UBIQUITY OF
COMPOSITIONAL DATA

Compositional: Relating to parts of some whole

Proportions
Parts per million
Percentages

X+Y+Z=k
And all Positive

RELATIVE *DATA*

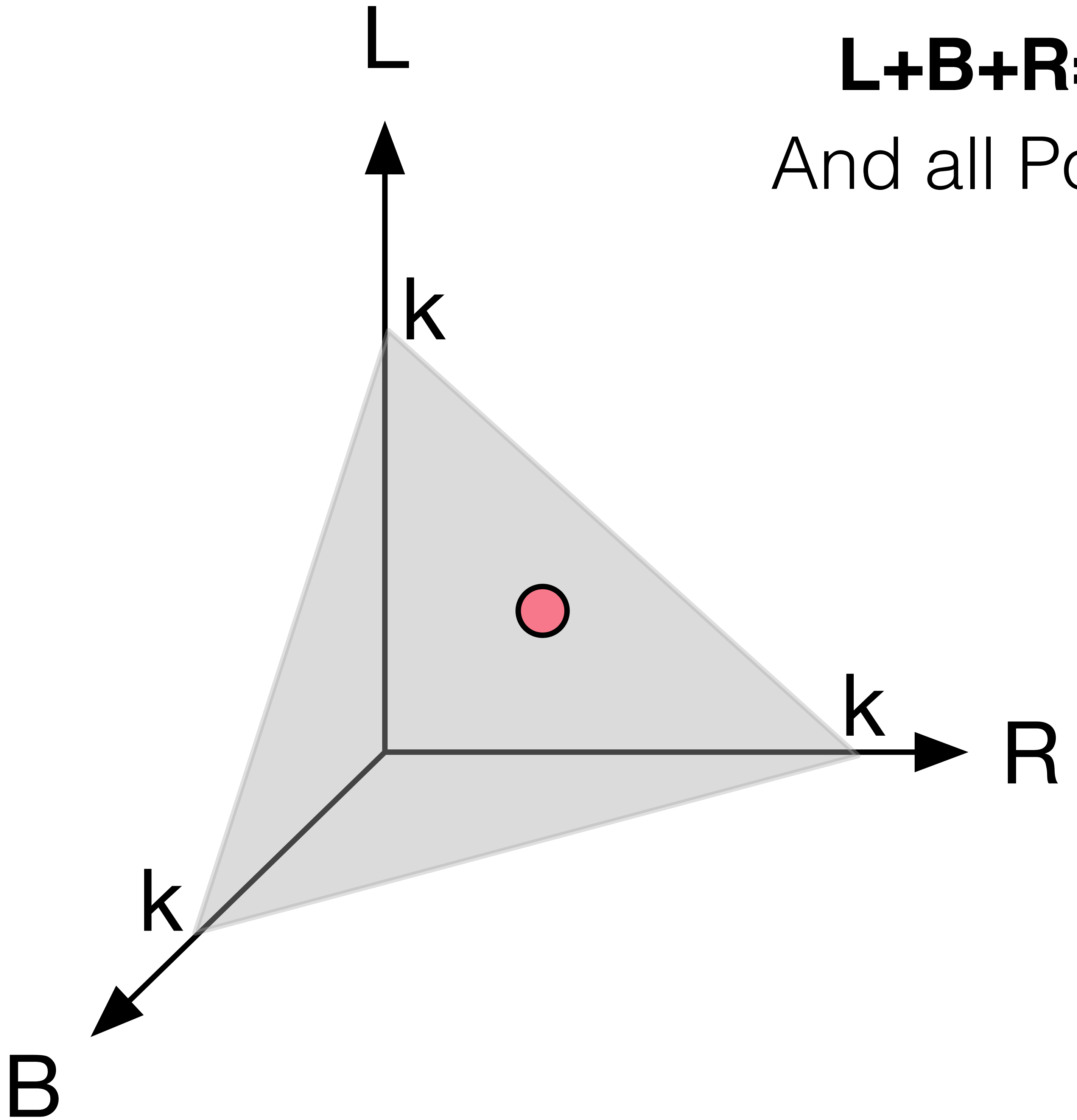
Simple Examples

- Does hongite have more calcium than struvite? (*e.g., parts per million*)
- Have I been spending more of my day in the bathroom since I ate that sandwich? (*e.g., percentage of your day*)
- Does my cow produce higher protein milk when I feed her that sandwich? (*e.g.,* proportion of calories from protein)

