Variable selection in microbiome compositional data analysis

Antoni Susin, Yiwen Wang, Kim-Anh Lê Cao, M.Luz Calle

SUPPLEMENTAL MATERIAL

We considered two different schemes for generating the dependent variable Y:

1. Log-contrast method:

We defined the values of Y (1 or 0) according to the probability of $Y = 1$ given by a logistic model with a log-contrast function of the true associated taxa $(X_1, \dots X_{k_1})$ given by

$$P(Y = 1) = 1/(1 + \exp(-(S - \bar{S})))$$

where $S = \sum_{j=1}^{k_1} b_j \log X_j$ with the constraint $\sum_{j=1}^{k_1} b_j = 0$

Table of b_j coefficients

Dataset	number associated (k1)	number not associated (k2)	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
Crohn	3	10	1	-0.5	-0.5							
	3	20	1	-0.5	-0.5							
	3	30	1	-0.5	-0.5							
	3	40	1	-0.5	-0.5							
	5	10	3	3	-2	-2	-2					
	5	20	3	3	-2	-2	-2					
	5	30	3	3	-2	-2	-2					
	5	40	3	3	-2	-2	-2					
	10	10	1	2	3	4	5	-1	-2	-3	-4	-5
	10	20	1	2	3	4	5	-1	-2	-3	-4	-5
	10	30	1	2	3	4	5	-1	-2	-3	-4	-5
	10	40	1	2	3	4	5	-1	-2	-3	-4	-5
HFHS	3	100	1	-0.5	-0.5							
	3	200	1	-0.5	-0.5							
	3	300	1	-0.5	-0.5							
	3	400	1	-0.5	-0.5							
	5	100	3	3	-2	-2	-2					
	5	200	3	3	-2	-2	-2					
	5	300	3	3	-2	-2	-2					
	5	400	3	3	-2	-2	-2					
	10	100	1	2	3	4	5	-1	-2	-3	-4	-5
	10	200	1	2	3	4	5	-1	-2	-3	-4	-5
	10	300	1	2	3	4	5	-1	-2	-3	-4	-5
	10	400	1	2	3	4	5	-1	-2	-3	-4	-5

2. K-means method:

We computed the Aitchison distance of the samples restricted to the first k_1 columns of **X** (associated taxa) and performed the K-means clustering method with K=2. This process divided the samples into two groups according to their similar profile with respect to the associated taxa. We defined $Y = 0$ for the individuals in one cluster and $Y = 1$ for the individuals in the other.