Regression with Supervised Log-Ratio

Jing Ma

Fred Hutchinson Cancer Research Center
Public Health Sciences Division

February 16, 2021

Goal

To predict a health outcome from microbiome data.

To select individual or groups of OTUs that are predictive of a health outcome.

Data

Microbiome data are compositional: $X = [X_1, \ldots, X_p] \in \mathbb{R}^p$

$$X_1 + \ldots X_p = 1.$$

Such data are also called relative abundances.

Data

Microbiome data are compositional: $X = [X_1, \ldots, X_p] \in \mathbb{R}^p$

$$X_1 + \ldots X_p = 1.$$

Such data are also called relative abundances.

If the absolute abundances are $W = [W_1, \ldots, W_p]$, and

$$X_j = \frac{W_j}{\sum_{k=1}^{p} W_k},$$

knowing only X indicates a loss of scale (i.e. the total). This motivates the need for methods that are scale invariant.

Additive Log-Ratio

Ratios circumvent the limitation of not knowing the absolute abundances in microbiome studies. For example,

$$\text{alr}(X) = \left(\log \frac{X_1}{X_p}, \ldots, \log \frac{X_{p-1}}{X_p} \right).$$

is scale-invariant.

Regression with ALR:

$$y = \sum_{j=1}^{p-1} \beta_j \log \frac{X_j}{X_p} + \varepsilon.$$

- ALR requires a reference. It can be difficult to know which reference to use.

Compositional Lasso

Compositional Lasso translates the ALR regression into a constraint-based model and removes the need for a reference.

$$
\begin{aligned}
y =& \sum_{j=1}^{p-1} \beta_j \log X_j - \left(\sum_{j=1}^{p-1} \beta_j \right) \log X_p + \varepsilon \\
=& \beta^\mathsf{T} \log X + \varepsilon,
\end{aligned}
$$

subject to $1^\mathsf{T} \beta = 0$.

- Method in Lin et al. (14) is not scale-invariant w.r.t. any subgroups.
- Refinements in Shi et al. (16) require knowledge of subgroups in order to meet the additional scale-invariance requirement.

Log-Ratio Regression

What about including all possible log-ratios? Bates and Tibshirani (19) proposed

$$y = \sum_{1 \le j < k \le p} \theta_{j,k} \log \frac{X_j}{X_k} + \varepsilon.$$

- This model is not identifiable, because the predictors are perfectly co-linear, e.g. $\log X_1/X_2, \log X_1/X_3, \log X_2/X_3$.

Balances

Rivera-Pinto et al. (18) proposed `selbal` to iteratively select subsets (A, B) such that

$$y = \beta_0 + \beta_1 \log \frac{g(X_A)}{g(X_B)} + \varepsilon$$

where $g(X)$ is the geometric mean function.

- The method starts with A, B both being of size 1 each and gradually add more parts to each subset.
- This method is computationally intense.

Balances from Orthonormal Bases

To reduce the search space for balances, we consider an orthonormal basis of the simplex $\mathcal{S}^p$ derived from a bifurcating tree.

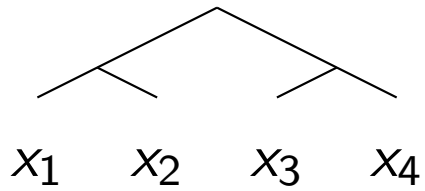


Figure 1: A simple bifurcating tree of four variables (x_1, x_2, x_3, x_4).

For each internal node of the tree, one can define a balance as the log-ratio between the geometric means of the two branches.

$$\log \frac{(x_1 x_2)^{1/2}}{(x_3 x_4)^{1/2}}, \quad \frac{1}{\sqrt{2}} \log \frac{x_1}{x_2}, \quad \frac{1}{\sqrt{2}} \log \frac{x_3}{x_4}.$$

Balances from Orthonormal Bases

Here is a more mathematical interpretation:

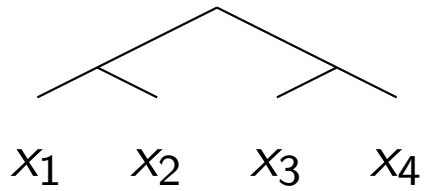


Figure 2: A simple bifurcating tree of four variables (x_1, x_2, x_3, x_4).

