Supporting Information for "It's All Relative: Regression Analysis with Compositional Predictors"

by Gen Li, Yan Li, Kun Chen

A Additional Simulation Studies

A.1 Log-Contrast Settings with Varying n, p, Zero Proportion, and SNR

In this study, we consider a model misspecified setting, where the response is generated from the log-contrast model. More specifically, the relative abundances $\boldsymbol{X}$, the truncated version of the relative abundances $\boldsymbol{X}_0$ (when applicable), and the read counts are generated in the same way as in Study I. Then we generate the response from the following log-contrast model with the true relative abundances in $\boldsymbol{X}$

$$y_i = \beta_0 + \beta_1 \log x_{i1} + \cdots + \beta_p \log x_{ip} + \varepsilon_i,$$

where $\sum_{j=1}^{p} \beta_j = 0$. The true coefficient vector is $\boldsymbol{\beta} = (1, -0.8, 0.6, 0, 0, -1.5, -0.5, 1.2, 0, 0, \mathbf{0}_{p-10}^T)^T$, where $\mathbf{0}_q$ is a length-q vector of zeros. Namely, only 6 out of the first 10 features have nonzero coefficients while all the remaining features have zero coefficients.

Varying n and p. We first consider settings with different n and p. In particular, the training sample size n_{train} is set to be 50 and 100, and the dimension p is set to be 100 and 200. The testing sample size n_{test} is fixed to be 400 throughout the settings. In addition, there is no zero truncation (i.e., $\boldsymbol{X} = \boldsymbol{X}_0$) and the SNR is fixed at 1. The MPSE comparison between different methods is shown in Figure S.1. In general, the results are similar across different settings. Although a misspecified scenario for the proposed method, RS-ES still has comparable prediction performance with the transformation-based methods. Figure S.2 shows the computing times for different methods under different settings. The proposed method is slightly more time-consuming than the competitors. The computing time of RS-ES is mostly affected by the dimension p, but not the sample size n.

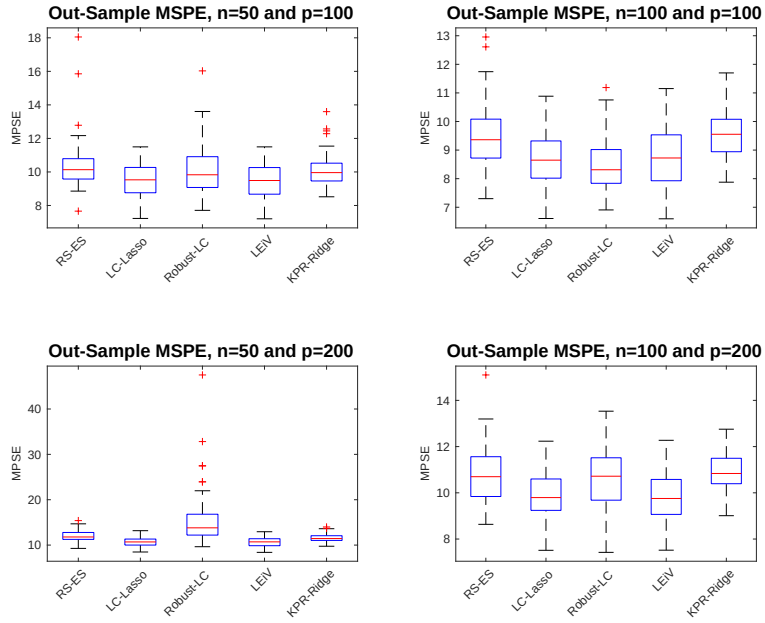


Figure S.1: Boxplots of MPSE in the log-contrast settings with varying n and p .

