Supplementary Material for Debiased Group Lasso for Multiple Compositional Data

Sujin Lee[1] and Sungkyu Jung[1,2]

[1]Department of Statistics, Seoul National University
[2]Institute for Data Innovation in Science, Seoul National University

The proof of theoretical results and additional simulation studies are discussed in this supplementary material.

S1 Rationale for linear constraints in compositional regression

In this section, we clarify the motivation and justification for imposing the linear constraint $\sum_{i=1}^{p} \beta_i = 0$ as well as utilizing the log transformation. Simply put, the constraint is imposed for model interpretability. For reference, we rewrite the definitions of three common modeling choices, listed in the main article:

- Modeling log-contrasts (with a reference component x_p) (Aitchison, 1982, Aitchison and Bacon-Shone, 1984, Bates and Tibshirani, 2019, Sun et al., 2020):

$$y = \alpha + \sum_{i=1}^{p-1} \beta_i \log(x_i/x_p) + \epsilon, \tag{S1}$$

  where $\mathbf{x} = (x_1, \ldots, x_p)$ is the compositional variable ($\sum_{i=1}^{p} x_i = 1$).

- Modeling clr-transformed compositions (Susin et al., 2020) :

$$y = \alpha + \sum_{i=1}^{p} \beta_i [\log(x_i) - p^{-1} \sum_{j=1}^{p} \log(x_j)] + \epsilon, \tag{S2}$$

  where $\mathsf{clr}(\mathbf{x}) = (\log(x_i) - p^{-1} \sum_{j=1}^{p} \log(x_j))_{i=1,\ldots,p}$ is called the centered log-ratio (clr) transformation of $\mathbf{x}$.

- Modeling log-transformed compositions (Greenacre, 2019, Lee et al., 2023, Lin et al., 2014, Lu et al., 2019, Shi et al., 2016, Wang and Zhao, 2017) :

$$y = \alpha + \sum_{i=1}^{p} \beta_i \log(x_i) + \epsilon. \tag{S3}$$

Other choices involving the ilr- or alr-transformations and utilizing the raw compositions have also been discussed in the literature (Egozcue et al., 2003, Li et al., 2023).

The log-contrast model (S1) is equivalent to log-transformed model (S3) if and only if the linear constraint $\sum_{i=1}^{p} \beta_i = 0$ is imposed on (S3). Under the same constraint imposed on both (S2) and (S3), the clr model is equivalent to the log-transformed model. Furthermore, the clr transformation maps a p-part composition $\mathbf{x}$ to a p-dimensional real vector with zero-sum constraint, meaning the transformed vector $\mathsf{clr}(\mathbf{x})$ lies in a $(p-1)$-dimensional hyperplane, necessitating a linear constraint on the coefficient vector to ensure identifiability. See Section S1.1.

When the linear constraint is imposed, these three modeling choices are equivalent. We choose to work with the log-transformed model (S3) with the constraints, as this practice has become standard in compositional regression modeling with regularization (e.g., Lin et al., 2014). These (equivalent) models ensure subcompositional coherence and scale invariance, key principles in compositional data analysis (Bates and Tibshirani, 2019, Greenacre et al., 2021). This is further discussed in Section S1.2.

S1.1 Equivalence of models (S1), (S2) and (S3) under the linear constraint

Equivalence of (S1) and (S3): If (and only if) the linear constraint $\sum_{i=1}^{p} \beta_i = 0$ is imposed to (S3), then one can rewrite (S3) in the form of the log-contrast model (S1). It in fact provides a much greater flexibility in the choice of the reference variable. For example, if x_1 was the reference variable, the log-transformed regression model (S3) writes to

$$\begin{aligned} y &= \alpha + (-\sum_{i=2}^{p} \beta_i) \log(x_1) + \sum_{i=2}^{p} \beta_i \log(x_i) + \epsilon, \\ &= \alpha + \sum_{i=2}^{p} \beta_i \log(x_i/x_1) + \epsilon. \end{aligned}$$

Equivalence of (S2) and (S3): The clr model (S2) can be written as

$$\begin{aligned} y &= \alpha + \sum_{i=1}^{p} \beta_i [\log(x_i) - p^{-1} \sum_{j=1}^{p} \log(x_j)] + \epsilon \\ &= \alpha + \sum_{i=1}^{p} \beta_i \log(x_i) - (\sum_{i=1}^{p} \beta_i) p^{-1} \sum_{j=1}^{p} \log(x_j) + \epsilon, \end{aligned}$$

from which one can see that the constraint $\sum_{i=1}^{p} \beta_i = 0$ makes the model equivalent to (S3) with the same constraint. Furthermore, the clr-transformed composition $\mathbf{x}$ is

$$\begin{aligned}\mathsf{clr}(\mathbf{x}) &= \left(\log(x_i) - p^{-1} \sum_{i=1}^{p} \log(x_i) \right)_{i=1,\ldots,p} \\ &= \log(\mathbf{z}) - p^{-1}\mathbf{1}\mathbf{1}^T \log(\mathbf{x}) \\ &= (\mathbf{I}_p - \mathbf{P}_1) \log(\mathbf{x}).\end{aligned}$$

Thus, the clr-transformed covariates are linearly dependent.

S1.2 Subcompositional coherence and scale invariance

By using either of the equivlaent models listed earlier, the statistician chose to assume that the effect of variables to the response is linear with respect to the log scale, and its effect is relative; that is, the effect of x_1 (with respect to x_p) on y should be the same for both $(x_1, x_p) = (0.2, 0.1)$ and $(x_1, x_p) = (0.6, 0.3)$. This model thus captures *relative abundances* with respect to a reference component and is *scale-invariant*, thereby facilitating interpretation.

The requirement of *subcompositional coherence* refers to the following: any analysis of subcompositional data should have the property that the analysis is unchanged whether we are analyzing the entire data set or a subcomposition (see, e.g., Aitchison, 1982, 1986, Bates and Tibshirani, 2019).

We provide an example where the model (S3) with the linear constraint ensures the subcompositional coherence. Consider a compositional regression model with raw variables $\mathbf{w} = (w_1, \ldots, w_p)$. If these raw variables do not constitute a composition, the closure operation is applied:

$$\mathbf{x} = (x_1, \ldots, x_p) = \left(\frac{w_1}{\sum_{j=1}^{p} w_j}, \ldots, \frac{w_p}{\sum_{j=1}^{p} w_j} \right) \in \mathbb{S}^{p-1}.$$

If it is known that the first three components are related to the response, then one takes a *subcomposition*, consisting of the first three component:

$$\begin{aligned}\mathbf{x}' &= (x_1', x_2', x_3') \\ &= \left(\frac{x_1}{\sum_{j=1}^{3} x_j}, \frac{x_2}{\sum_{j=1}^{3} x_j}, \frac{x_3}{\sum_{j=1}^{3} x_j} \right) \\ &= \left(\frac{w_1}{\sum_{j=1}^{3} w_j}, \frac{w_2}{\sum_{j=1}^{3} w_j}, \frac{w_3}{\sum_{j=1}^{3} x_j} \right) \in \mathbb{S}^{3-1}.\end{aligned}$$

The subcompositional coherence dictates that, if the first three components are truly related to the response, then the model with all p components $\mathbf{x}$ should be equivalent to the model with the subcomposition $\mathbf{x}'$.

The model with all p components $\mathbf{x}$ is

$$\begin{aligned} y &= \alpha + \sum_{i=1}^{3} \beta_i \log(x_i) + \epsilon \\ &= \alpha + \sum_{i=1}^{3} \beta_i \log\left(w_i / (\sum_{j=1}^{p} w_j) \right) + \epsilon \\ &= \alpha + \sum_{i=1}^{3} \beta_i \log(w_i) - (\sum_{i=1}^{3} \beta_i) \log\left(\sum_{j=1}^{p} w_j \right) + \epsilon \end{aligned} \tag{S4}$$

where the coefficient β_i of $\log(x_i)$ for $i = 4, \ldots, p$ is zero. The model with the subcomposition $\mathbf{x}'$ is

$$\begin{aligned} y &= \alpha + \sum_{i=1}^{3} \beta_i \log(x'_i) + \epsilon \\ &= \alpha + \sum_{i=1}^{3} \beta_i \log(w_i) - (\sum_{i=1}^{3} \beta_i) \log(\sum_{j=1}^{3} w_j) + \epsilon. \end{aligned} \tag{S5}$$

Comparing (S4) and (S5), we conclude that the linear constraint $\sum_{i=1}^{p} \beta_i = 0$ is necessary to ensure that the two models are equivalent.

