Supervised Learning and Model Analysis with Compositional Data

Shimeng Huang, Elisabeth Ailer, Niki Kilbertus, Niklas Pfister

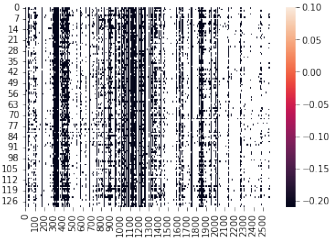
ICSDS2022
Florence, Italy
Dec. 16, 2022

UNIVERSITY OF
COPENHAGEN

Introduction

Motivation

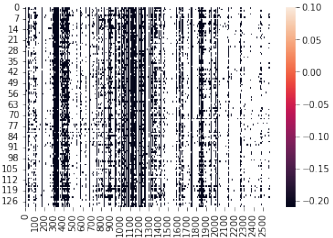
Consider having a matrix of relative abundance of microbiota and an outcome vector of whether a person is cirrhotic or not.

$$X = \quad \text{  } \quad y = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Motivation

Consider having a matrix of relative abundance of microbiota and an outcome vector of whether a person is cirrhotic or not.

$$X = \quad y = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$



Q: Which microbiota are important predictors?

Why is this difficult?

1. **Compositionality**
   Data points live on a simplex $\{x \in [0,1]^p \mid \sum x^j = 1\}$
   $\rightarrow$ ignoring this results in spurious correlations
2. **Non-linear effect**
   Complex underlying structure and interaction between microbiota
   $\rightarrow$ assuming linearity is unlikely to be realistic

Classical approach

Log-contrast model (Aitchison and Bacon-Shone, 1984):

$$Y = \sum_{j=1}^{p} \beta_j \log(X^j) + \epsilon \quad \text{and} \quad \sum_{j=1}^{p} \beta_j = 0,$$

Classical approach

Log-contrast model (Aitchison and Bacon-Shone, 1984):

$$Y = \sum_{j=1}^{p} \beta_j \log(X^j) + \epsilon \quad \text{and} \quad \sum_{j=1}^{p} \beta_j = 0,$$

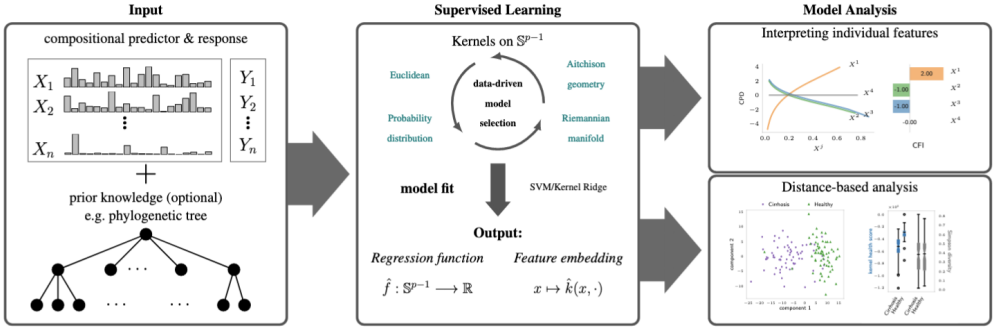
Although the log-contrast model is easy to fit and interpretable, it cannot:

- Incorporate zeros
- Complex signals (e.g. interactions)
- Include prior knowledge on the relation between components of X

Our proposal

Our proposal: `KernelBiome`

`KernelBiome`: A kernel-based nonparametric regression and classification framework for compositional data.



Supervised learning with kernels

We aim to estimate the conditional mean of Y:

$$f^* : x \mapsto \mathbb{E}[Y \mid X = x]$$

where we assume that $f^* \in \mathcal{F} \subseteq \{f | f : \mathbb{S}^{p-1} \to \mathbb{R}\}$.

Supervised learning with kernels

We aim to estimate the conditional mean of Y:

$$f^* : x \mapsto \mathbb{E}[Y \mid X = x]$$

where we assume that $f^* \in \mathcal{F} \subseteq \{f | f : \mathbb{S}^{p-1} \to \mathbb{R}\}$.

Why kernels?

- Targeted to the simplex (probability distribution, heat diffusion, Aitchision, and Riemannian manifold kernels)
- Able to capture complex signals (characteristic kernels)
- Suitable for proposed model analysis methods (differentiable $\hat{f}$)
- Provides an embedding $\hat{k}(x_i, \cdot)$ which can be used for post-analysis (not in this talk)

Interpreting individual components

Feature importance measures for complex models which do not respect the simplex structure can give misleading results.

$$f_1 : x \mapsto 10x^1 + 10x^2$$

$$f_2 : x \mapsto \frac{1 - x^2 - x^3}{1 - x^3}$$

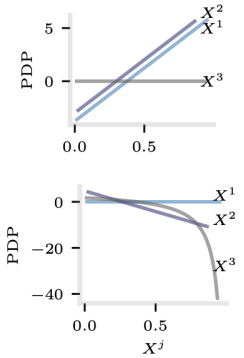
Interpreting individual components

Feature importance measures for complex models which do not respect the simplex structure can give misleading results.

$$f_1 : x \mapsto 10x^1 + 10x^2$$

$$f_2 : x \mapsto \frac{1 - x^2 - x^3}{1 - x^3}$$

	f_1			f_2		
	x^1	x^2	x^3	x^1	x^2	x^3
FI	3.76	2.99	**0.00**	0.00	-4.72	**-4.40**
PI	11.35	5.87	**0.00**	0.00	34.90	**25.39**



Top: f_1. Bottom: f_2.