COMPOSITONAL
SIMPLEX

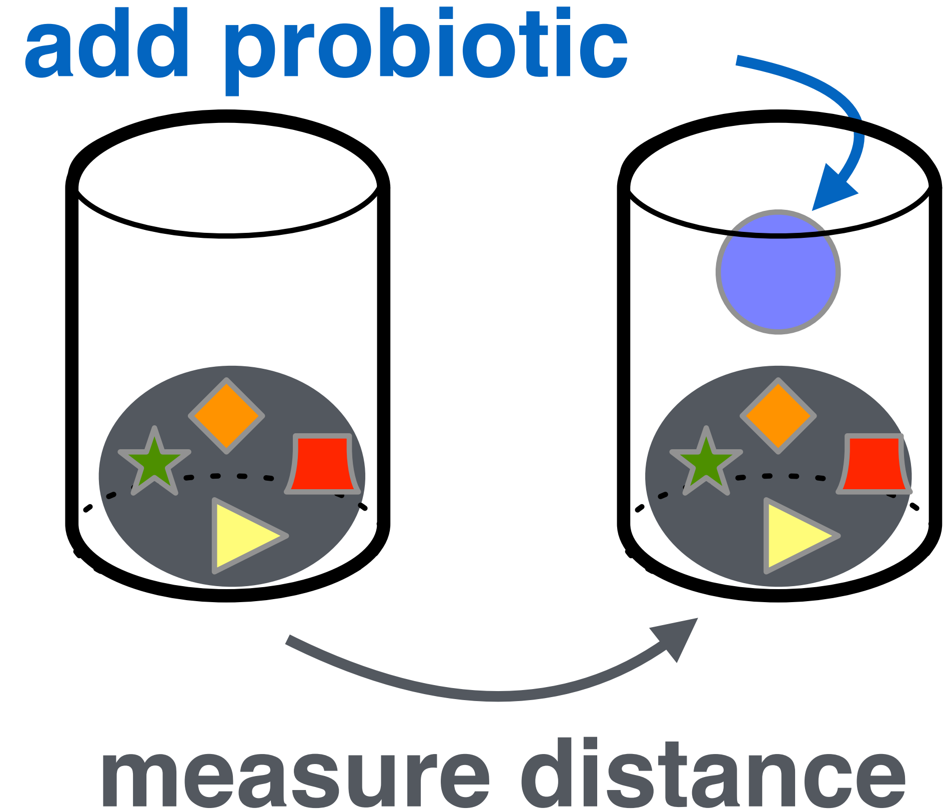
L+B+R=k

And all Positive

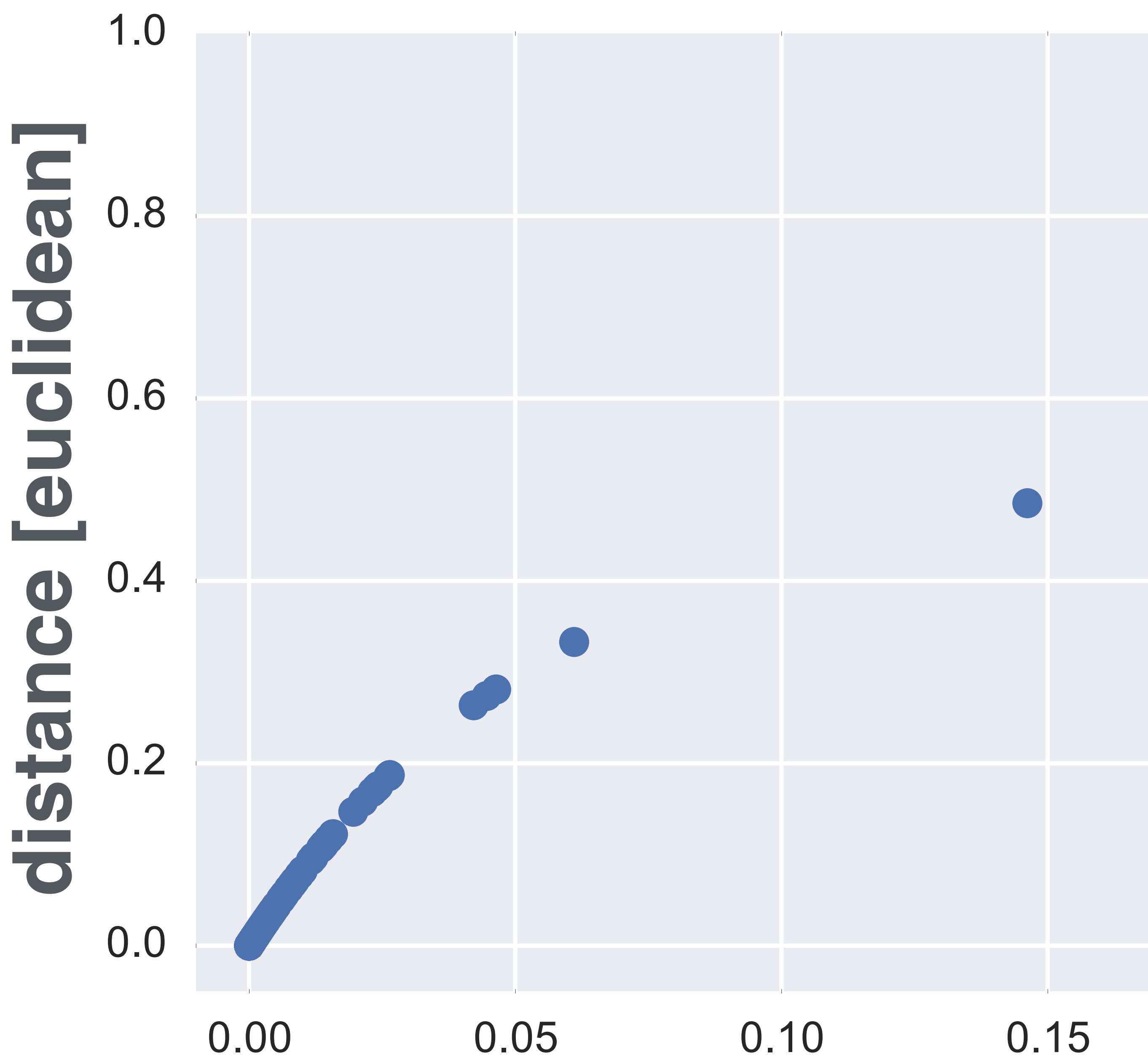


MODELING