S2 Proofs and Technical Details

Throughout, we use the following notation. For a natural number p, $[p] := \{1, \ldots, p\}$. For a vector $\mathbf{u} \in \mathbb{R}^d$, the ℓ_p norm is denoted by $\|\mathbf{u}\|_p = (\sum_{k=1}^{d} |u_k|^p)^{1/p}$, and $\|\mathbf{u}\|_\infty = \max_{1 \le k \le d} |u_k|$, and $\|\mathbf{u}\|_0 = \#\{j : u_j \neq 0\}$. For a matrix $\mathbf{A}_{m \times n}$, $\|\mathbf{A}\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{p} |a_{ij}|$ and define $|\mathbf{A}|_\infty = \max_{i,j} |a_{ij}|$ as element-wise maximum. In addition, for every symmetric semi-definite matrix $\mathbf{A}$, $\|\mathbf{A}\|_F$, and $\|\mathbf{A}\|$ denote the Frobenius and spectral norms of $\mathbf{A}$, respectively. Note that if $\rho_1, \ldots, \rho_k$ are the eigenvalues of $\mathbf{A}$, we have that $\text{trace}(\mathbf{A}) = \sum_{i=1}^{k} \rho_i$, $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{k} \rho_i^2}$ and $\|\mathbf{A}\| = \max_{i=1,\ldots,k} \rho_i$.

If $J \in [M]$, we define $\boldsymbol{\beta}_J$ as the vector obtained by selecting $\boldsymbol{\beta}_{G_j}$ for $j \in J$. Additionally, we defined $J(\boldsymbol{\beta}) = \{j : \boldsymbol{\beta}_{G_j} \neq \mathbf{0}, j \in [M]\}$ and $G(\boldsymbol{\beta}) = |J(\boldsymbol{\beta})|$ where $|J|$ denotes the cardinality of set $J \in [M]$. The set $J(\boldsymbol{\beta})$ contains the indices of the relevant groups and $G(\boldsymbol{\beta})$ denotes the number of such groups. For any $\boldsymbol{\beta} \in \mathbb{R}^p$, we define the the mixed $(2, p)$-norm of $\boldsymbol{\beta}$ for every $p \ge 1$ as

$$\|\boldsymbol{\beta}\|_{2,p} = \left(\sum_{j=1}^{M} \left(\sum_{k \in G_j} \beta_k^2 \right)^{p/2} \right)^{1/p} = \left(\sum_{i=1}^{M} \|\boldsymbol{\beta}_{G_j}\|_2^p \right)^{1/p},$$

and the $(2, \infty)$-norm of $\boldsymbol{\beta}$ as $\|\boldsymbol{\beta}\|_{2,\infty} = \max_{1 \le j \le M} \|\boldsymbol{\beta}_{G_j}\|_2$.

S2.1 On the validity of the constrained RE condition

This section provides additional discussion on the validity and behavior of the restricted eigenvalue (RE) condition when a linear constraint of the form $\mathbf{C}^T\boldsymbol{\beta} = \mathbf{0}$ is imposed, as in Equation (7) of the main article.

Recall that the RE(s) condition is said to hold with constant $\kappa > 0$ if the following inequality is satisfied:

$$\min_{\boldsymbol{\beta}} \left\{ \frac{\|\widetilde{\mathbf{Z}}\boldsymbol{\beta}\|_2}{\sqrt{n}\|\boldsymbol{\beta}_T\|_2} : |T| \leq s,\ \boldsymbol{\beta} \in \mathbb{R}^p \backslash \{0\},\ \boldsymbol{\beta} \in \mathcal{C}_0^{(G)}(T) \right\} \geq \kappa, \tag{S6}$$

where $\mathcal{C}^{(G)}(T)$ is a cone defined by group-sparsity constraints. In the unconstrained setting, this cone includes all vectors $\boldsymbol{\beta}$ satisfying a compatibility condition between components inside and outside T. When the linear constraint $\mathbf{C}^T\boldsymbol{\beta} = \mathbf{0}$ is introduced, the feasible domain is restricted to the intersection $\mathcal{C}^{(G)}(T) = \mathcal{C}_0^{(G)}(T) \cap \text{null}(\mathbf{C}^T)$. Although this intersection may no longer form a cone, the RE condition can still be meaningfully defined over this restricted domain.

Importantly, imposing the constraint does not weaken the RE condition. Since the feasible set under the constraint is a subset of the unconstrained domain, the minimum in (10) is taken over a smaller set. As such, if the RE(s) condition holds in the unconstrained case with constant κ, it necessarily holds in the constrained setting with the same or larger constant. This implies that the presence of the constraint does not invalidate the RE condition and may even strengthen it, depending on the geometry of the feasible set.

There are certain degenerate cases in which the constrained RE condition becomes ill-defined. For instance, if the active group set T is too small, e.g., $|T| = 1$, the set $\mathcal{C}^{(G)}(T)$ may be empty. In such cases, the left-hand side of (10) (with $\mathcal{C}_0^{(G)}(T)$ replaced by $\mathcal{C}^{(G)}(T)$) is undefined. To avoid these pathologies, we assume that such trivial cases are excluded.

While the constrained restricted eigenvalue (RE) condition plays a central role in our theoretical analysis, we note that directly verifying this condition for a given design matrix is generally infeasible. This difficulty arises from the fact that the RE condition involves a nonconvex optimization over coefficient vectors that lie in a sparsity-constrained cone intersected with a linear subspace induced by the constraint $\mathbf{C}^T\boldsymbol{\beta} = \mathbf{0}$. The resulting feasible set has a complex geometry that does not admit closed-form characterizations or tractable verification procedures.

Accordingly, it is common practice in high-dimensional analysis to assume the RE condition under general structural assumptions on the design matrix, without empirical verification. This approach enables rigorous theoretical guarantees while maintaining flexibility in modeling and analysis. Our theoretical results follow this standard convention.

S2.2 Proof of Theorem 1

We begin with intermediate lemmas.

Lemma S1. *(Cavalier et al., 2002) Let $\xi_1, \ldots, \xi_n$ be i.i.d. $N(0,1)$, $v = (v_1, \ldots, v_n) \neq 0, \eta_v = \frac{1}{\sqrt{2}\|v\|} \sum_{i=1}^n (\xi_i^2 - 1) v_i$ and $m(v) = \frac{\|v\|_\infty}{\|v\|}$. We have, for all $x > 0$, that*

$$\mathbb{P}(|\eta_v| > x) \leq 2 \exp\left(-\frac{x^2}{2(1 + \sqrt{2} x m(v))} \right).$$

Recall that the $p \times r$ constraint matrix $\mathbf{C}$ is defined as

$$\mathbf{C} = \begin{pmatrix} \mathbf{1}_{p_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{p_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{1}_{p_r} \end{pmatrix}, \tag{S7}$$

and $\mathbf{P}_\mathbf{C} = \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T = \mathrm{diag}(\mathbf{J}_{p_1}/p_1, \ldots, \mathbf{J}_{p_r}/p_r)$. Then, for any p, we have

$$\|\mathbf{I}_p - \mathbf{P}_\mathbf{C}\|_\infty \leq 1, \tag{S8}$$

for all p, and the diagonal elements of $\mathbf{I}_p - \mathbf{P}_\mathbf{C}$ are greater than zero, as long as p_i does not diverge. Next result by Shi et al. (2016) generalizes this result.

Lemma S2. *(Shi et al., 2016) For any matrix $\mathbf{A}$, we have*

$$\|(\mathbf{I}_p - \mathbf{P}_\mathbf{C})\mathbf{A}\|_\infty \leq |\mathbf{A}|_\infty.$$

See Shi et al. (2016) for a proof for Lemma S2.

