of the test. The condition $\text{rank}(\mathbf{S}_k^{\mathrm{T}}\mathbf{X}_k) = \text{rank}(\mathbf{X}_k)$ is required to guarantee the effectiveness of de-biasing, under which the role of $\mathbf{S}_k$ in the de-biasing procedure can be totally replaced by $\mathbf{P}_k$.

First, we provide the de-biased scaled iRRR estimator based on the notations defined in the main paper. With the scaled iRRR estimator $\hat{\mathbf{B}}^n = (\hat{\mathbf{B}}_1^{n\mathrm{T}}, \ldots, \hat{\mathbf{B}}_K^{n\mathrm{T}})^{\mathrm{T}}$ from (3), the de-biased estimator of $\mathbf{B}_k$ is

$$\hat{\mathbf{B}}_k = \hat{\mathbf{B}}_k^n + (\mathbf{S}_k'\mathbf{X}_k)^+\mathbf{S}_k^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n), \tag{26}$$

where $(\mathbf{S}_k'\mathbf{X}_k)^+$ is the Moore-Penrose inverse of $\mathbf{S}_k'\mathbf{X}_k$. For the group effect $\mathbf{X}_k\mathbf{B}_k$, the related de-biased estimator is

$$\mathbf{X}_k\hat{\mathbf{B}}_k = \mathbf{X}_k\hat{\mathbf{B}}_k^n + (\mathbf{P}_k\mathbf{Q}_k)^+\mathbf{P}_k(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^n). \tag{27}$$

Next, based on the de-biased estimator, we introduce a test statistic and derive its asymp-

totic distribution under the null. Note that, if $\text{rank}(\mathbf{X}_k) = p_k$ we have

$$(\mathbf{P}_k\mathbf{X}_k)(\hat{\mathbf{B}}_k - \mathbf{B}_k^*) = \mathbf{P}_k\mathbf{E} - \text{Rem}_k \tag{28}$$

with

$$\text{Rem}_k = \mathbf{P}_k \sum_{j\neq k}(\mathbf{X}_j\hat{\mathbf{B}}_j^n - \mathbf{X}_j\mathbf{B}_j^*), \tag{29}$$

and if $\text{rank}(\mathbf{X}_k) < p_k$, we can only make inference on $\mathbf{X}_k\mathbf{B}_k^*$ with

$$(\mathbf{P}_k\mathbf{Q}_k)(\mathbf{X}_k\hat{\mathbf{B}}_k - \mathbf{X}_k\mathbf{B}_k^*) = \mathbf{P}_k\mathbf{E} - \text{Rem}_k.$$

The effect of de-biasing in $\hat{\mathbf{B}}_k$ and $\mathbf{X}_k\hat{\mathbf{B}}_k$ is controlled by the approximation of $\mathbf{S}_k$ to $\mathbf{X}_k^\perp$ and the distance between $\hat{\mathbf{B}}^n$ and $\mathbf{B}^*$, where $\mathbf{X}_k^\perp$ is the best score matrix only available in the 'low-dimensional' scenario and is defined as the projection of $\mathbf{X}_k$ onto the orthogonal complement of the column space spanned by $(\mathbf{X}_1, \ldots, \mathbf{X}_{k-1}, \mathbf{X}_{k+1}, \ldots, \mathbf{X}_K)$. These two factors can be jointly measured by Rem_k. Therefore, once the magnitude of Rem_k is ignorable in the sense that

$$\sqrt{qr_k'}|\sigma/\hat{\sigma} - 1| + \|\text{Rem}_k/\sigma\|_F = o_p(1), \tag{30}$$