CHALLENGE

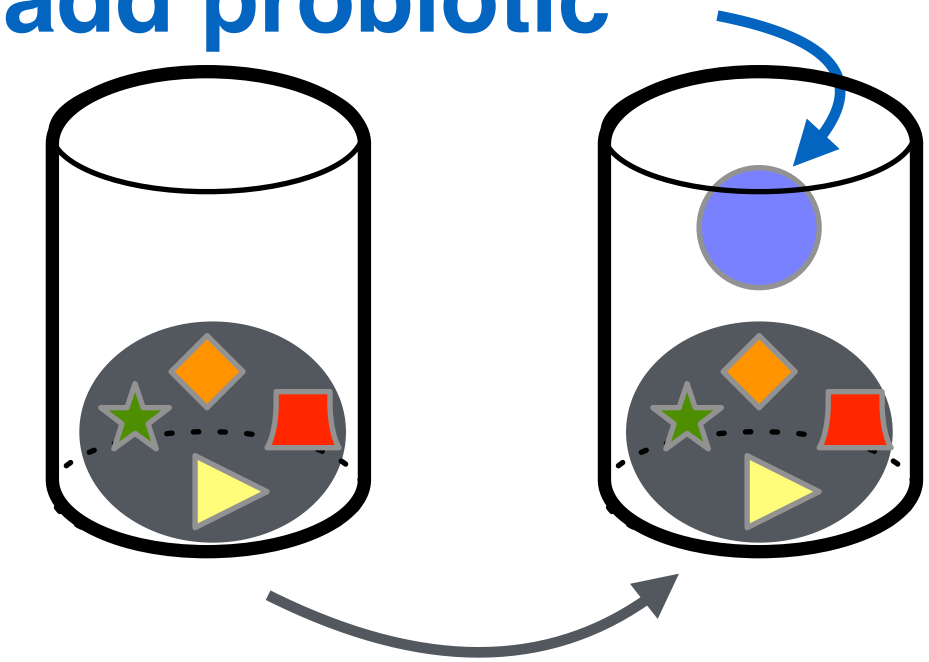


MODELING
CHALLENGE



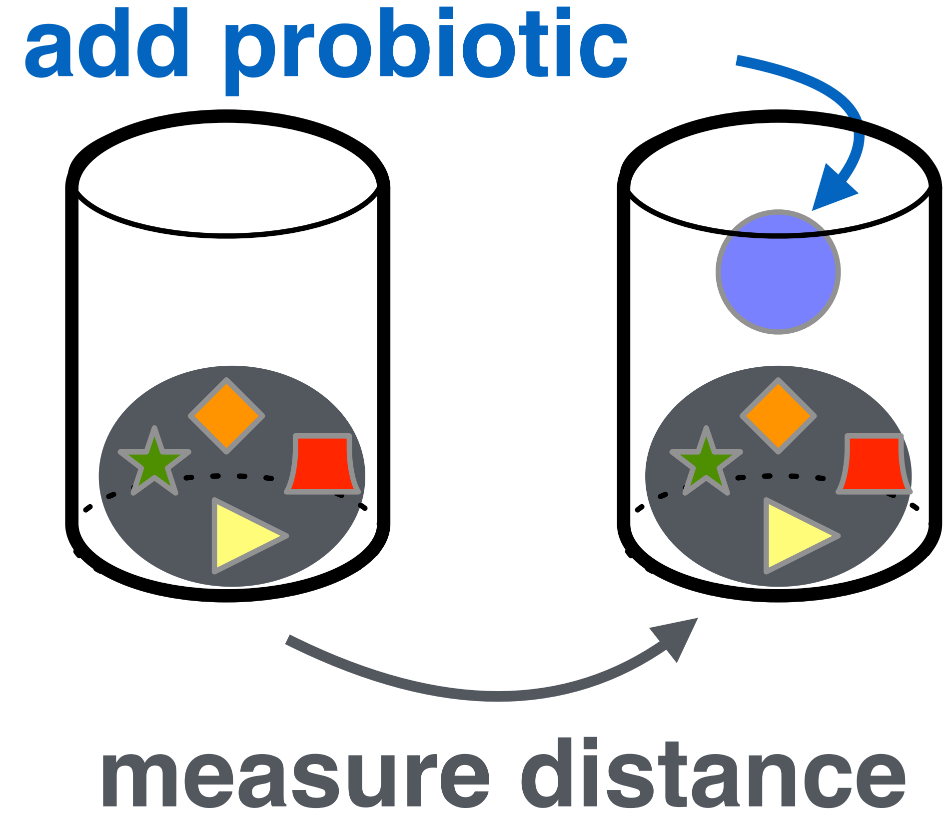
initial probiotic dosage [fraction]