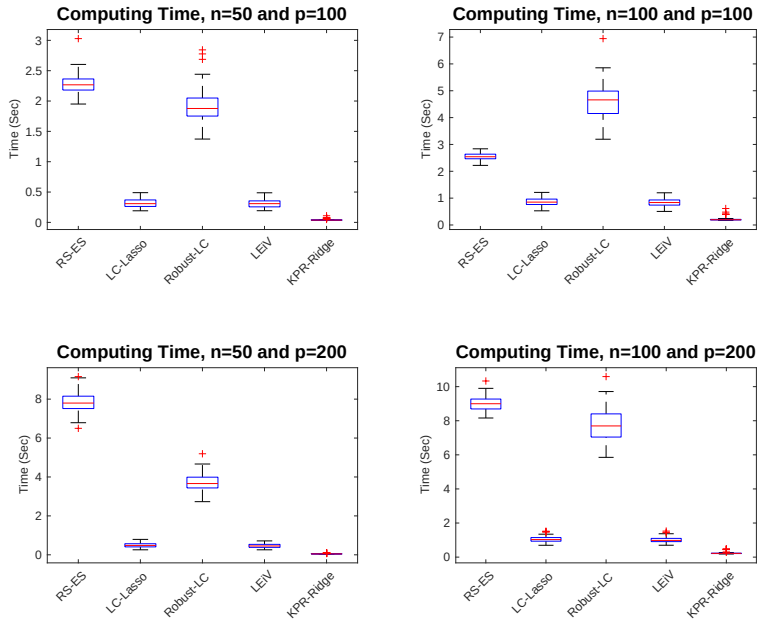


Figure S.2: Boxplots of computing time in the log-contrast settings with varying n and p.

Varying zero proportion and SNR. Next, we vary the zero proportion of $\boldsymbol{X}_0$ and SNR, and evaluate how the prediction performance is affected in different methods. Recall that the response is generated using the true relative abundances $\boldsymbol{X}$, so varying the zero proportion of $\boldsymbol{X}_0$ does not affect the response. The truncated relative abundances $\boldsymbol{X}_0$ are only used in generating raw read counts, which are subsequently used in model fitting and testing. We fix $n_{train} = 100$, $n_{test} = 400$, $p = 100$, and consider a range of zero proportions in $\boldsymbol{X}_0$ (i.e., 0%, 20%, 40%, 60%, and 80%). Moreover, we consider a low-signal setting (i.e., SNR=1) and a high-signal setting (i.e., SNR=10). The MPSE comparison results are shown in Figures S.3 and S.4.

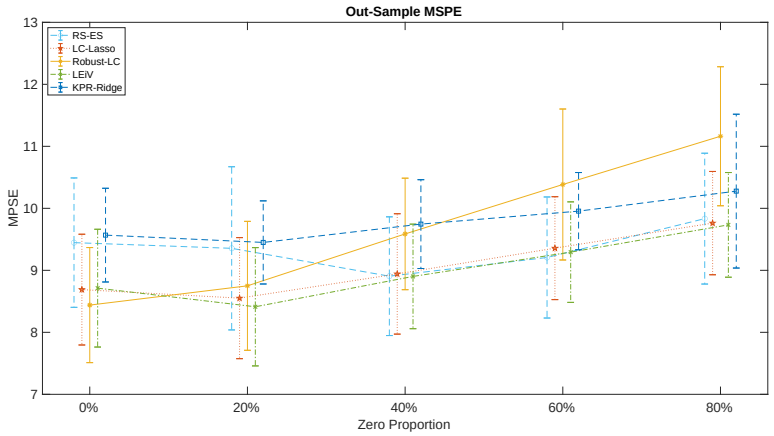


Figure S.3: Mean and standard deviation (i.e., error bar) of MPSE in the log-contrast settings with varying zero proportions (SNR=1).

We observe that the transformation-based methods (LC-Lasso, Robust-LC, LEiV, and KPR-Ridge) all have increasing MPSE when the zero proportion in data increases. In comparison, the proposed RS-ES method is relatively stable. This is especially evident in the high-signal setting (Figure S.4). In particular, when the zero proportion reaches 80%, RS-ES has the smallest MPSE among all, despite that the generative model (i.e., the log-contrast model) is misspecified for RS-ES. Therefore, we conclude that the proposed relative-shift framework is more robust against varying zero proportions than the transformation-based methods. This is likely due to the fact that the relative-shift model does not rely on any arbitrary zero-replacement strategy.