The following lemma plays an important building block in verifying Theorem 1. Recall that the optimization problem (7) is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\mathrm{argmin}} \left(\frac{1}{2n} \|\mathbf{y} - \widetilde{\mathbf{Z}}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{2,1} \right) \quad \text{subject to } \mathbf{C}^T\boldsymbol{\beta} = \mathbf{0}.$$

Lemma S3. *Let $\boldsymbol{\beta}^*$ be the true regression coefficient vector for the model (6) with $\mathbf{C}$ given as (S7). Assume that $M \geq 2, n \geq 1$. Let $\widehat{\boldsymbol{\Sigma}}_j = \widetilde{\mathbf{Z}}_{G_j}^T \widetilde{\mathbf{Z}}_{G_j}/n$ $(j = 1, \ldots, M)$, and $\widehat{\boldsymbol{\Sigma}} = \widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}/n$. Let $q > 0$ be any positive number. If the tuning parameter λ in the constrained optimization problem (7) is chosen so that*

$$\lambda \geq \max_{j \in [M]} \left\{ \frac{2\sigma}{\sqrt{n}} \sqrt{trace(\widehat{\boldsymbol{\Sigma}}_j) + 2\|\widehat{\boldsymbol{\Sigma}}_j\|(2q \log M + \sqrt{m_j q \log M})} \right\}, \tag{S9}$$

Then with probability at least $1 - 2M^{1-q}$, the following holds:

For any solution $\hat{\boldsymbol{\beta}}$ of (7) and for all $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfying $\mathbf{C}^T\boldsymbol{\beta} = \mathbf{0}$, we have, for all $j \in [M]$,

$$\begin{aligned} &\frac{1}{n}\|\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \lambda \sum_{j=1}^M \|\hat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}_{G_j}\|_2 \qquad \text{(S10)} \\ &\quad \leq \frac{1}{n}\|\widetilde{\mathbf{Z}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 + 4\lambda \sum_{j \in J(\boldsymbol{\beta})} \min(\|\boldsymbol{\beta}_{G_j}\|_2, \|\hat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}_{G_j}\|_2), \\ &\frac{1}{n}\|\widetilde{\mathbf{Z}}_{G_j}^T \widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \leq \lambda(k_0\sqrt{m_j M} + 1/2). \qquad \text{(S11)} \end{aligned}$$

where $m_j = |G_j|$.

Proof of Lemma S3. We largely follow the argument used in Lounici et al. (2011). For all $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfying $\mathbf{C}^T\boldsymbol{\beta} = \mathbf{0}$, we have

$$\frac{1}{n}\|\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}} - \mathbf{y}\|_2^2 + 2\lambda\sum_{j=1}^{M}\|\hat{\boldsymbol{\beta}}_{G_j}\|_2 \leq \frac{1}{n}\|\widetilde{\mathbf{Z}}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + 2\lambda\sum_{j=1}^{M}\|\boldsymbol{\beta}_{G_j}\|_2$$

and, using (6), we have

$$\frac{1}{n}\|\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \leq \frac{1}{n}\|\widetilde{\mathbf{Z}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 + \frac{2}{n}\boldsymbol{\epsilon}^T\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 2\lambda\sum_{j=1}^{M}(\|\boldsymbol{\beta}_{G_j}\|_2 - \|\hat{\boldsymbol{\beta}}_{G_j}\|_2).$$

By the Cauchy-Schwarz inequality, the second term of the right-hand side of the above inequality satisfies

$$\frac{1}{n}\boldsymbol{\epsilon}^T\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \frac{1}{n}\sum_{j=1}^{M}\|(\boldsymbol{\epsilon}^T\widetilde{\mathbf{Z}})_{G_j}\|_2\|\hat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}_{G_j}\|_2.$$

Consider the random event $\mathcal{A} = \cap_{j=1}^{M}\mathcal{A}_j$, where $\mathcal{A}_j = \{\frac{1}{n}\|\boldsymbol{\epsilon}_{G_j}^T\widetilde{\mathbf{Z}}\|_2 \leq \frac{\lambda}{2}\}$. We note that

$$\mathbb{P}(\mathcal{A}_j) = \mathbb{P}\left(\left\{\frac{1}{n^2}\boldsymbol{\epsilon}^T\widetilde{\mathbf{Z}}_{G_j}\widetilde{\mathbf{Z}}_{G_j}^T\boldsymbol{\epsilon} \leq \frac{\lambda^2}{4}\right\}\right) = \mathbb{P}\left(\left\{\frac{\sum_{i=1}^{n} v_{j,i}(\xi_i^2 - 1)}{\sqrt{2}\|v_j\|} \leq x_j\right\}\right),$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. $N(0,1)$, $v_{j,1}, \ldots, v_{j,n}$ denote the eigenvalues of the matrix $\widetilde{\mathbf{Z}}_{G_j}\widetilde{\mathbf{Z}}_{G_j}^T/n$, among which the positive ones are the same as those of $\widehat{\boldsymbol{\Sigma}}_j$, and x_j is defined as

$$x_j := \frac{\lambda^2 n/(4\sigma^2) - \mathrm{trace}(\widehat{\boldsymbol{\Sigma}}_j)}{\sqrt{2}\|\widehat{\boldsymbol{\Sigma}}_j\|_F}.$$

We apply Lemma S1 to upper bound the probability of the complement of the event $\mathcal{A}_j$. Specifically, we choose $v = v_j = (v_{j,1}, \ldots, v_{j,n}), x = x_j$ and $m(v) = \|\widehat{\boldsymbol{\Sigma}}_j\|/\|\widehat{\boldsymbol{\Sigma}}_j\|_F$ and conclude from Lemma S1 that

$$\mathbb{P}(A_j^c) \leq 2\exp\left(-\frac{x_j^2}{2(1 + \sqrt{2}x_j\|\widehat{\boldsymbol{\Sigma}}_j\|/\|\widehat{\boldsymbol{\Sigma}}_j\|_F)}\right).$$

We now choose the tuning parameter λ (which also affects x_j) so that the right-hand side of the above inequality is smaller than $2M^{-q}$ that yields

$$x_j \geq \sqrt{2}A + \sqrt{2A^2 + 2q\log M} \tag{S12}$$

where $A = \|\widehat{\boldsymbol{\Sigma}}_j\|/\|\widehat{\boldsymbol{\Sigma}}_j\|_F q\log M$. Using the subadditivity property of the square root and the inequality $\|\widehat{\boldsymbol{\Sigma}}_j\|_F \leq \sqrt{m_j}\|\widehat{\boldsymbol{\Sigma}}_j\|$, it is verified that for λ satisfying the condition (S9) satisfies (S12).

With λ chosen to satisfy (S9), we have $\mathbb{P}(A_j^c) \leq 2M^{-q}$ for all $j \in [M]$. A simple application of the union bound gives that $\mathbb{P}(\mathcal{A}^c) \leq 2M^{1-q}$.

With probability at least $1-2M^{1-q}$, we have the following:

$$\begin{aligned}
&\frac{1}{n}\|\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2+\lambda\sum_{j=1}^{M}\|\hat{\boldsymbol{\beta}}_{G_j}-\boldsymbol{\beta}_{G_j}\|_2\\
&\quad\le\frac{1}{n}\|\widetilde{\mathbf{Z}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2^2+2\lambda\sum_{j=1}^{M}(\|\hat{\boldsymbol{\beta}}_{G_j}-\boldsymbol{\beta}_{G_j}\|_2+\|\boldsymbol{\beta}_{G_j}\|_2-\|\hat{\boldsymbol{\beta}}_{G_j}\|_2)\\
&\quad\le\frac{1}{n}\|\widetilde{\mathbf{Z}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2^2+4\lambda\sum_{j\in J(\boldsymbol{\beta})}\min(\|\boldsymbol{\beta}_{G_j}\|_2,\|\hat{\boldsymbol{\beta}}_{G_j}-\boldsymbol{\beta}_{G_j}\|_2).
\end{aligned}$$

Then by the KKT condition of optimization problem (7), for every $j\in[M]$,

$$\frac{1}{n}\|\widetilde{\mathbf{Z}}_{G_j}^T(\mathbf{y}-\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}})+(\mathbf{C}\boldsymbol{\eta})_{G_j}\|_2\le\lambda,$$

for some $\boldsymbol{\eta}\in\mathbb{R}^r$. Then, we have

$$\|\widetilde{\mathbf{Z}}^T(\mathbf{y}-\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}})+\mathbf{C}\boldsymbol{\eta}\|_2^2=\sum_{j=1}^{M}\|\widetilde{\mathbf{Z}}_{G_j}^T(\mathbf{y}-\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}})+(\mathbf{C}\boldsymbol{\eta})_{G_j}\|_2^2\le Mn^2\lambda^2.$$

implying $\|\widetilde{\mathbf{Z}}^T(\mathbf{y}-\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}})+\mathbf{C}\boldsymbol{\eta}\|_\infty\le n\lambda\sqrt{M}$. Using the fact that $\widetilde{\mathbf{Z}}(\mathbf{I}-\mathbf{P}_\mathbf{C})=\widetilde{\mathbf{Z}}$, $\mathbf{C}^T\mathbf{C}=\mathbf{I}_r$, and Lemma S2, we have

$$\begin{aligned}
\|\widetilde{\mathbf{Z}}^T(\mathbf{y}-\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}})\|_\infty&=\|(\mathbf{I}-\mathbf{P}_\mathbf{C})(\widetilde{\mathbf{Z}}^T(\mathbf{y}-\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}})+\mathbf{C}\boldsymbol{\eta})\|_\infty\\
&\le\|\widetilde{\mathbf{Z}}^T(\mathbf{y}-\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}})+\mathbf{C}\boldsymbol{\eta}\|_\infty\le n\lambda\sqrt{M}.
\end{aligned}$$

Moreover, on the event $\mathcal{A}$, using (6) and triangle inequality, we have

$$\begin{aligned}
\frac{1}{n}\|\widetilde{\mathbf{Z}}_{G_j}^T\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2&\le\frac{1}{n}\|\widetilde{\mathbf{Z}}_{G_j}^T(\widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}-\mathbf{y})\|_2+\frac{1}{n}\|\widetilde{\mathbf{Z}}_{G_j}^T\boldsymbol{\epsilon}\|_2\\
&\le\sqrt{m_j}\lambda\sqrt{M}+\lambda/2=\lambda(\sqrt{m_jM}+1/2).
\end{aligned}$$

□

We are now ready to give a proof of Theorem 1.