we have the approximation $\|\mathbf{P}_k\mathbf{E} - \text{Rem}_k\|_F^2/\hat{\sigma}^2 \to \|\mathbf{P}_k\mathbf{E}/\sigma\|_F^2$, which together with the normal assumption on the random error matrix implies $\|\mathbf{P}_k\mathbf{E} - \text{Rem}_k\|_F^2/\hat{\sigma}^2 \to \chi^2_{r_k'q}$ where $r_k' = \text{rank}(\mathbf{P}_k) = \text{rank}(\mathbf{X}_k)$. If in the true model $\mathbf{B}_k^* = \mathbf{0}$ or $\mathbf{X}_k\mathbf{B}_k^* = \mathbf{0}$, then since $\mathbf{Y} = \mathbf{X}_{-k}\mathbf{B}_{-k}^* + \mathbf{E}$ we have

$$\mathbf{P}_k\mathbf{E} - \text{Rem}_k = \mathbf{P}_k(\mathbf{Y} - \mathbf{X}_{-k}\hat{\mathbf{B}}_{-k}^n). \tag{31}$$

Therefore, the test statistic is

$$T_k = \frac{1}{\hat{\sigma}^2} \left\| \mathbf{P}_k (\mathbf{Y} - \sum_{j \neq k} \mathbf{X}_j \hat{\mathbf{B}}_j^n) \right\|_F^2 \overset{H_0}{\sim} \chi^2_{r'_k q} \tag{32}$$

asymptotically. We shall note that if $\text{rank}(\mathbf{X}_k) < p_k$, $\mathbf{B}_k^*$ is not identifiable, the method is only applicable to test $\text{H}_0 : \mathbf{X}_k \mathbf{B}_k^* = \mathbf{0}$ vs. $\text{H}_1 : \mathbf{X}_k \mathbf{B}_k^* \neq \mathbf{0}$ and when $\text{rank}(\mathbf{X}_k) = p_k$, the method is also applicable to test $\text{H}_0 : \mathbf{B}_k^* = \mathbf{0}$ vs. $\text{H}_1 : \mathbf{B}_k^* \neq \mathbf{0}$.

In order to implement and validate this test procedure, in addition to the scaled iRRR estimator $(\hat{\mathbf{B}}^n, \hat{\sigma})$, we also need to find $\mathbf{P}_k$ and verify condition (30). One key ingredient to verify (30) is to make $\|\text{Rem}_k/\sigma\|_F = o_p(1)$. Recall the form of Rem_k, we have

$$\begin{aligned}
\frac{\|\text{Rem}_k\|_F}{\sigma\sqrt{nq}} &\leq \frac{\sum_{j \neq k} \|\mathbf{P}_k \mathbf{X}_j (\hat{\mathbf{B}}_j^n - \mathbf{B}_j^*)\|_F}{\sigma\sqrt{nq}} \\
&\leq \sum_{j \neq k} \frac{d_1(\mathbf{P}_k \mathbf{Q}_j)}{w_{*,j}\sigma\sqrt{nq}} w_{*,j} \|\mathbf{X}_j \hat{\mathbf{B}}_j^n - \mathbf{X}_j \mathbf{B}_j^*\|_F \\
&\leq \max_{j \neq k} \frac{d_1(\mathbf{P}_k \mathbf{Q}_j)}{w_{*,j}} \sum_{j=1}^K \frac{w_{*,j}}{\sigma\sqrt{nq}} \|\mathbf{X}_j \hat{\mathbf{B}}_j^n - \mathbf{X}_j \mathbf{B}_j^*\|_F \\
&= O_p(q \sum_{j=1}^K r_j w_j^2) \max_{j \neq k} \frac{d_1(\mathbf{P}_k \mathbf{Q}_j)}{w_{*,j}},
\end{aligned}$$

which leads to

$$\frac{\|\text{Rem}_k\|_F}{\sigma} = O_p\left(\sum_{k=1}^K \frac{r_k \{p_k q + 2\log(K/\epsilon)\}}{\sqrt{nq}} \right) \eta_k \tag{33}$$

where $\eta_k = \max_{j \neq k} d_1(\mathbf{P}_k \mathbf{Q}_j)/w_{*,j}$ is dominated by $d_1(\mathbf{P}_k \mathbf{Q}_j)$. Thus, an ideal $\mathbf{P}_k$ needs to minimize the variance of the resulting de-biased estimator while control the magnitude of $d_1(\mathbf{P}_k \mathbf{Q}_j)$. Mitra and Zhang (2016) derived the following optimization framework