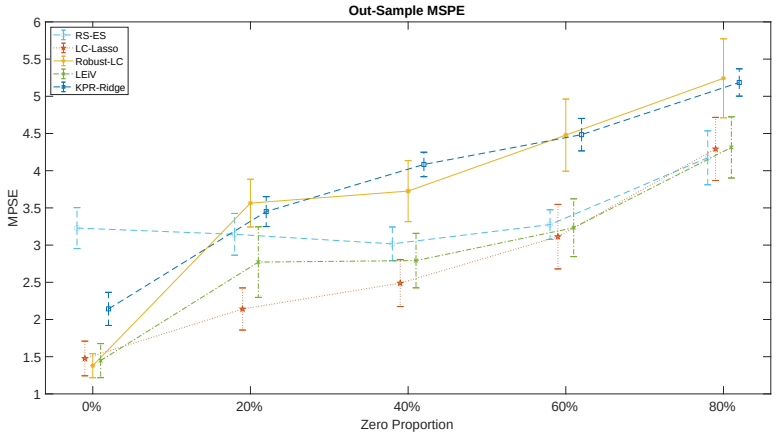


Figure S.4: Mean and standard deviation (i.e., error bar) of MPSE in the log-contrast settings with varying zero proportions (SNR=10).

A.2 Tree-Guided Equi-Sparsity Setting with High Dimension

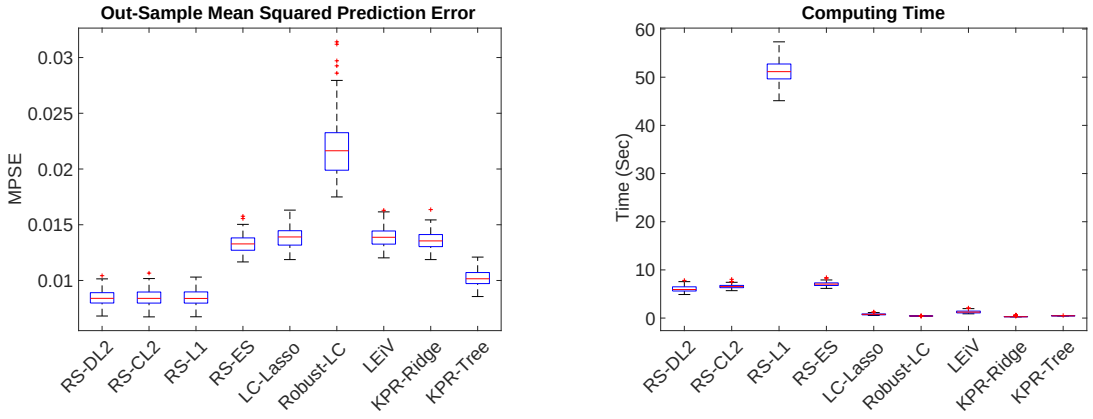
In this study, we follow the same data generation procedure as in Study II (tree-guided equi-sparsity setting), but increase the dimension p from 100 to 400 and 1000, respectively. The auxiliary tree structure and the tree-guided true coefficient vector are similar in essence to those in Study II, but proportionally expanded to a higher dimension. For example, when $p = 400$, we create the tree structure so that every $10 \times 4 = 40$ consecutive leaf nodes share a common parent node and so on. The upper-level structure of the tree remains the same as in Figure 3 in the main paper. Correspondingly, the true coefficient vector is set to be

$$\boldsymbol{\beta} = \left(\mathbf{1}_{80}^T,\ -2 \times \mathbf{1}_{40}^T,\ 0.5 \times \mathbf{1}_{40}^T,\ 2 \times \mathbf{1}_{160}^T,\ \boldsymbol{\xi}_{80}^T\right)^T.$$

The same goes for $p = 1000$. The comparison of MPSE and computing times between different methods is presented in Figure S.5.