Proof of Theorem 1. Let $T=\{j:\boldsymbol{\beta}_{G_j}\neq\mathbf{0}\}$ and let $\mathbf{h}=\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*$. By inequality (S10), on the event $\mathcal{A}=\cap_{j=1}^{M}\{\frac{1}{n}\|\boldsymbol{\epsilon}_{G_j}^T\widetilde{\mathbf{Z}}\|_2\le\frac{\lambda}{2}\}$, we have

$$\frac{1}{n}\|\widetilde{\mathbf{Z}}\mathbf{h}\|_2^2\le4\lambda\sum_{j\in T}\|\mathbf{h}_{G_j}\|_2\le4\lambda\sqrt{s}\|\mathbf{h}_T\|_2. \tag{S13}$$

Furthermore, by the same inequality, on the event $\mathcal{A}$, we have that $\sum_{j=1}^{M}\|\mathbf{h}_{G_j}\|_2\le4\sum_{j\in T}\|\mathbf{h}_{G_j}\|_2$, implying $\sum_{j\in T^c}\|\mathbf{h}_{G_j}\|_2\le3\sum_{j\in T}\|\mathbf{h}_{G_j}\|_2$. Since the RE($s$) condition is satisfied with κ_1, we have

$$\|\mathbf{h}_T\|_2\le\frac{\|\widetilde{\mathbf{Z}}\mathbf{h}\|_2}{\kappa_1\sqrt{n}}. \tag{S14}$$

Then, using (S13) and (S14) we have the following:

$$\left(\frac{1}{\sqrt{n}}\|\widetilde{\mathbf{Z}}\mathbf{h}\|_2\right)^2 \leq \left(\frac{4}{\kappa_1}\lambda\sqrt{s}\right)^2,$$

which proves (12). Moreover, inequality (13) follows by noting that

$$\|\mathbf{h}\|_{2,1} = \sum_{j=1}^{M}\|\mathbf{h}_{G_j}\|_2 \leq 4\sum_{j\in T}\|\mathbf{h}_{G_j}\|_2 \leq 4\sqrt{s}\|\mathbf{h}_T\|_2 \leq 4\sqrt{s}\frac{\|\widetilde{\mathbf{Z}}\mathbf{h}\|_2}{\kappa_1\sqrt{n}} \leq \frac{16\lambda s}{\kappa_1^2}.$$

Let T' be the set of indices in T^c corresponding to the s largest values of $\|\mathbf{h}_{G_j}\|$. Consider the set $T_{2s} = T \cup T'$. Note that $|T_{2s}| \leq 2s$. Let $j(k)$ be the index of the kth largest element of the set $\{\|\mathbf{h}_{G_j}\| : j \in T^c\}$. Then, we have

$$\|\mathbf{h}_{G_{j(k)}}\|_2 \leq \sum_{j\in T^c}\|\mathbf{h}_{G_j}\|_2/k.$$

This inequality, along with the fact that $\sum_{j\in T^c}\|\mathbf{h}_{G_j}\|_2 \leq 3\sum_{j\in T}\|\mathbf{h}_{G_j}\|_2$ on the event $\mathcal{A}$, implies the following:

$$\begin{aligned}\sum_{j\in T_{2s}^c}\|\mathbf{h}_{G_j}\|_2^2 &\leq \sum_{k=s+1}^{\infty}\frac{(\sum_{\ell\in T^c}\|\mathbf{h}_{G_\ell}\|_2)^2}{k^2}\\ &\leq \frac{(\sum_{\ell\in T^c}\|\mathbf{h}_{G_\ell}\|_2)^2}{s} \leq \frac{9(\sum_{\ell\in T}\|\mathbf{h}_{G_\ell}\|_2)^2}{s}\\ &\leq \frac{9s\|\mathbf{h}_T\|_2^2}{s} \leq 9\|\mathbf{h}_{T_{2s}}\|_2^2.\end{aligned}$$

Therefore,

$$\|\mathbf{h}_{T_{2s}^c}\|_2^2 \leq 9\|\mathbf{h}_{T_{2s}}\|_2^2$$

which leads to the following inequality

$$\|\mathbf{h}\|_2^2 \leq \|\mathbf{h}_{T_{2s}}\|_2^2 + \|\mathbf{h}_{T_{2s}^c}\|_2^2 \leq 10\|\mathbf{h}_{T_{2s}}\|_2^2. \tag{S15}$$

Next, from (S13), we observe that

$$\frac{1}{n}\|\widetilde{\mathbf{Z}}\mathbf{h}\|_2^2 \leq 4\lambda\sqrt{s}\|\mathbf{h}_{T_{2s}}\|_2. \tag{S16}$$

Additionally, $\sum_{j\in T^c}\|\mathbf{h}_{G_j}\|_2 \leq 3\sum_{j\in T}\|\mathbf{h}_{G_j}\|_2$ directly implies

$$\sum_{j\in T_{2s}^c}\|\mathbf{h}_{G_j}\|_2 \leq 3\sum_{j\in T_{2s}}\|\mathbf{h}_{G_j}\|_2.$$

Combining RE($2s$) condition is satisfied with κ_2 with inequality (S16), on the event $\mathcal{A}$, we have

$$\|\mathbf{h}_{T_{2s}}\|_2 \leq \frac{4\lambda\sqrt{s}}{\kappa_2}.$$

By putting all these results together, we obtain the following ℓ_2 estimation bound for group compositional lasso as follows:

$$\|\mathbf{h}\|_2 \leq \frac{4\lambda\sqrt{10s}}{\kappa_2},$$

which completes the proof. □

S2.3 Poof of Lemma 1

Proof. Write the ℓth row of $\widetilde{\mathbf{Z}}$ by $\widetilde{\mathbf{Z}}_\ell^T$, so that $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{\ell=1}^n \widetilde{\mathbf{Z}}_\ell \widetilde{\mathbf{Z}}_\ell^T$. Since $\mathbf{\Sigma}^{1/2}\Theta^{1/2}\widetilde{\mathbf{Z}}_\ell = (\mathbf{I}_p - \mathbf{P_C})\widetilde{\mathbf{Z}}_\ell = \widetilde{\mathbf{Z}}_\ell$, we have

$$\begin{aligned}\Theta\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - \mathbf{P_C}) &= \frac{1}{n}\sum_{\ell=1}^n \{\Theta\widetilde{\mathbf{Z}}_\ell\widetilde{\mathbf{Z}}_\ell^T - (\mathbf{I}_p - \mathbf{P_C})\} \\ &= \frac{1}{n}\sum_{\ell=1}^n \{\Theta^{1/2}\Theta^{1/2}\widetilde{\mathbf{Z}}_\ell\widetilde{\mathbf{Z}}_\ell^T\Theta^{1/2}\mathbf{\Sigma}^{1/2} - (\mathbf{I}_p - \mathbf{P_C})\}.\end{aligned}$$

Define $v_\ell^{(ij)} = \Theta_{i,\cdot}^{1/2}\Theta^{1/2}\widetilde{\mathbf{Z}}_\ell\widetilde{\mathbf{Z}}_\ell^T\Theta^{1/2}\mathbf{\Sigma}_{\cdot,j}^{1/2} - (\mathbf{I}_p - \mathbf{P_C})_{i,j}$, so that

$$\left(\Theta\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - \mathbf{P_C})\right)_{i,j} = \frac{1}{n}\sum_{\ell=1}^n v_\ell^{(ij)}.$$

Since $\mathbb{E}\Theta\widetilde{\mathbf{Z}}_\ell\widetilde{\mathbf{Z}}_\ell^T = \Theta\mathbf{\Sigma} = (\mathbf{I}_p - \mathbf{P_C})$, we have $\mathbb{E}v_\ell^{(ij)} = 0$, for any $i, j = 1, \ldots, p$ and for any $\ell = 1, \ldots, n$.