$$\mathbf{P}_k = \arg\min_{\mathbf{P}} \{ d_1(\mathbf{P}(\mathbf{I} - \mathbf{Q}_k)) : \mathbf{P} = \mathbf{P}^2 = \mathbf{P}^{\text{T}},\ d_1(\mathbf{P}\mathbf{Q}_j) \leq w'_j, \forall j \neq k \} \tag{34}$$

to solve out $\mathbf{P}_k$. In (34), w_j' is an upper bound of $d_1(\mathbf{P}\mathbf{Q}_j)$ and $d_1(\mathbf{P}_k(\mathbf{I}-\mathbf{Q}_k))$ measures the distance between the subspaces spanned by $\mathbf{P}_k$ and $\mathbf{I}-\mathbf{Q}_k$, which may inflate the variance of the de-biased estimator. The feasibility of (34) with a given w_j' has been verified for random designs with sub-Gaussian rows, refer to Theorem 4 and Lemma 1 in Mitra and Zhang (2016) for details. Since (34) has not been solved yet, in practice, $\mathbf{P}_k$ can be estimated from a penalized multivariate regression (2). Then with the conditions in Theorem 2 of the main paper, we can verify (30) thus validate the inference procedure.

Web Appendix E Proof of Theorem 2

Proof of Theorem 2. First we get the rate of $\|\text{Rem}_k\|_F/\sigma$. From the KKT condition of (2), we have $d_1(\mathbf{Q}_j\mathbf{S}_k/\sqrt{n}) \leq \xi w_j''$, which implies $d_1(\mathbf{Q}_j\mathbf{P}_k/\sqrt{n})d_{min}(\mathbf{S}_k) \leq \xi w_j''$. If we let $w_j'' = w_{*,j}$, then together with (33) and the condition

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

we have $\|\text{Rem}_k\|_F/\sigma = o_p(1)$. Then we consider the rate of $|1-\sigma/\hat{\sigma}|$. From (24) we can get

$$\left|1-\frac{\sigma}{\hat{\sigma}}\right| = O_p\left(\frac{1}{\sqrt{nq}}\right),$$

which together with $r_k'/n \to 0$ implies

$$\left|1-\frac{\sigma}{\hat{\sigma}}\right| = o_p\left(\frac{1}{\sqrt{r_k'q}}\right).$$

Combine these two results we complete the verification of (30). □

Web Appendix F Simulation with Compositional Data

F.1 Simulation with generated compositional data

We conduct simulations based on generated compositional data with a similar setting as the preterm infant dataset. Specifically, we let $n = 300$, $p = 60$, $q = 10$, $K = 10$ and each group is of size 6. Similar to Shi et al. (2016), we obtain vectors of count $\mathbf{w}_i = (\mathbf{w}_{1,i}^{\mathrm{T}}, \ldots, \mathbf{w}_{K,i}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p,\ i = 1, \ldots, n$ from a log-normal distribution $\ln \mathcal{N}_p(\boldsymbol{\mu}^w, \boldsymbol{\Sigma}^w)$. In order to reflect the difference in abundance of each taxon in the microbiome counts observation, we let $\boldsymbol{\mu}_k^w = (10, 1, 1, 1, 1, 1)$ for $k = 1, \ldots, 5$ and $\boldsymbol{\mu}_k^w = (1, 1, 1, 1, 1, 1)$ for $k = 6, \ldots, 10$, where $\boldsymbol{\mu}_k^w$ is the mean vector corresponding to the k-th group. To simulate the commonly existing correlation among counts of taxa, we let $\boldsymbol{\Sigma}^w = (\rho_x^{|i-j|})$ with $\rho_x \in \{0.2, 0.5\}$. We then transform the count data into sub-compositional data, i.e., $z_{k,i,j} = w_{k,i,j} / \sum_{j=1}^{p_k} w_{k,i,j},\ k = 1 \ldots, K;\ i = 1, \ldots, n$, where $w_{k,i,j}$ is the count of the j-th taxon within the k-th group of the i-th subject. To see the potential effect of the existence of highly abundant taxon on the group inference results, we select the first and the sixth group as two predictive groups, and let $\mathrm{rank}(\mathbf{B}_1^*) = \mathrm{rank}(\mathbf{B}_6^*) = 1$ with all the other groups have no contribution. After further rescaling $\mathbf{B}_k^*$ to make its largest entry to be 0.2, we obtain $\mathbf{Y}$ from (2) with $\mathrm{SNR} \in \{0.2, 0.4, 0.8\}$. The noise level estimation results and the testing results for two predictive groups and two irrelevant groups are summarized from 300 replications.