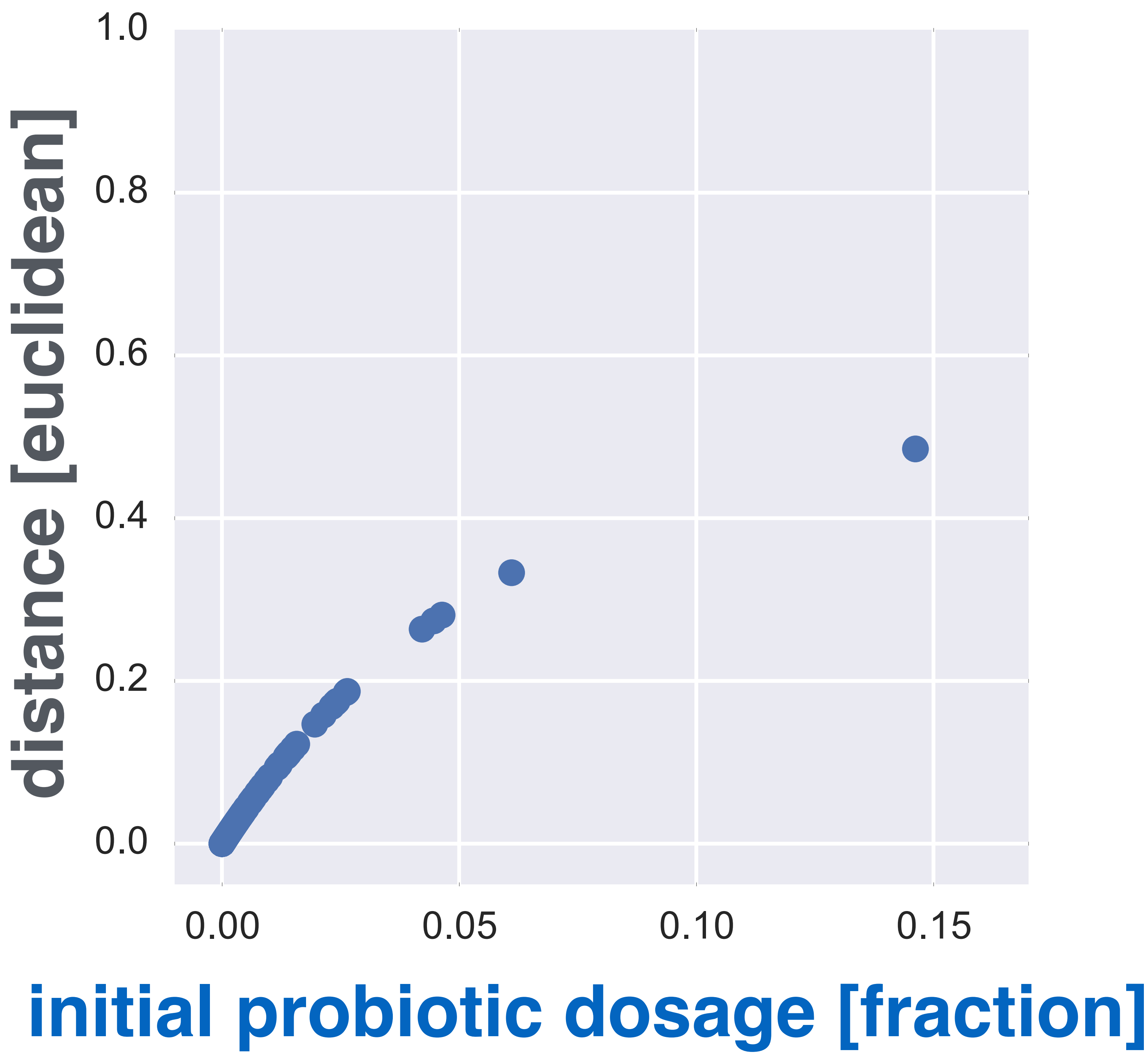
add probiotic



measure distance

MODELING
CHALLENGE



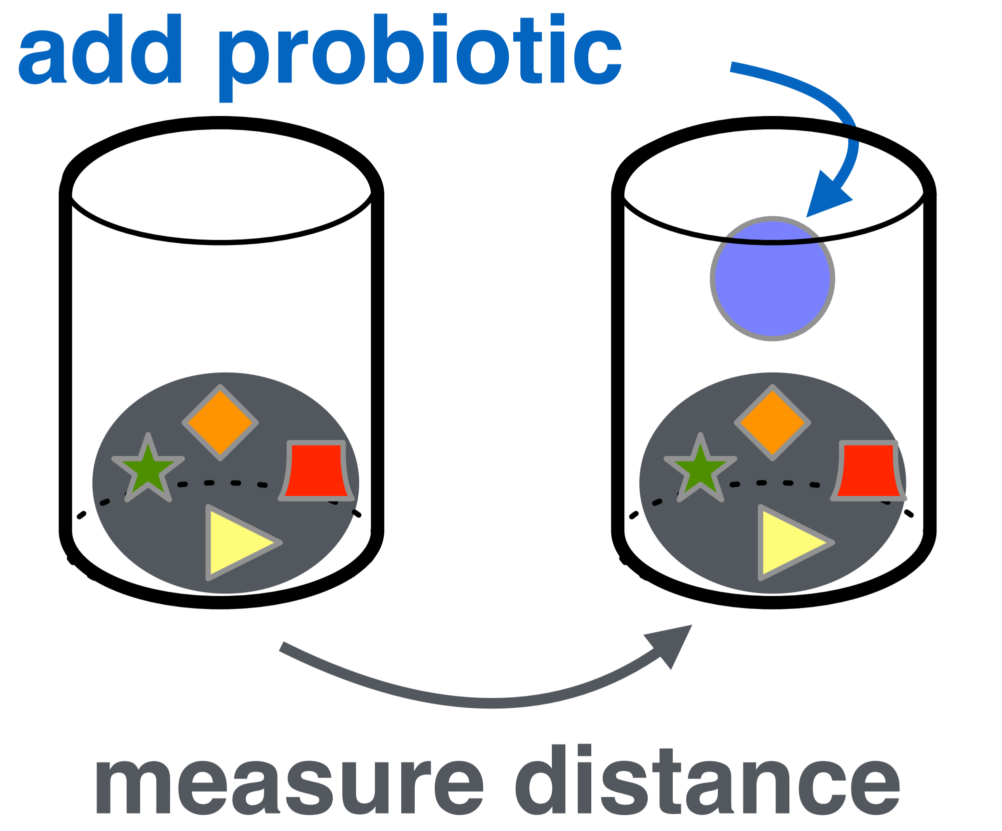


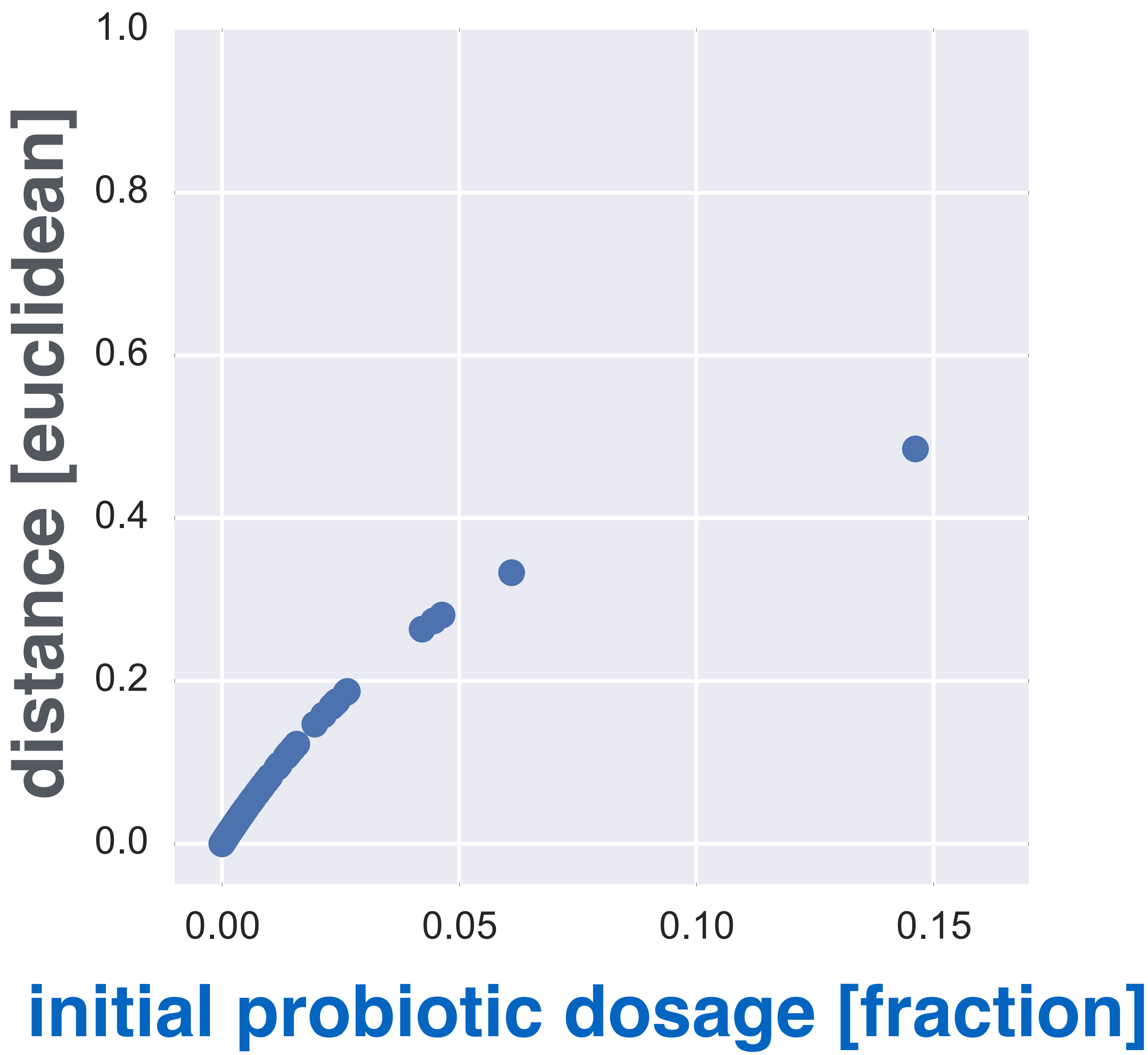
Challenges:

- Probiotic **addition alone** shifts community composition

MODELING

CHALLENGE





Challenges:

- Probiotic **addition alone** shifts community composition
- Shifts are **biased by probiotic dosage**

APPROACHES
TO CORRECTING DISTANCE

Transform statistics:

APPROACHES
TO CORRECTING DISTANCE

Transform statistics:

Aitchison distance

$$d(x_i, x_j) = \left(\sum_{k=1}^{D} (log(\frac{x_{ik}}{g(\mathbf{x}_i)}) - log(\frac{x_{jk}}{g(\mathbf{x}_j)}))^2 \right)^{\frac{1}{2}}$$

Aitchison, 1986

APPROACHES
TO CORRECTING DISTANCE

Transform statistics:

Aitchison distance

$$d(x_i, x_j) = \left(\sum_{k=1}^{D} (log(\frac{x_{ik}}{g(\mathbf{x}_i)}) - log(\frac{x_{jk}}{g(\mathbf{x}_j)}))^2 \right)^{\frac{1}{2}}$$