Interpreting individual components

Thinking “causal” — what are some reasonable interventions on the simplex?

Interpreting individual components

Thinking “causal” — what are some reasonable interventions on the simplex?

1. **Multiplying a component**
   For any $j \in \{1, \ldots, p\}$, $x \in \mathbb{S}^{p-1}$ and $c > 0$, define

$$\psi_j(x, c) := s_c(x^1, \cdots, x^{j-1}, cx^j, x^{j+1}, \cdots, x^p) \in \mathbb{S}^{p-1},$$

   where $s_c = 1/(\sum_{\ell \neq j}^{p} x^\ell + cx^j)$.

Interpreting individual components

Thinking “causal” — what are some reasonable interventions on the simplex?

1. **Multiplying a component**
   For any $j \in \{1, \dots, p\}$, $x \in \mathbb{S}^{p-1}$ and $c > 0$, define

$$\psi_j(x, c) := s_c(x^1, \cdots, x^{j-1}, cx^j, x^{j+1}, \cdots, x^p) \in \mathbb{S}^{p-1},$$

   where $s_c = 1/(\sum_{\ell \neq j}^{p} x^\ell + cx^j)$.

2. **Setting a component**
   For any $j \in \{1, \dots, p\}$, $x \in \mathbb{S}^{p-1}$ with $\sum_{\ell \neq j}^{p} x^\ell > 0$ and $c \in [0, 1]$, define the intervened composition by

$$\phi_j(x, c) := (sx^1, \cdots, sx^{j-1}, c, sx^{j+1}, \cdots, sx^p) \in \mathbb{S}^{p-1},$$

   where $s = (1 - c)/(\sum_{\ell \neq j}^{p} x^\ell)$.

Interpreting individual components

Two measures for feature importance on the simplex based on the above:

1. **Compositional feature importance** (CFI)

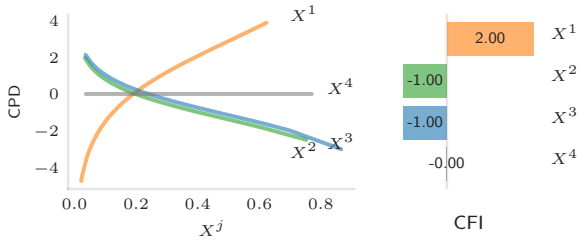
$$I_f^j := \mathbb{E}\left[\frac{\partial}{\partial c} f(\psi_j(X, c)) \mid_{c=1}\right]$$

2. **Compositional partial dependence** (CPD)

$$S_f^j : z \mapsto \mathbb{E}\big[f(\phi_j(X, z))\big] - \mathbb{E}\big[f(X)\big]$$

Example: log-contrast model

$$f : x \mapsto 2\log(x^1) - \log(x^2) - \log(x^3).$$



CFI and CPD using the true function f based on $n = 100$ i.i.d. compositional log-normal samples

Back to the previous example

$$f_1 : x \mapsto 10x^1 + 10x^2$$

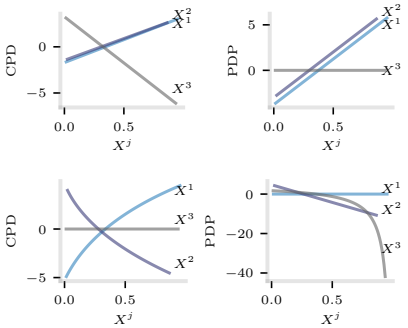
$$f_2 : x \mapsto \frac{1 - x^2 - x^3}{1 - x^3}$$

Back to the previous example

$$f_1 : x \mapsto 10x^1 + 10x^2$$

$$f_2 : x \mapsto \frac{1 - x^2 - x^3}{1 - x^3}$$

	f_1			f_2		
	x^1	x^2	x^3	x^1	x^2	x^3
CFI	0.85	0.87	**-1.72**	1.94	-1.94	**0.00**
FI	3.76	2.99	**0.00**	0.00	-4.72	**-4.40**
PI	11.35	5.87	**0.00**	0.00	34.90	**25.39**



Top: f_1. Bottom: f_2.

Conclusions

Conclusions

- Microbiome data analysis can be tricky
  - Misleading correlations, tricky interpretation
- Log-ratio approach solves some of the issues but applicability is limited
  - Zero-inflation, interactions, $\cdots$
- `KernelBiome`: a pipeline for prediction + post-analysis tageting the simplex
  - Non-parametric framework with interpretability
  - Can incorporate prior knowledge via weighting (omitted today)
  - Distance-based analysis (omitted today)
  - Competitive performance on public microbiome datasets (omitted today)
  - `pip install kernelbiome`

Scan for preprint ☺

References

J. Aitchison and J. Bacon-Shone. Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330, 1984. ISSN 00063444. doi:10.1093/biomet/71.2.323.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. doi:10.1214/aos/1013203451.

S. Huang, E. Ailer, N. Kilbertus, and N. Pfister. Supervised learning and model analysis with compositional data. *arXiv preprint*, 2022. doi:10.48550/ARXIV.2205.07271.