F.2 Simulation with real compositional data

The data collected from the preterm infant study have the following structure, $p = 62$, $q = 10$, $K = 11$ and the group size is $(p_1, \ldots, p_K) = (3, 2, 3, 4, 7, 15, 2, 9, 3, 2, 12)$. From all the 11 groups, we select the first and the sixth group to be predictive to the response with $r_1^* = r_6^* = 1$, and all the remaining groups have no prediction contribution, i.e., $r_k^* = 0,\ k \notin \{1, 6\}$. To control the signal strength, we make the largest entry in $\mathbf{B}^*$ to be 0.2. By resampling with replacement, each time we obtain a dataset with sample size $n = 300$

and then apply the proposed method. The whole procedure is repeated 300 times, and the results are displayed in Web Table 1. In general, when the dataset has a weak signal strength, the test power is low; when the signal strength becomes larger, the power of the test also increases. Moreover, the multivariate method has a larger test power than the univariate method. As expected, the magnitude of $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ affects the performance of the test. Specifically, groups 1, 2, 3, 7, 9, 10 have smaller $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ values than groups 4, 5, 6, 8, 11 whose $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ values are close to 1. The magnitude of inflation of the false positive rate for the groups with larger $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ values is larger than that for groups with smaller $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ values. Therefore, it is necessary to pay more attention to the groups that have large $d_1(\mathbf{P}_k(\mathbf{I}_n - \mathbf{Q}_k))$ values.

Web Table 1 Simulation results based on the resampled real microbiome compositional data across 300 replications. The performance of noise level estimation is displayed in terms of the mean ($\times 100$) and standard error ($\times 100$, in parenthesis) of $\hat{\sigma}/\sigma - 1$ and $|\hat{\sigma}/\sigma - 1|$, respectively. Each group is denoted as "G" followed by its group number.

SNR	$\hat{\sigma}/\sigma - 1$	$\lvert\hat{\sigma}/\sigma - 1\rvert$	G1 TP	G2 FP	G3 FP	G4 FP	G5 FP	G6 TP	G7 FP	G8 FP	G9 FP	G10 FP	G11 FP
Multivariate Inference													
0.25	0.74 (1.36)	1.23 (0.94)	0.05	0.06	0.05	0.06	0.06	0.14	0.06	0.07	0.05	0.07	0.10
0.50	3.65 (1.40)	3.66 (1.37)	0.09	0.04	0.04	0.07	0.04	0.55	0.05	0.09	0.03	0.07	0.37
1.00	6.18 (5.61)	6.36 (5.41)	0.32	0.04	0.02	0.05	0.03	1.00	0.06	0.22	0.02	0.07	0.63
Univariate Inference (Bonferroni)													
0.25			0.05	0.04	0.04	0.05	0.02	0.10	0.03	0.06	0.04	0.05	0.06
0.50			0.07	0.03	0.04	0.04	0.02	0.31	0.03	0.07	0.04	0.05	0.19
1.00			0.28	0.02	0.02	0.06	0.03	0.98	0.03	0.16	0.02	0.08	0.86
Univariate Inference (HMP)													
0.25			0.04	0.04	0.05	0.06	0.02	0.10	0.03	0.07	0.04	0.05	0.07
0.50			0.07	0.04	0.03	0.05	0.02	0.33	0.03	0.07	0.04	0.05	0.21
1.00			0.29	0.02	0.02	0.06	0.03	0.99	0.03	0.17	0.02	0.07	0.91

