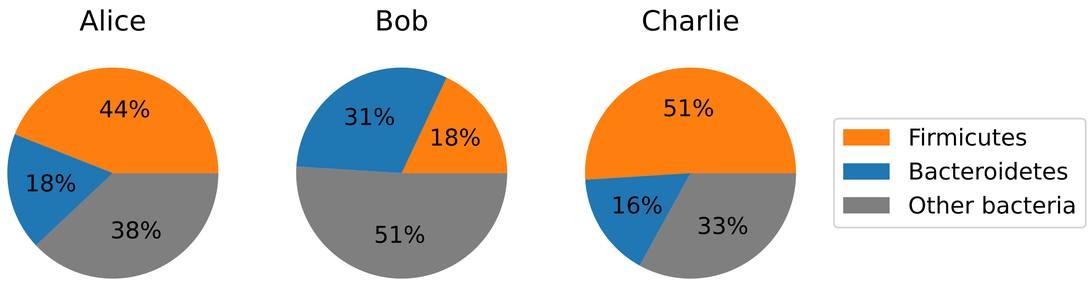
Data Augmentation for Compositional Data: Advancing Predictive Models of the Microbiome

Elliott Gordon-Rodriguez, Thomas P. Quinn, and John P. Cunningham

Columbia University

Compositional Data

Compositional Data (CoDa) describe the parts of a whole:

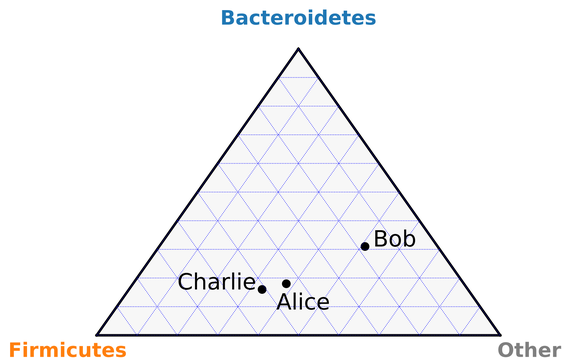


Equivalently, CoDa can be thought of as simplex-valued data:

$$\mathcal{D} = \{\mathbf{x}_i \in \Delta^K\}_{i=1}^n,$$

where Δ^K denotes the simplex:

$$\Delta^K = \left\{ \mathbf{x} \in \mathbb{R}_+^K : \sum_{k=1}^K x_k = 1 \right\}.$$



The Human Microbiome

Methods: Aitchison Mixup

Aitchison (1984) established the following Hilbert space structure on the simplex:

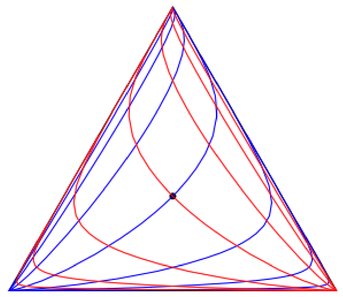
$$\mathbf{v} \oplus \mathbf{x} = \frac{1}{\sum_{j=1}^p v_j x_j}[v_1 x_1, \ldots, v_p x_p],$$

$$\lambda \odot \mathbf{x} = \frac{1}{\sum_{j=1}^p x_j^\lambda}[x_1^\lambda, \ldots, x_p^\lambda],$$

$$\langle \mathbf{v}, \mathbf{x} \rangle = \frac{1}{2p} \sum_{j=1}^p \sum_{k=1}^p \log\left(\frac{v_j}{v_k}\right) \log\left(\frac{x_j}{x_k}\right).$$

Given training points $\mathbf{x}_1$ and $\mathbf{x}_2$, and $\lambda \sim U(0,1)$, generate:

$$\mathbf{x}^{\text{aug}} := (\lambda \odot \mathbf{x}_1) \oplus ((1-\lambda) \odot \mathbf{x}_2)$$



Aitchison Geometry on Δ^2 (Orthogonal grid)

Experiments: Supervised Learning

Aitchison Mixup improves existing microbiome learning pipelines across 12 standard benchmarks.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
Crohn's (Ileum)	0.72	**0.79**	0.76	**0.79**	0.72	**0.74**	0.73	**0.79**	**0.74**	**0.74**
Crohn's (Rectum)	0.78	**0.82**	**0.81**	0.80	**0.80**	**0.80**	0.78	**0.83**	**0.74**	**0.74**
Gastrointestinal	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Female/Male	0.60	**0.64**	**0.57**	**0.57**	0.56	**0.58**	**0.58**	**0.58**	0.50	**0.51**
Stool/Tongue	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Plaque	0.81	**0.83**	0.82	**0.83**	**0.84**	0.83	0.78	**0.82**	0.75	**0.76**
Colorectal Cancer	**0.68**	0.67	0.67	**0.69**	0.73	**0.74**	0.63	**0.73**	**0.59**	0.54
Diabetes	0.62	**0.65**	0.66	**0.68**	0.64	**0.65**	0.45	**0.70**	**0.64**	**0.64**
Cirrhosis	**0.93**	**0.93**	0.94	**0.95**	0.92	**0.93**	0.84	**0.90**	0.76	**0.82**
Black/Hispanic	0.53	**0.60**	0.57	**0.61**	0.61	**0.62**	0.62	**0.63**	**0.63**	0.61
Nugent Score	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.96**	0.95
Black/White	0.55	**0.61**	0.58	**0.65**	**0.61**	**0.61**	**0.66**	0.65	0.58	**0.60**
Mean	0.77	**0.79**	0.78	**0.80**	0.78	**0.79**	0.75	**0.80**	**0.74**	**0.74**

Also in the paper but not shown here...

- Compositional Feature Dropout.
- Compositional CutMix.
- Contrastive Learning.

Acknowledgements

We thank Samuel Lippl, Richard Zemel, and the anonymous reviewers for helpful discussions, and the Simons Foundation 542963, Sloan Foundation, McKnight Endowment Fund, NSF DBI-1707398, and the Gatsby Charitable Foundation for support.