Aitchison, 1986

Transform data (x↔y):

APPROACHES

TO CORRECTING DISTANCE

Transform statistics:

Aitchison distance

$$d(x_i, x_j) = \left(\sum_{k=1}^{D} (log(\frac{x_{ik}}{g(\mathbf{x}_i)}) - log(\frac{x_{jk}}{g(\mathbf{x}_j)}))^2 \right)^{\frac{1}{2}}$$

Aitchison, 1986

Transform data (x↔y):

Euclidean distance

$$d(y_i, y_j) = \left(\sum_{k=1}^{N} (y_{ik} - y_{jk})^2 \right)^{\frac{1}{2}}$$

requirements for a proper analysis

- **scale invariance:** the analysis should not depend on the closure constant κ; proportional positive vectors are equivalent as compositions
- **permutation invariance:** the order of the parts should be irrelevant
- **subcompositional coherence:** studies performed on subcompositions should not stand in contradiction with those performed on the full composition

centred log-ratio (clr) transformation

Composition $\mathbf{x} \in \mathcal{S}^D$

clr transformation of $\mathbf{x}$ is the $\mathbb{R}^D$-vector

$$\mathrm{clr}(\mathbf{x}) = \mathbf{v} = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_i}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right]$$

$$g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}, \quad v_i = \ln \frac{x_i}{g(\mathbf{x})}, \quad \sum_{i=1}^{D} v_i = 0$$

clr inverse: back into the simplex $\mathcal{S}^D$

$$\mathbf{x} = \mathrm{clr}^{-1}(\mathbf{v}) = \mathcal{C} \exp[v_1, v_2, \ldots, v_D]$$

Properties of clr representation

- **clr coefficients are log-contrasts**

$$v_i = \sum_{k=1}^{D} \alpha_k \ln x_k \ , \ \alpha_i = 1 - \frac{1}{D} \ , \ \alpha_k = -\frac{1}{D}$$

- **isometry $\mathcal{S}^D \longrightarrow \mathbb{R}_0^D$**

  clr transforms $\oplus$, $\odot$ into $+$, $\cdot$

$$\mathrm{clr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \mathrm{clr}(\mathbf{x}_1) + \beta \cdot \mathrm{clr}(\mathbf{x}_2)$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2) \rangle$$

$$\|\mathbf{x}_1\|_a = \|\mathrm{clr}(\mathbf{x}_1)\| \quad , \quad d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2))$$

Orthonormal basis

Definition
Compositions $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_{D-1}$ in $\mathcal{S}^D$ are an orthonormal basis if

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = \langle \mathrm{clr}(\mathbf{e}_i), \mathrm{clr}(\mathbf{e}_j) \rangle = \delta_{ij}$$

Basis contrast matrix: clr matrix of the basis $(D-1, D)$

$$\mathbf{\Psi} = \begin{pmatrix} \mathrm{clr}(\mathbf{e}_1) \\ \mathrm{clr}(\mathbf{e}_2) \\ \dots \\ \mathrm{clr}(\mathbf{e}_{D-1}) \end{pmatrix} \quad , \quad \mathbf{\Psi}\mathbf{\Psi}' = I_{D-1} \quad , \quad \mathbf{\Psi}'\mathbf{\Psi} = I_D - (1/D)\mathbf{1}'_D\mathbf{1}_D$$

Coordinates

Given an orthonormal basis $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_{D-1}$ in $\mathcal{S}^D$,
Expression in coordinates

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i \ , \ x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a$$

Isometric log-ratio: assigns coordinates to a composition
$\mathrm{ilr} : \mathcal{S}^D \to \mathbb{R}^{D-1}$ is one-to-one.

$$\mathbf{x} \overset{\mathrm{ilr}}{\longrightarrow} \mathbf{x}^* = [x_1^*, x_2^*, \ldots, x_{D-1}^*]$$

Properties of ilr-coordinates

Given an orthonormal basis $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_{D-1}$ in $\mathcal{S}^D$

ilr and ilr^{-1}

$$\mathbf{x}^* = \mathrm{ilr}(\mathbf{x}) = \mathrm{clr}(\mathbf{x}) \cdot \boldsymbol{\Psi}' \quad , \quad \mathbf{x} = \mathcal{C}\left(\exp(\mathbf{x}^* \boldsymbol{\Psi})\right)$$

Isometry: $\mathrm{ilr} : \mathcal{S}^D \longrightarrow \mathbb{R}^{D-1}$

$$\mathrm{ilr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \mathrm{ilr}(\mathbf{x}_1) + \beta \cdot \mathrm{ilr}(\mathbf{x}_2) = \alpha \cdot \mathbf{x}_1^* + \beta \cdot \mathbf{x}_2^*$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathrm{ilr}(\mathbf{x}_1), \mathrm{ilr}(\mathbf{x}_2) \rangle = \langle \mathbf{x}_1^*, \mathbf{x}_2^* \rangle$$

$$\|\mathbf{x}\|_a = \|\mathrm{ilr}(\mathbf{x})\| \quad , \quad d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\mathrm{ilr}(\mathbf{x}_1), \mathrm{ilr}(\mathbf{x}_2))$$