Let $\|X\|_{\psi_1}$ be the sub-exponential norm of a random variable X, which is defined as

$$\|X\|_{\psi_1} = \sup_{q\geq 1} q^{-1}(\mathbb{E}|X|^q)^{1/q}.$$

It is shown on p. 2897 of Javanmard and Montanari (2014) that for two random variables X and Y, $\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}$. This fact and Remark 5.18 of Vershynin (2012) yield that

$$\begin{aligned}\|v_\ell^{(ij)}\|_{\psi_1} &\leq 2\|\Theta_{i,\cdot}^{1/2}\Theta^{1/2}\widetilde{\mathbf{Z}}_\ell\widetilde{\mathbf{Z}}_\ell^T\Theta^{1/2}\mathbf{\Sigma}_{\cdot,j}^{1/2}\|_{\psi_1} \\ &\leq 2\|\Theta_{i,\cdot}^{1/2}\Theta^{1/2}\widetilde{\mathbf{Z}}_\ell\|_{\psi_2}\|\mathbf{\Sigma}_{\cdot,j}^{1/2}\Theta^{1/2}\widetilde{\mathbf{Z}}_\ell\|_{\psi_2} \\ &\leq 2\|\Theta_{i,\cdot}^{1/2}\|_2\|\mathbf{\Sigma}_{\cdot,j}^{1/2}\|_2\|\Theta^{1/2}\widetilde{\mathbf{Z}}_\ell\|_{\psi_2}\|\Theta^{1/2}\widetilde{\mathbf{Z}}_\ell\|_{\psi_2} \\ &\leq 2\sqrt{\sigma_{\max}(\mathbf{\Sigma})\sigma_{\max}(\Theta)}\phi^2 \\ &\leq 2\sqrt{C_{\max}/C_{\min}}\phi^2 \equiv \phi',\end{aligned}$$

Let $\mathbf{a}_i = (\mathbf{I}_p - \mathbf{P_C})\mathbf{e}_i$. Then, for any $i \in [p], j \in [M]$, we use the sub-Gaussian quadratic form concentration inequality (Cai et al., 2022, Lemma 6) to obtain

$$\mathbb{P}\left(\|(\mathbf{a}_i)_{G_j} - \frac{1}{n}\widetilde{\mathbf{Z}}_{G_j}^T\widetilde{\mathbf{Z}}\Theta\mathbf{e}_i\|_2 \geq \gamma\right) \leq 2\exp\left\{Cm - \frac{n}{6}\min\left((\frac{\gamma}{e\phi'})^2, \frac{\gamma}{e\phi'}\right)\right\}, \tag{S17}$$

where $m = \max_{j\in[M]} m_j$. Take $\gamma = c\sqrt{(\log M)/n}$ we have

$$\mathbb{P}\left(\|(\mathbf{a}_i)_{G_j} - \frac{1}{n}\widetilde{\mathbf{Z}}_{G_j}^T\widetilde{\mathbf{Z}}\Theta\mathbf{e}_i\|_2 \geq c\sqrt{\frac{\log M}{n}}\right) \leq 2e^{Cm}M^{-c^2/(6e^2\phi'^2)}$$
$$= 2e^{Cm}M^{-(c^2C_{\min})/(24e^2\phi^4C_{\max})}.$$

Therefore, by union bounding over all pairs of $i \in [p]$ and $j \in [M]$,

$$\mathbb{P}\left\{\|\Theta\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - \mathbf{P_C})\|_{2,\infty} \geq c\sqrt{\frac{\log M}{n}}\right\} \leq 2pe^{Cm}M^{-(c^2C_{\min})/(24e^2\phi^4C_{\max})+1}.$$

□

S2.4 Proof of Theorem 2

Proof. Note that using our model (6), we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^* &= \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* + \frac{1}{n}\widehat{\mathbf{\Theta}}\widetilde{\mathbf{Z}}^T\boldsymbol{\epsilon} + \frac{1}{n}\widehat{\mathbf{\Theta}}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) \\ &= \frac{1}{n}\widehat{\mathbf{\Theta}}\widetilde{\mathbf{Z}}^T\boldsymbol{\epsilon} + (\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - \mathbf{I}_p)(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) \\ &= \frac{1}{n}\widehat{\mathbf{\Theta}}\widetilde{\mathbf{Z}}^T\boldsymbol{\epsilon} + (\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - \mathbf{P_C}))(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}), \text{ since } \mathbf{C}^T\boldsymbol{\beta}^* = \mathbf{C}^T\hat{\boldsymbol{\beta}} = 0.\end{aligned}$$

Thus, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = B + \Delta$ where $B = \frac{1}{\sqrt{n}}\widehat{\mathbf{\Theta}}\widetilde{\mathbf{Z}}^T\boldsymbol{\epsilon}$. Also note that $(\mathbf{I}_p - \mathbf{P_C})\widehat{\mathbf{\Sigma}}(\mathbf{I}_p - \mathbf{P_C}) = \widehat{\mathbf{\Sigma}}$, and $B = \frac{1}{\sqrt{n}}\widehat{\mathbf{\Theta}}\widetilde{\mathbf{Z}}^T\boldsymbol{\epsilon} = \frac{1}{\sqrt{n}}\widehat{\mathbf{\Theta}}(\mathbf{I}_p - \mathbf{P_C})\mathbf{Z}^T\boldsymbol{\epsilon}$. Thus,

$$B|\mathbf{Z} \sim N(\mathbf{0}, \sigma^2\widehat{\mathbf{\Theta}}(\mathbf{I}_p - \mathbf{P_C})\widehat{\mathbf{\Sigma}}(\mathbf{I}_p - \mathbf{P_C})\widehat{\mathbf{\Theta}}^T) = N(\mathbf{0}, \sigma^2\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Theta}}^T).$$

Note that

$$|\langle\mathbf{x}, \mathbf{y}\rangle| = |\sum_{j=1}^{M}\langle\mathbf{x}_{G_j}, \mathbf{y}_{G_j}\rangle| \leq \sum_{j=1}^{M}\|\mathbf{x}_{G_j}\|_2\|\mathbf{y}_{G_j}\|_2 \leq \|\mathbf{x}\|_{2,\infty}\|\mathbf{y}\|_{2,1}.$$

Using the above relationship, we have

$$\begin{aligned}\|\Delta\|_\infty &= \max_i |\langle(\mathbf{I}_p - \mathbf{P_C})\mathbf{e}_i - \widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}}\mathbf{e}_i, (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\rangle| \\ &\leq \sqrt{n}\|\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - \mathbf{P_C})\|_{2,\infty}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{2,1}.\end{aligned}$$

By Lemma 1, when choosing $\gamma = c\sqrt{(\log M)/n}$, $\widehat{\mathbf{\Theta}}$ is a feasible solution of the optimization problem (16) with probability at least $1 - 2pe^{Cm}M^{-c'}$. Therefore, $\|\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - \mathbf{P_C})\|_{2,\infty} \leq c\sqrt{(\log M)/n}$ with probability at least $1 - 2pe^{Cm}M^{-c'}$. By Theorem 1, take

$$\lambda = \max_{j\in[M]}\frac{2\sigma}{\sqrt{n}}\sqrt{\text{trace}(\widehat{\mathbf{\Sigma}}_j) + 2\|\widehat{\mathbf{\Sigma}}_j\|(2q\log M + \sqrt{m_j q\log M})},$$

then we have

$$\mathbb{P}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_{2,1} \leq \frac{16\lambda s}{\kappa^2}\right) \geq 1-2M^{1-q},$$

where q is a positive parameter. Altogether, we have

$$\begin{aligned}
&\mathbb{P}\left\{\|\Delta\|_\infty > \frac{16c\lambda s\sqrt{\log M}}{\kappa^2\sqrt{n}}\right\} \\
&\leq \mathbb{P}\left\{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_{2,1} \leq \frac{16\lambda s}{\kappa^2}\right\} + \mathbb{P}\left\{\|\Theta\widehat{\boldsymbol{\Sigma}}-(\mathbf{I}_p-\mathbf{P}_{\mathbf{C}})\|_{2,\infty} \leq c\sqrt{\frac{\log M}{n}}\right\} \\
&\leq 2M^{1-q} + 2pe^{Cm}M^{-c'}.
\end{aligned}$$

□

S2.5 Poof of Theorem 3

Theorem 3 is verified by the following lemma, which closely follows the results of Sun and Zhang (2012) and Mitra and Zhang (2016). First, define

$$\mu(\lambda,\xi) = \frac{2\xi\lambda^2}{\kappa}, \quad \tau_- = \frac{2\mu(\lambda,\xi)(\xi-1)}{\xi+1}, \quad \tau_+ = \frac{\tau_-}{2} + \mu(\lambda,\xi)(k_0\sqrt{mM}+1/2),$$

where $m = \max_{j\in[M]} m_j$. Lastly, define a constant

$$A^* = \frac{(\xi+1)/(\xi-1)}{\{\sqrt{1-2\mu(\lambda^*,\xi)(\xi+1)/(\xi-1)}\}_+}.$$

Then, based on the idea of Theorem 6 in Mitra and Zhang (2016), we have the following lemma.