The internal nodes on the bifurcating tree in Figure 2 define a set of orthogonal vectors

$$\boldsymbol{u}_1^\intercal = (\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}), \quad \boldsymbol{u}_2^\intercal = \sqrt{\frac{1}{2}}(1, -1, 0, 0), \quad \boldsymbol{u}_3^\intercal = \sqrt{\frac{1}{2}}(0, 0, 1, -1),$$

such that $\text{ilr}(x) = (\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3)^\intercal \log(x)$.

Regression with ILR

Let $\boldsymbol{U}^\mathsf{T} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{p-1}) \in \mathbb{R}^{p \times (p-1)}$ denote the orthogonal basis in the Euclidean space associated with a sequential binary partition of the simplex $\mathcal{S}^p$. We have $\text{ilr}(X) = \boldsymbol{U} \log(X)$.

$$y = (\theta^{\text{ilr}})^\mathsf{T} \text{ilr}(X) + \varepsilon,$$

and $\beta = \boldsymbol{U}^\mathsf{T} \theta^{\text{ilr}}$. One can verify that $\sum_j \beta_j = 0$.

Given data $\{y_i, X_i\}_{i=1}^n$, one can impose a structured penalty to solve for θ^{ilr} (and β)

$$\min \sum_{i=1}^{n} \{y_i - (\theta^{\text{ilr}})^\mathsf{T} \text{ilr}(X_i)\}^2 + P_\lambda(\theta^{ilr}).$$

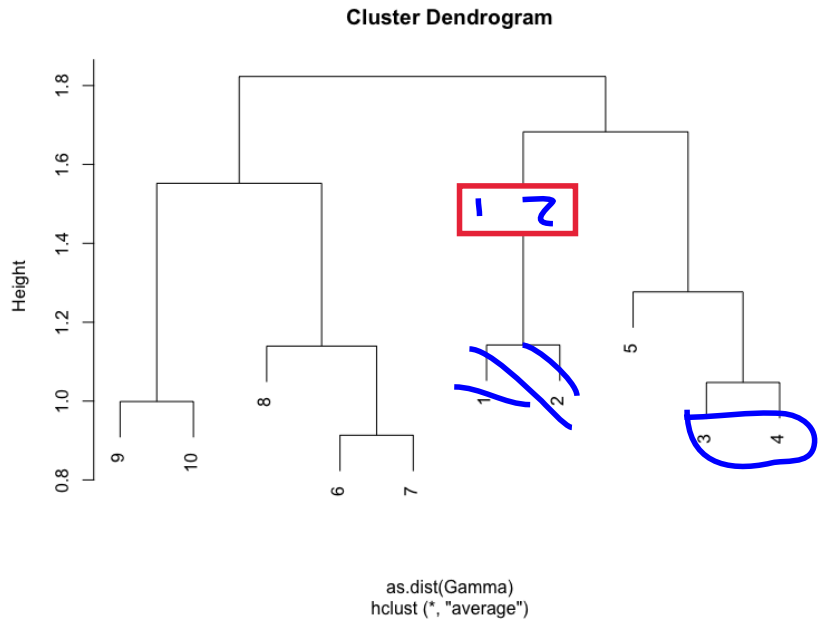
Regression with ILR

How to define the tree?

Regression with ILR

How to define the tree?

Want a tree that leads to easy-to-interpret balances being selected.



$\log \frac{X_1}{X_2}$ is easier to interpret than $\log \frac{(X_1 X_2)^{1/2}}{(X_3 X_4 X_5)^{1/3}}$, because the former consists of only two taxa.

Regression with ILR

How to define the tree?

Regression with ILR

How to define the tree?

Quinn and Erb (2020) suggested hierarchical clustering of the Aitchison variation matrix

$$T_{j,k} = \text{Var}(\log \frac{X_j}{X_k})$$

and choosing the balances that consist of 2 or 3 taxa. But these are not necessarily the most predictive!

Regression with ILR

How to define the tree?