Building an orthonormal basis of balances

the intuitive approach

example: for $\mathbf{x} \in \mathcal{S}^5$ define a sequential binary partition and obtain the coordinates in the corresponding orthonormal basis

order	x_1	x_2	x_3	x_4	x_5	coordinate
1	$+1$	-1	$+1$	$+1$	-1	$y_1 = \sqrt{\frac{3\cdot 2}{3+2}} \ln \frac{(x_1 \cdot x_3 \cdot x_4)^{1/3}}{(x_2 \cdot x_5)^{1/2}}$
2	0	$+1$	0	0	-1	$y_2 = \sqrt{\frac{1\cdot 1}{1+1}} \ln \frac{x_2}{x_5}$
3	$+1$	0	-1	-1	0	$y_3 = \sqrt{\frac{1\cdot 2}{1+2}} \ln \frac{x_1}{(x_3 \cdot x_4)^{1/2}}$
4	0	0	$+1$	-1	0	$y_4 = \sqrt{\frac{1\cdot 1}{1+1}} \ln \frac{x_3}{x_4}$

Balances and balancing elements

Coordinates in an orthonormal basis obtained from a sequential binary partition:

$$y_i = \sqrt{\frac{r_i \cdot s_i}{r_i + s_i}} \ln \frac{\left(\prod_{j \in R_i} x_j\right)^{1/r_i}}{\left(\prod_{\ell \in S_i} x_\ell\right)^{1/s_i}}$$

where $i =$ order of partition, R_i and S_i index sets, r_i the number of indices in R_i, s_i the number in S_i
The corresponding balancing element is

$$\mathbf{e}_i = \mathcal{C}(\exp[\psi_{i1}, \psi_{i2}, \ldots, \psi_{iD}])$$

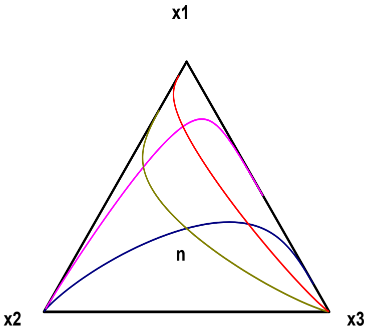
$$\psi_{i+} = +\sqrt{\frac{s_i}{r_i(r_i + s_i)}} \quad , \quad \psi_{i-} = -\sqrt{\frac{r_i}{s_i(r_i + s_i)}} \quad , \quad \psi_{i0} = 0$$

parallel lines

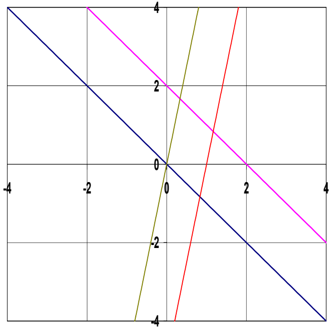
Processes of exponential growth or decay are straight-lines:

$$x_i(t) = x_i(0) \cdot \exp(\lambda_i t)\ ,\ \ i = 1, 2, \ldots, D$$

$$\mathbf{x}(t) = \mathbf{x}(0) \oplus (t \odot \exp(\boldsymbol{\lambda}))$$

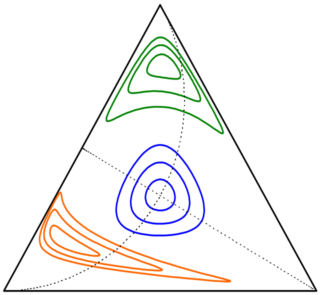


in $\mathcal{S}^3$

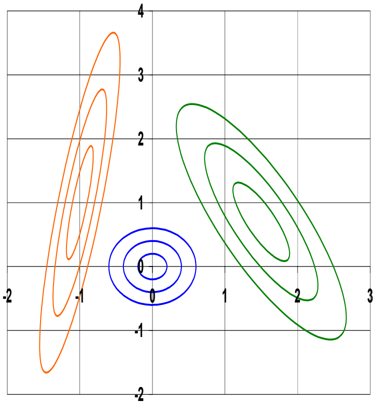


coordinate representation

circles and ellipses



in $\mathcal{S}^3$



coordinate representation

variability, centre and variance

$\mathbf{X}$ random composition in $\mathcal{S}^D$
$\mathbf{X}^*$ random ilr-coordinates in $\mathbb{R}^{D-1}$
$\mathbf{z} \in \mathcal{S}^D$

variability with respect to z

$$\mathrm{Var}[\mathbf{X};\mathbf{z}] = \int_{\mathbb{R}^{D-1}} d^2(\mathbf{X}^*,\mathbf{z}^*)\ f_{\mathbf{X}^*}\ d\mathbf{x}^*$$

centre and total variance

$$\mathrm{Cen}[\mathbf{X}] = \operatorname*{argmin}_{\mathbf{z}} \mathrm{Var}[\mathbf{X};\mathbf{z}] \quad , \quad \mathrm{TotVar}[\mathbf{X}] = \min_{\mathbf{z}} \mathrm{Var}[\mathbf{X};\mathbf{z}]$$