Lemma S4. *Let $(\hat{\boldsymbol{\beta}},\hat{\sigma})$ be a solution of the optimization problem* (20) *and $\boldsymbol{\beta}^*$ be a vector with $supp(\boldsymbol{\beta}^*) \subset G_{S^*}$ for some $S^* \subset T = [M]$. Let $\xi > 1$. Suppose $RE(s)$ condition holds with κ and $\tau_+ < 1$. Define the following event*

$$\mathcal{E} = \left\{\max_{j\in[M],\mathbf{C}^T\boldsymbol{\beta}^*=\mathbf{0}} \frac{\widetilde{\mathbf{Z}}_{G_j}^T(\mathbf{y}-\widetilde{\mathbf{Z}}\boldsymbol{\beta}^*)}{\lambda n\sigma^*/\sqrt{1+\tau_-}} < \frac{\xi-1}{\xi+1}\right\}, \tag{S18}$$

where $\sigma^* = \|\mathbf{y}-\widetilde{\mathbf{Z}}\boldsymbol{\beta}^*\|_2/\sqrt{n}$ is the oracle noise level. Suppose $\lambda \geq A\|\mathbf{X}_{G_j}/\sqrt{n}\|_S\lambda_*$ with $A \geq A^*$. Then,

$$\mathbb{P}(\mathcal{E}) \geq 1-\delta \tag{S19}$$

with the event $\mathcal{E}$ is in (S18). *Moreover, if $\sqrt{n}\mu(\lambda,\xi) \to 0$, then $\hat{\sigma}$ is a consistent estimator of σ, in the sense that for any $\epsilon > 0$,*

$$\lim_{n,p\to\infty} \mathbb{P}\left(\left|\frac{\hat{\sigma}}{\sigma}-1\right| \geq \epsilon\right) = 0.$$

Proof of Lemma S4

Proof. We follow the proof in Mitra and Zhang (2016), Sun and Zhang (2012). Let $t \geq \sigma^*/\sqrt{1+\tau_-}$, $\mathbf{h}_{G_j} = \hat{\boldsymbol{\beta}}_{G_j}(t\lambda) - \boldsymbol{\beta}^*_{G_j}$, and $\mathbf{h} = \hat{\boldsymbol{\beta}}(t\lambda) - \boldsymbol{\beta}^*$. Then, using $\mathbf{y} = \widetilde{\mathbf{Z}}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, we have

$$\begin{aligned}(\sigma^*)^2 - \|\mathbf{y} - \widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}(t\lambda)\|_2^2/n &= (\widetilde{\mathbf{Z}}\mathbf{h})^T(2\boldsymbol{\epsilon} - \widetilde{\mathbf{Z}}\mathbf{h})/n \\ &= (\widetilde{\mathbf{Z}}\mathbf{h})^T\boldsymbol{\epsilon}/n + (\widetilde{\mathbf{Z}}\mathbf{h})^T(\mathbf{y} - \widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}(t\lambda))/n. \end{aligned} \tag{S20}$$

Suppose $\mathcal{E}$ happens so that $\|\widetilde{\mathbf{Z}}_{G_j}\boldsymbol{\epsilon}\|_2/n \leq t\lambda(\xi-1)/(\xi+1)$. It follows that

$$|(\widetilde{\mathbf{Z}}\mathbf{h})^T\boldsymbol{\epsilon}/n| = \left|\sum_{j=1}^M \mathbf{h}_{G_j}^T\widetilde{\mathbf{Z}}_{G_j}^T\boldsymbol{\epsilon}/n\right| \leq \sum_{j=1}^M \|\mathbf{h}_{G_j}\|_2\|\widetilde{\mathbf{Z}}_{G_j}\boldsymbol{\epsilon}\|_2/n \leq \frac{\xi-1}{\xi+1}t\lambda\|\mathbf{h}\|_{2,1}.$$

Moreover, the KKT condition implies

$$|\mathbf{h}^T\widetilde{\mathbf{Z}}^T(\mathbf{y} - \widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}(t\lambda))/n| \leq \|\widetilde{\mathbf{Z}}^T(\mathbf{y} - \widetilde{\mathbf{Z}}\boldsymbol{\beta}(t\lambda))\|_{2,\infty}\|\mathbf{h}\|_{2,1}.$$

Note that by Lemma S3, we have

$$\begin{aligned}\|\widetilde{\mathbf{Z}}^T(\mathbf{y} - \widetilde{\mathbf{Z}}\boldsymbol{\beta}(t\lambda))/n\|_{2,\infty} &= \max_{j\in[M]} \|\widetilde{\mathbf{Z}}_{G_j}^T(\mathbf{y} - \widetilde{\mathbf{Z}}\boldsymbol{\beta}(t\lambda))/n\|_2 \\ &\leq \max_{j\in[M]} \|\widetilde{\mathbf{Z}}_{G_j}^T(\widetilde{\mathbf{Z}}\mathbf{h})\|_2\|\widetilde{\mathbf{Z}}_{G_j}^T\boldsymbol{\epsilon}\|_2 \\ &\leq t\lambda\left(k_0\sqrt{mM} + \frac{1}{2} + \frac{\xi-1}{\xi+1}\right).\end{aligned}$$

Hence, we have the following inequality

$$|\mathbf{h}^T\widetilde{\mathbf{Z}}^T(\mathbf{y} - \widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}(t\lambda))/n| \leq t\lambda\left(k_0\sqrt{mM} + \frac{1}{2} + \frac{\xi-1}{\xi+1}\right)\|\mathbf{h}\|_{2,1}.$$

As $(\widetilde{\mathbf{Z}}\mathbf{h})^T(2\boldsymbol{\epsilon} - \widetilde{\mathbf{Z}}\mathbf{h})/n \leq 2(\widetilde{\mathbf{Z}}\mathbf{h})^T\boldsymbol{\epsilon}/n$, inserting these inequalities to (S20) yields

$$\begin{aligned}-t\lambda\left(k_0\sqrt{mM} + \frac{1}{2} + 2\frac{\xi-1}{\xi+1}\right)\|\mathbf{h}\|_{2,1} &\leq (\sigma^*)^2 - \|\mathbf{y} - \widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}(t\lambda)\|_2^2/n \\ &\leq 2\frac{\xi-1}{\xi+1}t\lambda\sum_{j=1}^M\|\mathbf{h}\|_{2,1}.\end{aligned}$$

A rescaled version $\hat{\boldsymbol{\beta}}(t\lambda)$ can be written as

$$\frac{\hat{\boldsymbol{\beta}}(t\lambda)}{t} = \underset{\mathbf{C}^T\mathbf{b}=\mathbf{0}}{\operatorname{argmin}}\left\{\frac{1}{2n}\|\mathbf{y}/t - \widetilde{\mathbf{Z}}\mathbf{b}\|_2^2 + \lambda\sum_{j=1}^M\|\mathbf{b}_{G_j}\|_2\right\},$$

as the group Lasso estimator with target $\boldsymbol{\beta}^*/t$ and noise vector $\boldsymbol{\epsilon}/t$. As $t \geq \sigma^*/\sqrt{1+\tau_-}$,

$$\frac{\lambda}{t}\sum_{j=1}^M\|\mathbf{h}_{G_j}\|_2 = \sum_{j=1}^M\|\hat{\boldsymbol{\beta}}_{G_j}(t\lambda)/t - \boldsymbol{\beta}^*_{G_j}/t\|_2 < \mu(\lambda,\xi).$$