Quinn and Erb (2020) suggested hierarchical clustering of the Aitchison variation matrix

$$T_{j,k} = \text{Var}(\log \frac{X_j}{X_k})$$

and choosing the balances that consist of 2 or 3 taxa. But these are not necessarily the most predictive!

This motivates us to define a tree whose leaf nodes are the most predictive!

Supervised Tree

- For each pair of taxa (j, k), compute the correlation between the log-ratio $Z_{j,k} = \log \frac{X_j}{X_k}$ and the outcome y:

$$\gamma_{j,k} = \text{Cor}(Z_{j,k}, y).$$

- Form a distance matrix using $(1 - |\gamma_{j,k}|)$.
- Apply hierarchical clustering with average linkage to the distance in the previous step.
- Use balances obtained from the hierarchical clustering tree to predict the outcome.

Regression with ILR

What penalty to use? In practice, scientists might be interested in signatures such as the ratio between two groups of taxa, and/or individual taxa.

To achieve simultaneous selection of balances and taxa, one approach is to impose sparsity on both θ and β via

$$\min \sum_{i=1}^{n} \{y_i - \theta^{\mathsf{T}} \mathrm{ilr}(X_i)\}^2 + \lambda \left(\alpha \sum_{j=1}^{p-1} w_j |\theta_j| + (1-\alpha) \sum_{j=1}^{p} \eta_j |\beta_j| \right),$$

subject to $\beta = \boldsymbol{U}^{\mathsf{T}} \theta$. This can be solved using the algorithm in Bien et al. (20). Note the difference between our framework and the method by Bien et al. (20).

Summary of Models

$$\mathbb{E}[y \mid X] = \begin{cases} \beta^{\mathsf{T}} \log X, & 1^{\mathsf{T}}\beta = 0 \\ \displaystyle\sum_{1 \le j < k \le p} \theta^{lr}_{j,k} \log \frac{X_j}{X_k} & \\ \beta_0 + \beta_1 \log \dfrac{g(X_A)}{g(X_B)} & \\ (\theta^{\mathsf{ilr}})^{\mathsf{T}} \mathsf{ilr}(X). & \end{cases}$$

Simulations I

- Generate $W_i \sim N(\mu, \Sigma)$. The first five entries in μ are $\log(p)$, with all remaining ones being 0. Entries in the covariance matrix satisfy $\Sigma_{j,k} = \rho^{|j-k|}$ with $\rho = 0.2$ or 0.5.
- Compute $X_i = W_{ij} / \sum_k W_{ik}$
- $\beta = (1, 0.4, 1.2, -1.5, -0.8, -0.3, 0, \ldots, 0)$ with the first 6 entries being nonzero.
- $y_i = \beta^\intercal \log(X_i) + \varepsilon_i$ with $\varepsilon_i \sim N(0, 0.5^2)$.
- The sample size $n = 100$ and number of predictors $p = 200$.

Simulations II

- Generate $W_i \sim N(\mu, \Sigma)$. The first five entries in μ are $\log(p)$, with all remaining ones being 0. Entries in the covariance matrix satisfy $\Sigma_{j,k} = \rho^{|j-k|}$ with $\rho = 0.2$ or 0.5.
- Compute $X_i = W_{ij} / \sum_k W_{ik}$
- Compute Aitchison variation and cluster $T_{j,k} = \text{Var}(\log \frac{X_j}{X_k})$
- Derive the transformation matrix $\boldsymbol{U}$ and $\text{ilr}(X_i)$
- Generate sparse $\theta \in \mathbb{R}^{p-1}$ with (1) $\beta = \boldsymbol{U}^\top \theta$ is sparse and (2) β is not sparse.
- $y_i = \theta^\top \text{ilr}(X_i) + \varepsilon_i$ with $\varepsilon_i \sim N(0, 0.5^2)$.
- The sample size $n = 100$ and number of predictors $p = 200$.

Model Fitting

Apply cross-validation to select the tuning parameters.

Apply supervised log-ratio with the tree obtained from (1) clustering the Aitchison variation and (2) clustering the 1-correlation distance.

Model Evaluation

FRED HUTCH
CURES START HERE™

Mean squared prediction error and variable selection performance using ROC curve. See Buhlmann et al. (2013)
https://arxiv.org/pdf/1209.5908.pdf