The results are similar to those in the main paper. The relative-shift methods with different tree-guided regularization have very similar MPSE and significantly outperform other competing methods (including the RS-ES method without the tree information). KPR-Tree has the second best performance due to its ability to account for the tree structure. In terms of the computing time, the proposed methods (especially RS-L1) are slightly slower than the competing methods,

(a) High-Dimensional Setting ($p = 400$)



(b) High-Dimensional Setting ($p = 1000$)

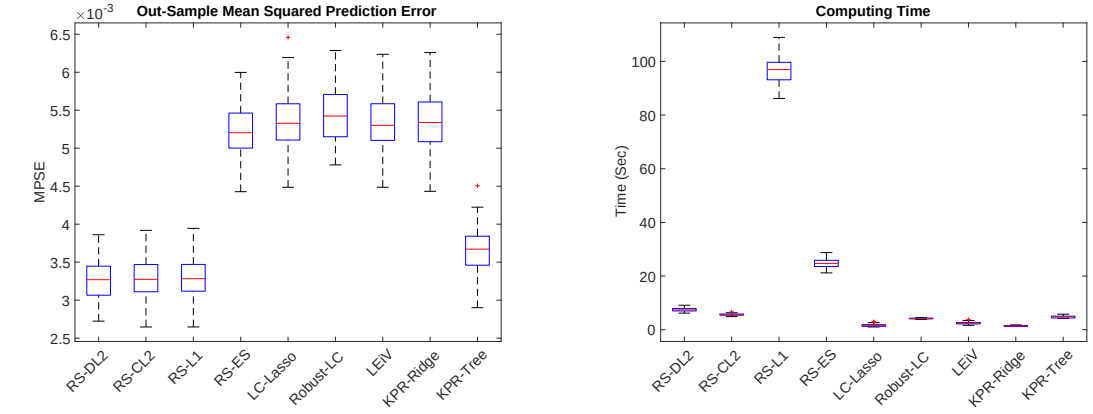


Figure S.5: Boxplots of MPSE and computing time in the tree-guided equi-sparsity setting with high dimension.

but the model fitting procedure (with cross validation for tuning parameter selection) always finishes within 1 minute in each simulation run (except for RS-L1 when $p = 1000$). Considering the size of the problem, we deem the methods to be computationally efficient.

A.3 Comparison between RS-L1, RS-CL2, and RS-DL2

In this study, we further investigate the differences between RS-L1, RS-CL2, and RS-DL2. As proof of concept, we focus on a small-scale setting with $p = 6$ variables. The tree structure among the variables is shown in Figure S.6. We further assume the true coefficients are $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.5$ and $\beta_5 = \beta_6 = 2$. By definition, if we fix the intermediate coefficient for the root node γ_{11} to be zero, then the most sparse setting would be $\gamma_9 = 0.5$, $\gamma_{10} = 2$, and $\gamma_1 = \cdots = \gamma_8 = 0$, which we treat as the true values for $\boldsymbol{\gamma}$. Following Study I, we simulate the relative abundance data from a logistic Gaussian distribution and generate the response from a relative-shift model (with SNR being 1). We further truncate the compositional data to get excessive zeros, and use them as input data with measurement errors.

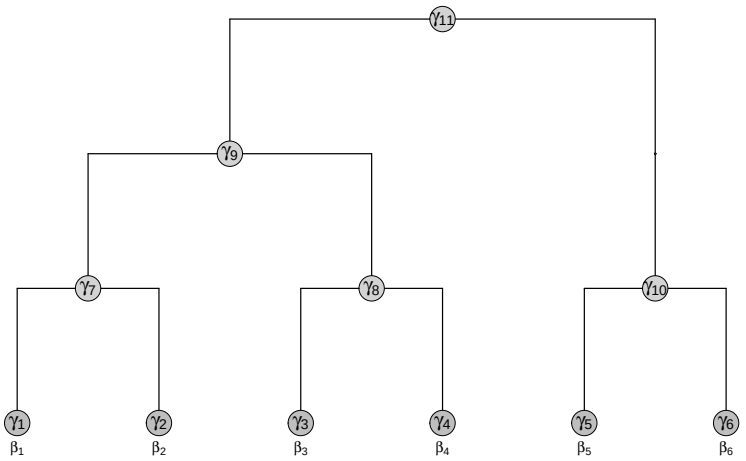


Figure S.6: The taxonomic tree structure among variables.