As $\tau_- = 2\mu(\lambda,\xi)(\xi-1)/(\xi+1)$ and $\tau_+ = \mu(\lambda,\xi)\left(k_0\sqrt{mM} + \frac{1}{2} + 2\frac{\xi-1}{\xi+1}\right)$, we have

$$-\tau_+ t^2 = -\left(k_0\sqrt{mM} + \frac{1}{2} + 2\frac{\xi-1}{\xi+1}\right) t^2\mu(\lambda,\xi) < (\sigma^*)^2 - \|\mathbf{y} - \widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}(t\lambda)\|_2^2/n$$
$$< 2\frac{\xi-1}{\xi+1} t^2\mu(\lambda,\xi) = \tau_- t^2.$$

The upper bound above for $t = \sigma^*/\sqrt{1+\tau_-}$ implies

$$t^2 - \|\mathbf{y} - \widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}(t\lambda)\|_2^2/n < t^2 - (\sigma^*)^2 + \tau_- t^2 = 0,$$

so that $\hat{\sigma} > t = \sigma^*/\sqrt{1+\tau_-}$. Similarly, combining the results from the lower bound yields

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

This implies with the condition on $\mu(\lambda,\xi)$, $\sqrt{n}\mu(\lambda,\xi) \to 0$,

$$|\hat{\sigma}/\sigma^* - 1| = o_P(\mu(\lambda,\xi)) = o_P(n^{-1/2})$$

Therefore, by the central limit theorem and $\sigma^*/\sigma \sim \chi_n/\sqrt{n}$, which completes the proof. □

S2.6 Proof of Theorem 4

Proof. In Theorem 2, when choosing proper λ, then we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^d_{G_j} - \boldsymbol{\beta}^*_{G_j}) \sim N_{m_j}(0, \sigma^2\mathbf{V}_{G_j,G_j}),$$

where $\Delta = \sqrt{n}(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}} - (\mathbf{I}_p - \mathbf{P}_\mathbf{C}))(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})$ and $\mathbf{V} = \sigma^2\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}^T$. Additionally, using Theorem S4, we have

$$\lim_{n,p\to\infty} \mathbb{P}\left(\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \epsilon\right) = 0.$$

Since $\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}^T$ is not random, we can conclude that

$$\lim_{n,p\to\infty} \mathbb{P}\left(\|\widehat{\mathbf{V}} - \mathbf{V}\|_\infty \geq \epsilon\right) = 0,$$

which completes the proof. □

S2.7 Computation complexity

In this section, we provide the computation complexity of the proposed methods, Compositional Group Lasso (CGL) and the debiased CGL, CGL(D). In short, the computation complexities of these methods are:

- CGL: $O(kLT(np + Mm^2))$,
- CGL(D): $O(p^4/c + FT(np + Mm^2) + np^2)$.

Here, n, p, and M denote the numbers of samples, variables, and groups, respectively, and m is the maximum group size. k is the number of cross-validation folds, L is the number of candidate values for λ, and (T, F) denote the numbers of iterations for the block coordinate descent and the scaled group lasso operation, respectively.

CGL. We implemented the CGL estimator using the `R` package `gglasso` (Yang et al., 2024), solving the group lasso optimization problem with compositional constraints (Equation (4) in the main article). The package employs block coordinate descent. The complexity per iteration is approximately

$$O(np) + O(Mm^2),$$

where $O(np)$ accounts for residual computation, and $O(Mm^2)$ arises from group-wise soft-thresholding and updates. Over T iterations, the total cost is

$$O(T(np + Mm^2)).$$

Considering L candidate values for λ and k-fold cross-validation, the total complexity becomes

$$O(kLT(np + Mm^2)).$$

CGL(D). Obtaining the debiased compositional group lasso estimates involves three steps:

Step 1: Precision matrix estimation (Algorithm 1)

Step 2: Noise variance estimation (Scaled group lasso; Equation (17) in the main article)

Step 3: Debiasing (Equation (15) in the main article)

Step 1. Precision matrix estimation (Algorithm 1) is performed by solving p constrained convex programs using the `CVXR` package. Each problem has a complexity of $O(p^3)$, resulting in a total cost of

$$O(p^4).$$

Additionally, computing the sample covariance matrix incurs a cost of $O(np^2)$.

Step 2. Scaled group lasso requires F outer iterations to jointly estimate the noise level and regression coefficients, each invoking a group lasso solver over T iterations. Thus, the total complexity is

$$O(FT(np + Mm^2)).$$

Step 3. The debiasing step (Equation (15)) involves matrix-vector operations, with total complexity $O(np + p^2)$.

Therefore, the total cost of CGL(D) is

$$O(p^4 + FT(np + Mm^2) + np^2).$$

Since the precision matrix estimation dominates the overall cost, we note that parallelizing this step across c cores reduces the total cost to

$$O(p^4/c + FT(np + Mm^2) + np^2).$$

S3 Additional Tables

Table S1: Average performance measures (with standard errors) of the four methods for the same signal pattern, Model 1 when $(\rho_1, \rho_2) = (0, 0)$.

τ	(n, M)	Method	$\|M(\hat{\boldsymbol{\beta}})\|_0$	sensitivity	specificity	PPV
3	(100, 10)	CL	5.98 (0.15)	1 (0.00)	0.57 (0.02)	0.54 (0.02)
		CL(D)	1.17 (0.04)	0.39 (0.01)	1 (0.00)	0.99 (0.01)
		GL	6.89 (0.18)	1 (0.00)	0.44 (0.03)	0.47 (0.02)
		CGL	8.06 (0.16)	1 (0.00)	0.28 (0.02)	0.39 (0.01)
		GL(D)	2.49 (0.05)	0.83 (0.02)	1 (0.00)	1 (0.00)
		CGL(D)	4.06 (0.13)	1 (0.00)	0.85 (0.02)	0.8 (0.02)
	(300, 20)	CL	5.55 (0.16)	1 (0.00)	0.85 (0.01)	0.59 (0.02)
		CL(D)	2.58 (0.05)	0.86 (0.02)	1 (0.00)	1 (0.00)
		GL	9.32 (0.34)	1 (0.00)	0.63 (0.02)	0.37 (0.01)
		CGL	11.99 (0.28)	1 (0.00)	0.47 (0.02)	0.27 (0.01)
		GL(D)	3 (0.00)	1 (0.00)	1 (0.00)	1 (0.00)
		CGL(D)	3.83 (0.1)	1 (0.00)	0.95 (0.01)	0.83 (0.02)
	(500, 20)	CL	5.42 (0.17)	1 (0.00)	0.86 (0.01)	0.61 (0.02)
		CL(D)	3 (0.00)	1 (0.00)	1 (0.00)	1 (0.00)
		GL	9.17 (0.32)	1 (0.00)	0.64 (0.02)	0.38 (0.02)
		CGL	12.07 (0.28)	1 (0.00)	0.47 (0.02)	0.26 (0.01)
		GL(D)	3 (0.00)	1 (0.00)	1 (0.00)	1 (0.00)
		CGL(D)	3.78 (0.11)	1 (0.00)	0.95 (0.01)	0.85 (0.02)
0.3	(100, 10)	CL	0.55 (0.05)	0.1 (0.02)	0.97 (0.01)	0.3 (0.05)
		CL(D)	0.03 (0.02)	0.01 (0.00)	1 (0.00)	0.02 (0.01)
		GL	3.09 (0.27)	0.43 (0.03)	0.74 (0.03)	0.42 (0.04)
		CGL	3.88 (0.32)	0.53 (0.04)	0.67 (0.03)	0.36 (0.03)
		GL(D)	0.26 (0.05)	0.08 (0.02)	1 (0.00)	0.2 (0.04)
		CGL(D)	1.2 (0.15)	0.22 (0.03)	0.92 (0.01)	0.35 (0.04)
	(300, 20)	CL	0.21 (0.04)	0.05 (0.01)	1 (0.00)	0.14 (0.03)
		CL(D)	0.12 (0.04)	0.03 (0.01)	1 (0.00)	0.06 (0.02)
		GL	6.58 (0.42)	0.7 (0.03)	0.74 (0.02)	0.44 (0.03)
		CGL	9.65 (0.37)	0.87 (0.02)	0.59 (0.02)	0.3 (0.02)
		GL(D)	1.01 (0.08)	0.31 (0.02)	1 (0.00)	0.72 (0.04)
		CGL(D)	2.05 (0.15)	0.46 (0.02)	0.96 (0.01)	0.7 (0.04)
	(500, 20)	CL	0.25 (0.04)	0.08 (0.01)	1 (0.00)	0.25 (0.04)
		CL(D)	0.47 (0.07)	0.14 (0.02)	1 (0.00)	0.33 (0.05)
		GL	7.22 (0.42)	0.87 (0.02)	0.73 (0.02)	0.48 (0.02)
		CGL	11.3 (0.33)	0.97 (0.01)	0.51 (0.02)	0.29 (0.01)
		GL(D)	1.67 (0.07)	0.52 (0.02)	0.99 (0.00)	0.96 (0.02)
		CGL(D)	2.54 (0.12)	0.66 (0.02)	0.97 (0.01)	0.84 (0.02)

Table S2: Average performance measures (with standard errors) of the four methods for the different signal pattern, Model 2 when $(\rho_1, \rho_2) = (0, 0)$.