Web Appendix G Additional Application Results

Web Table 2 lists the estimated coefficients of the control variables from the overall model. In terms of the sub-scale stress/abstinence, the signs of the estimated coefficients are the same as the results from Sun et al. (2020). Stress/abstinence is the amount of stress and abstinence

Web Table 2 Estimated coefficients of control variables from the overall model (coefficient of birth weight is multiplied by 1000 and all other coefficients are multiplied by 10).

	Intercept	MBM	Female	Vaginal	PROM	SNAPPE-II	Birth weight
Habituation	62.372	12.694	2.394	3.317	-3.020	0.069	-0.517
Attention	59.368	-8.039	1.308	10.392	5.081	-0.032	-1.103
Handling	4.907	-0.139	-0.739	-1.508	-0.441	0.015	0.132
Qmovement	39.393	4.187	-3.711	3.535	2.411	0.011	-0.067
Regulation	59.774	0.448	-2.745	4.578	3.821	-0.256	-0.665
Nonoptref	34.916	8.465	-4.702	-7.855	-4.039	0.264	0.998
Stress	1.901	-0.032	-0.088	-0.265	0.029	0.015	-0.029
Arousal	33.183	-3.236	3.362	2.113	-5.717	-0.081	0.189
Excitability	3.087	-13.781	7.811	-3.680	-9.722	0.179	2.397
Lethargy	7.598	34.721	-0.171	-23.384	-0.771	0.233	2.549

signs observed in the neurodevelopmental examination procedure (Lester and Tronick, 2004), and a lower value indicates a better neurodevelopment situation. Based on the fitting results, female infants generally perform better in the neurodevelopment examination than male infants. Vaginal delivery and a higher percentage of feeding with mother's breast milk also bring benefit to the neurodevelopment of preterm infants. Moreover, the estimated coefficient of birth weight is -0.029 after multiplying by 1000, which indicates that infants with larger birth weights are more likely to have a better neurological development. SNAPPE-II is one kind of illness severity score, and a higher SNAPPE-II score is often observed among expired infants (Harsha and Archana, 2015). Thus, it is reasonable to observe that the SNAPPE-II score is positively related to the stress/abstinence score. As for the PROM, it is a major cause of premature birth and could be very dangerous to both mother and infant. The method provides a positive coefficient estimate of PROM, which matches well with the intuition that a pregnant who did not experience PROM is more likely to give birth to a healthier baby. The effects of control variables on other sub-scale scores of NNNS can be similarly interpreted based on the estimated coefficients.

Web Appendix H Additional Tables and Figures

Web Table 3 $d_1(\mathbf{P_k}(\mathbf{I_n} - \mathbf{Q_k}))$ values obtained in simulation studies. The results are displayed in terms of the mean and standard error (in parenthesis) across 300 replications. Each group is denoted as "G" followed by its group number.