The comparisons of RS-L1, RS-CL2, and RS-DL2 are shown in Figures S.7 and S.8. In particular, Figure S.7 shows the box-plots of intermediate coefficient estimates from different methods. Both RS-DL2 and RS-CL2 are slightly better than RS-L1 in obtaining sparse estimation of the intermediate coefficients γ_1 to γ_8. This is likely due to the (overlapping) group structure

of the penalty terms in *Descendant* ℓ_2 and *Child* ℓ_2. Figure S.8 further shows the comparison of different methods in $\boldsymbol{\beta}$ coefficient estimation, prediction, and computing time. For $\boldsymbol{\beta}$ estimation, RS-DL2 has the smallest mean squared error, followed by RS-CL2. The difference between each pair of methods is not dramatic, but is statistically significant from a paired t-test at the nominal significance level of 0.05. The prediction performances of different methods are similar. The small differences in estimation accuracy may be obscured by the measurement errors in the input data. For computing time, RS-L1 is the best, likely because this is a rather small-scale setting. Overall, all three methods provide similar results, and the group penalties in *Descendant* ℓ_2 and *Child* ℓ_2 may potentially lead to more accurate coefficient estimate.

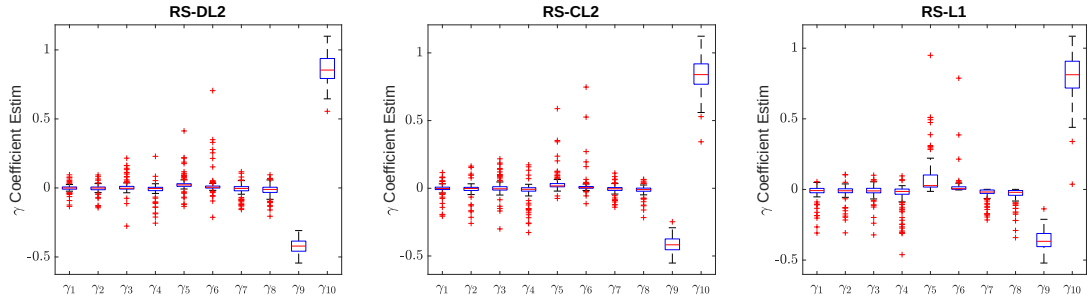


Figure S.7: The box-plots of intermediate coefficient estimates from RS-DL2, RS-CL2, and RS-L1.

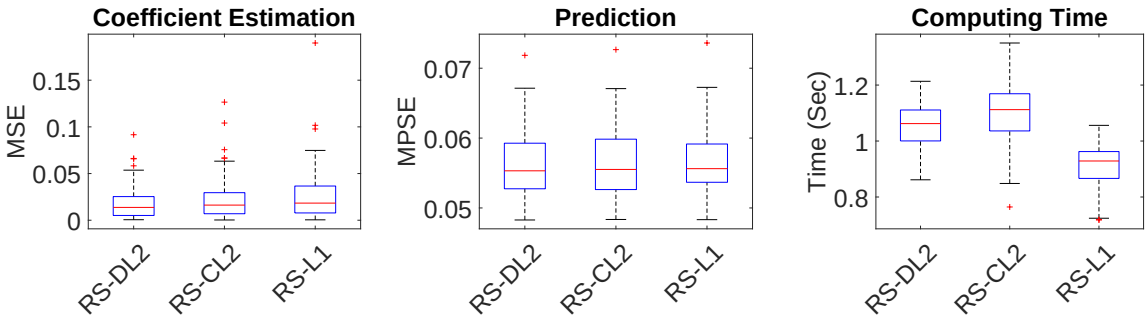


Figure S.8: The box-plots of $\boldsymbol{\beta}$ estimation, prediction, and computing time of RS-DL2, RS-CL2, and RS-L1.