computation of centre

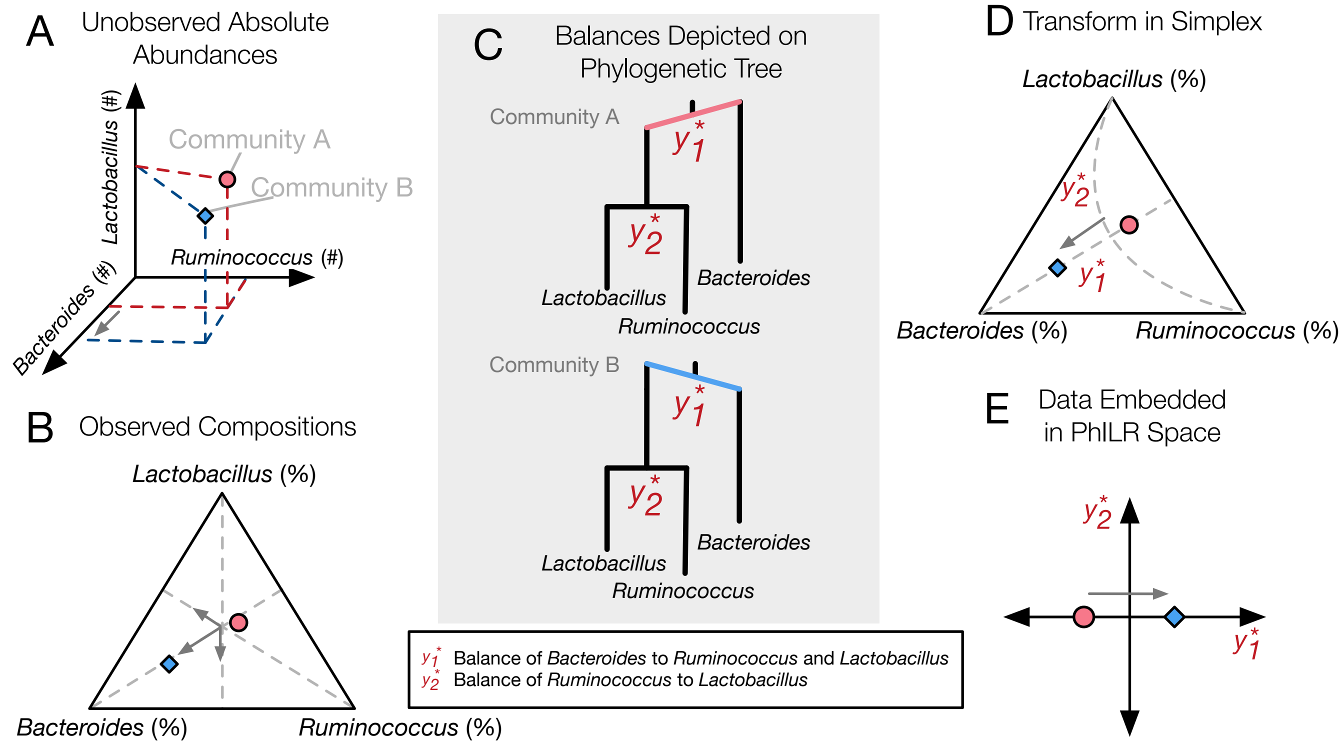
$$\mathrm{Cen}[\mathbf{X}] = \mathrm{ilr}^{-1}(\mathrm{E}[\mathrm{ilr}(\mathbf{X})]) = \mathcal{C}\exp(\mathrm{E}[\ln(\mathbf{x})])$$

three decompositions of total variance

$\mathbf{X}$ random composition in $\mathcal{S}^D$
$\mathbf{X}^* = \mathrm{ilr}(\mathbf{X})$ random ilr-coordinates in $\mathbb{R}^{D-1}$
the decomposition of total variance

$$\begin{aligned}
\mathrm{TotVar}[\mathbf{X}] &= \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathrm{Var}\left[\ln\frac{x_i}{x_j}\right] \\
&= \sum_{i=1}^{D}\mathrm{Var}[\mathrm{clr}_i(\mathbf{X})] \\
&= \sum_{j=1}^{D-1}\mathrm{Var}[\mathrm{ilr}_j(\mathbf{X})]
\end{aligned}$$

basic in exploratory analysis and linear modelling



A
Unobserved Absolute Abundances
Lactobacillus (#)
Community A
Community B
Ruminococcus (#)
Bacteroides (#)
B
Observed Compositions
Lactobacillus (%)
Bacteroides (%)
Ruminococcus (%)
C
Balances Depicted on Phylogenetic Tree
Community A
y_1^*
y_2^*
Lactobacillus
Ruminococcus
Bacteroides
Community B
y_1^*
y_2^*
Lactobacillus
Ruminococcus
Bacteroides
y_1^* Balance of Bacteroides to Ruminococcus and Lactobacillus
y_2^* Balance of Ruminococcus to Lactobacillus
D
Transform in Simplex
Lactobacillus (%)
y_2^*
y_1^*
Bacteroides (%)
Ruminococcus (%)
E
Data Embedded in PhILR Space
y_2^*
y_1^*

some specific references

Egozcue, J.J., V. Pawlowsky-Glahn, G. Mateu-Figueras and C. Barceló-Vidal: Isometric logratio transformations for compositional data analysis, Mathematical Geology, 35, 3, 279-300, 2003.

Egozcue, J.J. and V. Pawlowsky-Glahn: Groups of parts and their balances in compositional data analysis, Mathematical Geology, 37, 7, 799-832, 2005.

Egozcue, J. J. and V. Pawlowsky-Glahn: Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V., (eds) Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, (ISBN: 1-86239-205-6), Special Publications, 264, 67-77, 2006.

Egozcue, J. J. and Pawlowsky-Glahn, V.: Basic concepts and procedures. Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J.: The principle of working on coordinates.

both In Pawlowsky-Glahn, V. and Buccianti A. (Eds.) *Compositional Data Analysis: Theory and Applications*, ISBN-10: 0-470-71135-3, Wiley, Chichester UK, 2011.