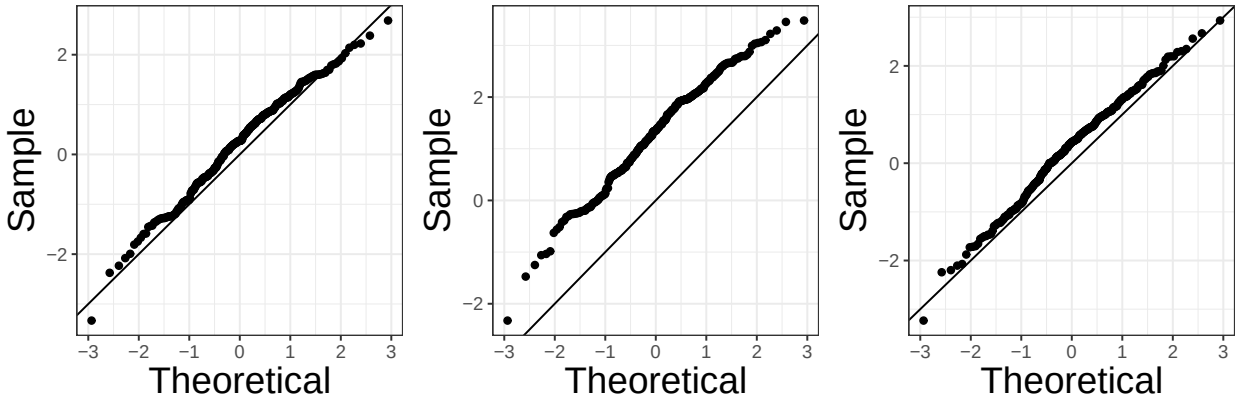
Section 4.1: Setting 1				
(Correlation type, ρ_x)	G1	G2	G3	
(within-group, 0.0)	0.355 (0.020)	0.353 (0.019)	0.354 (0.018)	
(within-group, 0.5)	0.333 (0.020)	0.332 (0.019)	0.333 (0.019)	
(among-group, 0.0)	0.355 (0.020)	0.353 (0.019)	0.354 (0.018)	
(among-group, 0.5)	0.532 (0.030)	0.554 (0.023)	0.552 (0.024)	
Section 4.1: Setting 2				
(Correlation type, ρ_x)	G1	G2	G3	
(within-group, 0.0)	0.783 (0.009)	0.782 (0.009)	0.783 (0.010)	
(within-group, 0.5)	0.743 (0.013)	0.742 (0.014)	0.742 (0.013)	
(among-group, 0.0)	0.783 (0.009)	0.782 (0.009)	0.783 (0.010)	
(among-group, 0.5)	0.754 (0.014)	0.762 (0.013)	0.762 (0.014)	
Section 4.2: Simulation with Generated Compositional Data				
(Correlation type, ρ_x)	G1	G2	G6	G7
(among-group, 0.0)	0.994 (0.001)	0.994 (0.001)	0.433 (0.025)	0.434 (0.026)
(among-group, 0.5)	0.995 (0.000)	0.996 (0.000)	0.507 (0.034)	0.507 (0.033)

Web Table 4 Simulation results with **X** being generated from the among-group correlation setup. 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.

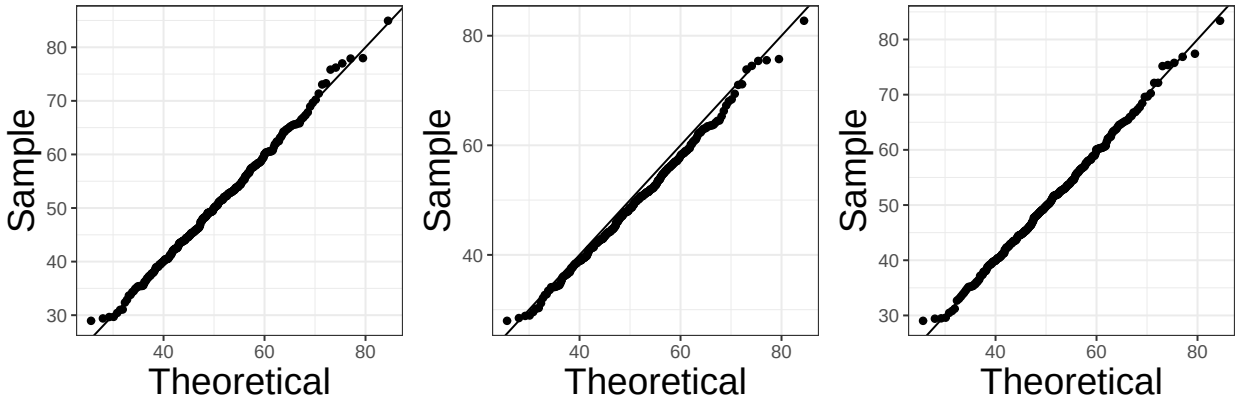
Design (SNR, ρ_x)	$\hat{\sigma}/\sigma - 1$	$\lvert\hat{\sigma}/\sigma - 1\rvert$	Multivariate			Univariate (Bonf)			Univariate (HMP)		
			G1	G2	G3	G1	G2	G3	G1	G2	G3
Setting 1											
(0.1,0.0)	0.34 (1.44)	1.17 (0.90)	0.65	0.05	0.05	0.51	0.03	0.03	0.54	0.03	0.03
(0.1,0.0)	0.34 (1.44)	1.17 (0.90)	0.61	0.05	0.05	0.50	0.04	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
(0.2,0.0)	1.81 (1.45)	1.97 (1.23)	1.00	0.04	0.04	1.00	0.04	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
(0.4,0.0)	1.11 (1.61)	1.53 (1.22)	1.00	0.03	0.05	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	0.07	0.11	0.04	0.05	0.11	0.04	0.04
(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)	0.85	0.03	0.05	0.65	0.03	0.04	0.71	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)	1.54 (1.71)	1.87 (1.33)	1.00	0.03	0.05	1.00	0.02	0.04	1.00	0.02	0.04