B Proofs of the Theoretical Results

Proof of Theorem 1. Let $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ be a solution to (9). We have that

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \lambda\mathcal{P}_T(\widehat{\boldsymbol{\gamma}}) \leq \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\mathcal{P}_T(\boldsymbol{\gamma}),$$

for any $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ such that $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}$. Recall that $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ are the coefficient vectors of the true model satisfying $\boldsymbol{\beta}^* = \boldsymbol{A}\boldsymbol{\gamma}^*$. Let $\widehat{\boldsymbol{\Delta}}^{\gamma} = \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$, and $\widehat{\boldsymbol{\Delta}}^{\beta} = \boldsymbol{A}\widehat{\boldsymbol{\Delta}}^{\gamma} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. By using $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, we get

$$\begin{aligned}\frac{1}{2n}\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|^2 &\leq \lambda\mathcal{P}_T(\boldsymbol{\gamma}^*) - \lambda\mathcal{P}_T(\widehat{\boldsymbol{\gamma}}) + \frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\widehat{\boldsymbol{\Delta}}^{\beta} \\ &= \lambda(\mathcal{P}_T(\boldsymbol{\gamma}^*) - \mathcal{P}_T(\widehat{\boldsymbol{\gamma}})) + \frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\boldsymbol{A}\widehat{\boldsymbol{\Delta}}^{\gamma}. \end{aligned} \tag{S.1}$$

We mainly consider the group penalty forms *Child* ℓ_2 and *Descendant* ℓ_2 in (5) and (6) of the main paper, as the results for the *Node* ℓ_1 penalty in (4) of the main paper can be directly derived from Yan and Bien (2021). For each node $u \in I(T)$, define $\boldsymbol{P}_u$ as a $\mathbb{R}^{(|T|-1)\times(|T|-1)}$ diagonal matrix indicating its child nodes, such that the diagonal elements are given by

$$(\boldsymbol{P}_u)_{vv} = \begin{cases} 1 & v \in \text{Child(u)}; \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, for each node $u \in I(T)$, define $\boldsymbol{M}_u$ as a $\mathbb{R}^{(|T|-1)\times(|T|-1)}$ diagonal matrix indicating the nodes of its descendants, such that the diagonal elements are given by

$$(\boldsymbol{M}_u)_{vv} = \begin{cases} 1 & v \in \text{Descendant(u)}; \\ 0 & \text{otherwise.} \end{cases}$$

Then, the *Child* ℓ_2 penalty in (5) of the main paper can be re-expressed as

$$\mathcal{P}_T(\boldsymbol{\gamma}) = \sum_{u\in I(T)} \|\boldsymbol{P}_u\boldsymbol{\gamma}\|,$$

and the *Descendant* ℓ_2 penalty in (6) of the main paper can be re-expressed as

$$\mathcal{P}_T(\boldsymbol{\gamma}) = \sum_{u\in I(T)} \|\boldsymbol{M}_u\boldsymbol{\gamma}\|.$$

Moreover, the following two identities hold

$$\sum_{u\in I(T)} \boldsymbol{P}_u\boldsymbol{P}_u = \boldsymbol{I}_{|T|-1}, \qquad \sum_{u\in I(T)} \boldsymbol{P}_u\boldsymbol{M}_u = \boldsymbol{I}_{|T|-1}. \tag{S.2}$$

With the above treatment of the penalty forms, we are now ready to treat the second term in (S.1). Let's focus on the *Child* ℓ_2 penalty first.

$$\begin{aligned}
|\boldsymbol{\varepsilon}^T\boldsymbol{X}\boldsymbol{A}\widehat{\boldsymbol{\Delta}}^{\gamma}| &= |\boldsymbol{\varepsilon}^T\boldsymbol{X}\boldsymbol{A}(\sum_{u\in I(T)}\boldsymbol{P}_u\boldsymbol{P}_u)\widehat{\boldsymbol{\Delta}}^{\gamma}| \\
&= |\sum_{u\in I(T)}(\boldsymbol{\varepsilon}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{P}_u)(\boldsymbol{P}_u\widehat{\boldsymbol{\Delta}}^{\gamma})| \\
&\le \sum_{u\in I(T)}\|\boldsymbol{\varepsilon}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{P}_u\|\|\boldsymbol{P}_u\widehat{\boldsymbol{\Delta}}^{\gamma}\| \\
&\le \max_{u\in I(T)}\{\|\boldsymbol{\varepsilon}^T\boldsymbol{X}\boldsymbol{A}\boldsymbol{P}_u\|\}\sum_{u\in I(T)}\|\boldsymbol{P}_u\widehat{\boldsymbol{\Delta}}^{\gamma}\| \\
&\le \max_{u\in I(T)}\{\|(\boldsymbol{X}\boldsymbol{A}\boldsymbol{P}_u)^T\boldsymbol{\varepsilon}\|\}\{\sum_{u\in I(T)}\|\boldsymbol{P}_u\widehat{\boldsymbol{\gamma}}\| + \sum_{u\in I(T)}\|\boldsymbol{P}_u\boldsymbol{\gamma}^*\|\}.
\end{aligned}$$