τ	(n, M)	Method	$\|M(\hat{\boldsymbol{\beta}})\|_0$	sensitivity	specificity	PPV
3	(100, 10)	CL	4.5 (0.14)	1 (0.00)	0.79 (0.02)	0.73 (0.02)
		CL(D)	2.15 (0.08)	0.72 (0.03)	1 (0.00)	1 (0.00)
		GL	7.39 (0.18)	1 (0.00)	0.37 (0.03)	0.43 (0.01)
		CGL	8.07 (0.14)	1 (0.00)	0.28 (0.02)	0.38 (0.01)
		GL(D)	2.85 (0.04)	0.95 (0.01)	1 (0.00)	1 (0.00)
		CGL(D)	3.96 (0.12)	1 (0.00)	0.86 (0.02)	0.81 (0.02)
	(300, 20)	CL	4.37 (0.13)	1 (0.00)	0.92 (0.01)	0.74 (0.02)
		CL(D)	3 (0.00)	1 (0.00)	1 (0.00)	1 (0.00)
		GL	9.74 (0.33)	1 (0.00)	0.6 (0.02)	0.35 (0.02)
		CGL	12.68 (0.27)	1 (0.00)	0.43 (0.02)	0.25 (0.01)
		GL(D)	3 (0.00)	1 (0.00)	1 (0.00)	1 (0.00)
		CGL(D)	3.79 (0.1)	1 (0.00)	0.95 (0.01)	0.84 (0.02)
	(500, 20)	CL	4.12 (0.12)	1 (0.00)	0.93 (0.01)	0.78 (0.02)
		CL(D)	3.02 (0.01)	1 (0.00)	1 (0.00)	1 (0.00)
		GL	9.98 (0.35)	1 (0.00)	0.59 (0.02)	0.34 (0.01)
		CGL	13.08 (0.27)	1 (0.00)	0.41 (0.02)	0.24 (0.01)
		GL(D)	3 (0.00)	1 (0.00)	1 (0.00)	1 (0.00)
		CGL(D)	3.5 (0.08)	1 (0.00)	0.97 (0.00)	0.89 (0.02)
0.3	(100, 10)	CL	0.59 (0.06)	0.15 (0.02)	0.98 (0.00)	0.4 (0.05)
		CL(D)	0.2 (0.05)	0.06 (0.01)	1 (0.00)	0.17 (0.04)
		GL	3.43 (0.31)	0.53 (0.04)	0.74 (0.03)	0.47 (0.04)
		CGL	4.73 (0.31)	0.65 (0.04)	0.6 (0.03)	0.38 (0.03)
		GL(D)	0.26 (0.05)	0.08 (0.02)	1 (0.00)	0.22 (0.04)
		CGL(D)	1.78 (0.21)	0.31 (0.03)	0.88 (0.02)	0.42 (0.04)
	(300, 20)	CL	0.77 (0.1)	0.24 (0.03)	1 (0.00)	0.41 (0.05)
		CL(D)	0.82 (0.06)	0.25 (0.02)	1 (0.00)	0.68 (0.05)
		GL	6.83 (0.38)	0.78 (0.03)	0.74 (0.02)	0.45 (0.03)
		CGL	10.61 (0.36)	0.94 (0.02)	0.54 (0.02)	0.29 (0.01)
		GL(D)	1 (0.06)	0.31 (0.02)	1 (0.00)	0.78 (0.04)
		CGL(D)	2.26 (0.16)	0.5 (0.03)	0.96 (0.01)	0.74 (0.03)
	(500, 20)	CL	2.13 (0.13)	0.67 (0.04)	0.99 (0.00)	0.78 (0.04)
		CL(D)	1.57 (0.09)	0.5 (0.02)	1 (0.00)	0.96 (0.02)
		GL	8.43 (0.4)	0.9 (0.02)	0.66 (0.02)	0.41 (0.02)
		CGL	11.4 (0.26)	0.99 (0.01)	0.5 (0.02)	0.28 (0.01)
		GL(D)	1.74 (0.08)	0.55 (0.02)	0.99 (0.00)	0.96 (0.01)
		CGL(D)	2.72 (0.12)	0.74 (0.03)	0.97 (0.00)	0.87 (0.02)

Table S3: Average performance measures (with standard errors) for the same signal pattern, Model 1 when $(\rho_1, \rho_2) = (0.5, 0.1)$, and $\tau = 0.3$.

Method	(n, M)	$\|G(\hat{\boldsymbol{\beta}})\|_0$	sensitivity	specificity	PPV
CGL	(1000, 20)	11 (0.48)	1 (0.00)	0.53 (0.04)	0.28 (0.05)
	(2000, 20)	10.5 (0.76)	1 (0.00)	0.56 (0.04)	0.32 (0.03)
CGL(D)	(1000, 20)	3.4 (0.25)	0.87 (0.04)	0.95 (0.01)	0.83 (0.04)
	(2000, 20)	3.3 (0.29)	0.93 (0.02)	0.97 (0.02)	0.88 (0.04)

References

Aitchison, J. (1982), "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 44, 139–160.

— (1986), *The Statistical Analysis of Compositional Data*, Chapman & Hall.

Aitchison, J. and Bacon-Shone, J. (1984), "Log contrast models for experiments with mixtures," *Biometrika*, 71, 323–330.

Bates, S. and Tibshirani, R. (2019), "Log-ratio lasso: scalable, sparse estimation for log-ratio models," *Biometrics*, 75, 613–624.

Cai, T. T., Zhang, A. R., and Zhou, Y. (2022), "Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference," *IEEE transactions on information theory*, 68, 5975–6002.

Cavalier, L., Golubev, G., Picard, D., and Tsybakov, A. (2002), "Oracle inequalities for inverse problems," *The Annals of Statistics*, 30, 843–874.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003), "Isometric logratio transformations for compositional data analysis," *Mathematical geology*, 35, 279–300.

Greenacre, M. (2019), "Variable selection in compositional data analysis using pairwise logratios," *Mathematical Geosciences*, 51, 649–682.

Greenacre, M., Martínez-Álvaro, M., and Blasco, A. (2021), "Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation," *Frontiers in microbiology*, 12, 727398.

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

Lee, S., Jung, S., Lourenco, J., Pringle, D., and Ahn, J. (2023), "Resampling-based inferences for compositional regression with application to beef cattle microbiomes," *Statistical Methods in Medical Research*, 32, 151–164.

Li, G., Li, Y., and Chen, K. (2023), "It's all relative: Regression analysis with compositional predictors," *Biometrics*, 79, 1318–1329.

Lin, W., Shi, P., Feng, R., and Li, H. (2014), "Variable selection in regression with compositional covariates," *Biometrika*, 101, 785–797.

Lounici, K., Pontil, M., Van De Geer, S., and Tsybakov, A. B. (2011), "Oracle inequalities and optimal inference under group sparsity," *The Annals of Statistics*, 39, 2164–2204.

Lu, J., Shi, P., and Li, H. (2019), "Generalized linear models with linear constraints for microbiome compositional data," *Biometrics*, 75, 235–244.

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

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

Sun, T. and Zhang, C.-H. (2012), "Scaled sparse linear regression," *Biometrika*, 99, 879–898.

Sun, Z., Xu, W., Cong, X., Li, G., and Chen, K. (2020), "Log-contrast regression with functional compositional predictors: linking preterm infant's gut microbiome trajectories to neurobehavioral outcome," *The annals of applied statistics*, 14, 1535.

Susin, A., Wang, Y., Lê Cao, K.-A., and Calle, M. L. (2020), "Variable selection in microbiome compositional data analysis," *NAR Genomics and Bioinformatics*, 2, lqaa029.

Vershynin, R. (2012), "Introduction to the non-asymptotic analysis of random matrices," *Compressed Sensing: Theory and Applications*, 210–268.

Wang, T. and Zhao, H. (2017), "Structured subcomposition selection in regression and its application to microbiome data analysis," *The Annals of Applied Statistics*, 11, 771 – 791.

Yang, Y., Zou, H., and Bhatnagar, S. (2024), *gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm*, r package version 1.5.1.