Web Table 5 Corrected p-values from the univariate analysis adjusted by using BH adjustment. For each stage, we control the FDR of the tests related to the orders identified in multivariate analysis based on the corrected p-values. The values highlighted with an asterisk are the significant ones by controlling the FDR under 10%.

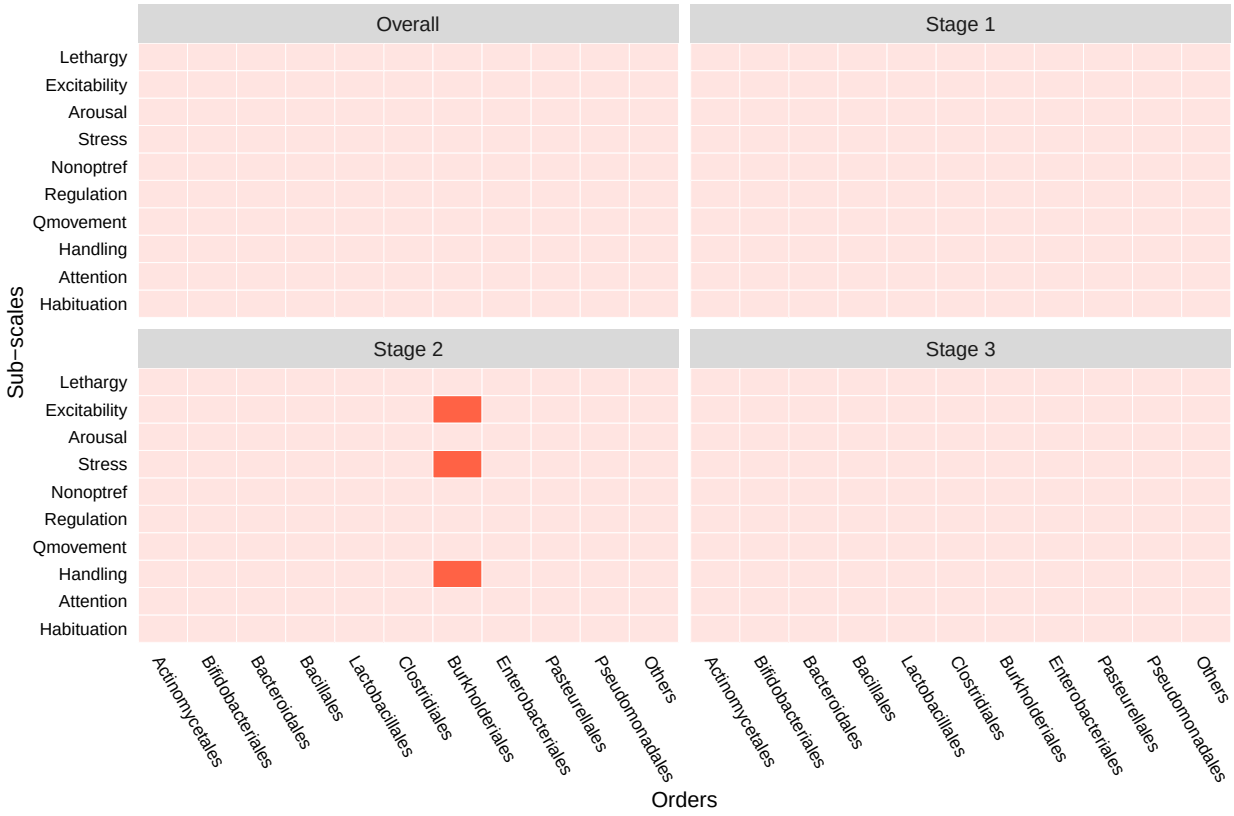
	Overall		Stage 2	Stage 3
	Clostridiales	Others	Burkholderiales	Actinomycetales
Habituation	0.334	0.334	0.214	0.136
Attention	0.231	0.231	0.247	0.168
Handling	0.471	0.531	0.002*	0.506
Qmovement	0.402	0.266	0.067*	0.451
Regulation	0.231	0.231	0.026*	0.074*
Nonoptref	0.591	0.280	0.710	0.935
Stress	0.221	0.471	0.002*	0.074*
Arousal	0.221	0.125	0.160	0.769
Excitability	0.231	0.257	0.000*	0.074*
Lethargy	0.401	0.221	0.710	0.684