That is,

$$|\frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\boldsymbol{A}\widehat{\boldsymbol{\Delta}}^{\gamma}| \le \max_{u\in I(T)}\{\|\frac{1}{n}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{P}_u)^T\boldsymbol{\varepsilon}\|\}(\mathcal{P}_T(\boldsymbol{\gamma}^*) + \mathcal{P}_T(\widehat{\boldsymbol{\gamma}})). \tag{S.3}$$

It is easy to see that the same inequality holds for the *Descendant* ℓ_2 penalty due to the second identity in (S.2).

Now we bound the stochastic term $\max_{u\in I(T)}\{\|(\boldsymbol{X}\boldsymbol{A}\boldsymbol{P}_u)^T\boldsymbol{\varepsilon}\|/n\}$. We will need the following results from Proposition 1 of Hsu et al. (2012).

Lemma S.1. *Let $\boldsymbol{Z}$ be an $m\times n$ matrix and let $\Sigma_z = \boldsymbol{Z}^T\boldsymbol{Z}$. Suppose $\boldsymbol{\varepsilon} = (\epsilon_1,\ldots,\epsilon_n)^T$ is a multivariate Gaussian random with mean zero and covariance $\sigma^2\boldsymbol{I}$. For all $t>0$,*

$$pr(\|\boldsymbol{Z}\boldsymbol{\varepsilon}\|^2 > \sigma^2\{\,Tr(\Sigma_z) + 2\sqrt{Tr(\Sigma_z^2)t} + 2\|\Sigma_z\|t\}) < e^{-t},$$

where $\|\Sigma\|$ denotes the spectral norm of the non-negative definite matrix Σ.

For each $u\in I(T)$, denote $\boldsymbol{X}_u\in\mathbb{R}^{n\times p_u}$ be the submatrix of $\boldsymbol{X}\boldsymbol{A}\boldsymbol{P}_u$ by keeping its p_u columns corresponding to the children of the node u; the rest of the columns in $\boldsymbol{X}\boldsymbol{A}\boldsymbol{P}_u$ are all zero

columns. It can be recognized that each column of $\boldsymbol{X}_u$ is obtained by aggregating the columns of $\boldsymbol{X}$ corresponding to the leaf nodes $L(T_v)$, $v \in$ Child(u). For example, in Figure 1 of the main paper, for the internal node γ_{10}, its $\boldsymbol{X}_u$ matrix contains 3 columns, $X_1 + X_2, X_3, X_4$. Since $\boldsymbol{X}$ is compositional, it follows that each element of the matrix $\boldsymbol{X}_u$ is in the interval $[0, 1]$, and the sum of each of its rows is less than or equal to 1. Write $\boldsymbol{X}_u = \{x_{ij}^u\}_{n \times p_u}$, it follows that

$$\begin{aligned}
&\mathrm{Tr}(\Sigma_u) = \sum_i^n \sum_j^{p_u} (x_{ij}^u)^2 \leq \sum_i^n (\sum_j^{p_u} x_{ij}^u)^2 \leq n; \\
&\mathrm{Tr}(\Sigma_u^2) \leq \mathrm{Tr}(\Sigma_u)^2 \leq n^2; \\
&\|\Sigma_u\| \leq \mathrm{Tr}(\Sigma_u) \leq n.
\end{aligned}$$

By Lemma S.1, we have that for any $u \in I(T)$ and any $t > 0$,

$$pr\{\|\boldsymbol{X}_u^T \boldsymbol{\varepsilon}\|^2 > n\sigma^2(1 + 2\sqrt{t} + 2t)\} < e^{-t}.$$

It follows that

$$pr\{\frac{1}{\sqrt{n}}\|\boldsymbol{X}_u^T \boldsymbol{\varepsilon}\| > 2\sqrt{2}\sigma\sqrt{t}\} < e^{-t} \text{ when } t > 1/2.$$

Upon taking $t = \log|I(T)|/\delta > 1/2$, we have

$$pr(\frac{1}{\sqrt{n}}\|\boldsymbol{X}_u^T \boldsymbol{\varepsilon}\| > 2\sqrt{2}\sigma\sqrt{\log(|I(T)|)/\delta}) < e^{-\log\frac{|I(T)|}{\delta}} = \frac{\delta}{|I(T)|}$$

By taking a union bound over all internal nodes, we get

$$pr(\max_{u \in I(T)} \frac{1}{\sqrt{n}}\|\boldsymbol{X}_u^T \boldsymbol{\varepsilon}\| > 2\sqrt{2}\sigma\sqrt{\log(|I(T)|)/\delta}) \leq \delta.$$

By (S.1) and (S.3), if we take $\lambda \geq 2\sqrt{2}\sigma\sqrt{\log(|I(T)|)/(\delta n)}$, then with probability at least $1 - \delta$,

$$\begin{aligned}
\frac{1}{2n}\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|^2 &\leq \lambda(\mathcal{P}_T(\boldsymbol{\gamma}^*) - \mathcal{P}_T(\widehat{\boldsymbol{\gamma}})) + \lambda(\mathcal{P}_T(\boldsymbol{\gamma}^*) + \mathcal{P}_T(\widehat{\boldsymbol{\gamma}})) \\
&= 2\lambda\mathcal{P}_T(\boldsymbol{\gamma}^*).
\end{aligned}$$

The results then follow because the above holds for any $\boldsymbol{\gamma}^*$ such that $\boldsymbol{\beta}^* = \boldsymbol{A}\boldsymbol{\gamma}^*$. This completes the proof.

□

Proof of Lemma 1. For the *Node* ℓ_1 penalty in (4) of the main paper, the result directly follows based on the construction of the coarsest aggregating set B^* and a corresponding choice of $\boldsymbol{\gamma}^*$, i.e., for each $u \in B^*$, $\gamma_u^* = \beta_j^*$ for any $j \in L(T_u)$, and otherwise $\gamma_u^* = 0$. It is easily seen that this $\boldsymbol{\gamma}^*$ satisfies $\boldsymbol{A}\boldsymbol{\gamma}^* = \boldsymbol{\beta}^*$. It then follows that $\sum_{u \in T_{-r}} |\gamma_u| \leq M|B^*|$.

For the *Child* ℓ_2 penalty, consider again the above construction of $\boldsymbol{\gamma}^*$ according to the coarsest aggregating set B^*. Immediately, we have that

$$\|\boldsymbol{\gamma}^*\| \leq M\sqrt{|B^*|}.$$

Define

$$B_p^* = \{u \in I(T); \text{Child(u)} \cap B^* \neq \emptyset.\}.$$

That is, B_p^* collects the parent nodes of all the nodes in B^*. We have that $\|(\gamma_v)_{v \in \text{Child(u)}}\| \neq 0$ only if $u \in B_p^*$. Since there is no overlap among Child(u), $u \in B_p^*$, we also have that $|B_p^*| \leq |B^*|$. It then follows that

$$\sum_{u \in I(T)} \|(\gamma_v)_{v \in \text{Child(u)}}\| = \sum_{u \in B_p^*} \|(\gamma_v)_{v \in \text{Child(u)}}\| \leq \sqrt{|B_p^*|}\|\boldsymbol{\gamma}^*\| \leq M\sqrt{|B_p^*||B^*|} \leq M|B^*|.$$

Here we use the fact that for any partition of a vector, i.e., $\boldsymbol{a} = (\boldsymbol{a}_1^T, \ldots, \boldsymbol{a}_s^T)^T$, it holds that $\sum_{i=1}^s \|\boldsymbol{a}_i\| \leq \sqrt{s}\|\boldsymbol{a}\|$. This completes the proof.

□

References

Hsu, D., Kakade, S., Zhang, T., et al. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability* **17,**.

Yan, X. and Bien, J. (2021). Rare feature selection in high dimensions. *Journal of the American Statistical Association* **116,** 887–900.