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



Web Figure 2 Simulation results for setting 1 with $\rho_x = 0$ from 300 simulation runs: from left to right are the Q-Q plots of $\|\mathbf{P}_3\mathbf{E} - \mathrm{Rem}_3\|_F^2/\hat{\sigma}^2$ versus $\chi^2_{r'_3 q}$ with SNR $= 0.1$, 0.2, and 0.4, respectively.



Web Figure 3 The identified predictive orders for each sub-scale score of NNNS when control the FDR under 10% for each time-specific analysis. The selected orders are marked in red, while the remaining orders are marked in pink.

References

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli 19*(4), 1212–1242.

Cai, J.-F., E. J. Candès, and Z. Shen (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization 20*(4), 1956–1982.

Chatterjee, A. and S. N. Lahiri (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics 41*(3), 1232–1259.

Harsha, S. S. and B. R. Archana (2015). Snappe-ii (score for neonatal acute physiology with perinatal extension-ii) in predicting mortality and morbidity in nicu. *Journal of Clinical and Diagnostic Research: JCDR 9*(10), SC10.

He, B., H. Yang, and S. Wang (2000). Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and applications 106*(2), 337–356.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of statistics 28*(5), 1356–1378.

Lester, B. M. and E. Z. Tronick (2004). The neonatal intensive care unit network neurobehavioral scale procedures. *Pediatrics 113*(Supplement 2), 641–667.

Li, G., X. Liu, and K. Chen (2019). Integrative multi-view regression: Bridging group-sparse and low-rank models. *Biometrics 75*(2), 593–602.

Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association 104*(488), 1671–1681.

Mitra, R. and C. H. Zhang (2016). The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics 10*(2), 1829–1873.

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics 45*(1), 158–195.

Shi, C., R. Song, Z. Chen, and R. Li (2019). Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics 47*(5), 2671–2703.

Shi, P., A. Zhang, and H. Li (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics 10*(2), 1019–1040.

Sun, Z., W. Xu, X. Cong, G. Li, and K. Chen (2020). Log-contrast regression with functional compositional predictors: Linking preterm infants' gut microbiome trajectories to neurobehavioral outcome. *Annals of Applied Statistics 14*(3), 1535–1556.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B 58*, 267–288.

Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *The Annals of Statistics 37*(5A), 2178–2201.

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications 170*, 33–45.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B 68*(1), 49–67.

Zhang, C. H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B 67*(2), 301–320.