UBIQUITY OF COMPOSITIONAL DATA

Compositional: Relating to parts of some whole
- Proportions
- Parts per million
- Percentages

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)
COMPOSITIONAL SIMPLEX

\[ L + B + R = k \]

And all Positive
MODELING CHALLENGE

add probiotic

measure distance
MODELING CHALLENGE

Isolate initial abundance [Iraction]

Community shift [euclidean distance]

Initial probiotic dosage [fraction]

Add probiotic

Measure distance
MODELING CHALLENGE

Challenges:
- Probiotic addition alone shifts community composition

Initial probiotic dosage [fraction] vs. distance [euclidean]
Isolate initial abundance

Community shift

Probiotic addition alone shifts community composition

Challenges:

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

MODELING CHALLENGE

add probiotic

measure distance

Challenges:
APPROACHES TO CORRECTING DISTANCE

Transform statistics:
**APPROACHES TO CORRECTING DISTANCE**

**Transform statistics:**

**Aitchison distance**

\[ d(x_i, x_j) = \left( \sum_{k=1}^{D} \left( \log\left( \frac{x_{ik}}{g(x_i)} \right) - \log\left( \frac{x_{jk}}{g(x_j)} \right) \right)^2 \right)^{\frac{1}{2}} \]

Aitchison, 1986
**Transform statistics:**

**Aitchison distance**

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

*Source: Aitchison, 1986*

---

**APPROACHES TO CORRECTING DISTANCE**

- Aitchison distance
- Euclidean distance
APPROACHES TO CORRECTING DISTANCE

Transform statistics:

Aitchison distance

\[ d(x_i, x_j) = \left( \sum_{k=1}^{D} \left( \log \left( \frac{x_{ik}}{g(x_i)} \right) - \log \left( \frac{x_{jk}}{g(x_j)} \right) \right)^2 \right)^{1/2} \]

Aitchison, 1986

Transform data \((x \leftrightarrow y)\):

Euclidean distance

\[ d(y_i, y_j) = \left( \sum_{k=1}^{N} (y_{ik} - y_{jk})^2 \right)^{1/2} \]
requirements for a proper analysis

- **scale invariance:** the analysis should not depend on the closure constant $\kappa$; 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
Composition $x \in S^D$

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

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

$$g(x) = \left( \prod_{i=1}^{D} x_i \right)^{1/D}$$

$$v_i = \ln \frac{x_i}{g(x)} \quad \sum_{i=1}^{D} v_i = 0$$

**clr inverse:** back into the simplex $S^D$

$$x = \text{clr}^{-1}(v) = C \exp[v_1, v_2, \ldots, v_D]$$
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 \( S^D \rightarrow \mathbb{R}^D_0 \)

  clr transforms \( \oplus, \odot \) into \( +, \cdot \)

  \[
  \text{clr}(\alpha \odot x_1 \oplus \beta \odot x_2) = \alpha \cdot \text{clr}(x_1) + \beta \cdot \text{clr}(x_2)
  \]

  \[
  \langle x_1, x_2 \rangle_a = \langle \text{clr}(x_1), \text{clr}(x_2) \rangle
  \]

  \[
  \| x_1 \|_a = \| \text{clr}(x_1) \| \ , \ \ d_a(x_1, x_2) = d(\text{clr}(x_1), \text{clr}(x_2))
  \]
Orthonormal basis

Definition
Compositions $e_1, e_2, ..., e_{D-1}$ in $S^D$ are an orthonormal basis if

$$
\langle e_i, e_j \rangle_a = \langle \text{clr}(e_i), \text{clr}(e_j) \rangle = \delta_{ij}
$$

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

$$
\psi = \begin{pmatrix}
\text{clr}(e_1) \\
\text{clr}(e_2) \\
\vdots \\
\text{clr}(e_{D-1})
\end{pmatrix}, \quad \psi \psi' = I_{D-1}, \quad \psi' \psi = I_D - (1/D) 1'_D 1_D
$$
Given an orthonormal basis $e_1, e_2, \ldots, e_{D-1}$ in $S^D$,

**Expression in coordinates**

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

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

$$\operatorname{ilr} \quad \mathbf{x} \rightarrow \mathbf{x}^* = [x_1^*, x_2^*, \ldots, x_{D-1}^*]$$
Properties of ilr-coordinates

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

$\text{ilr}$ and $\text{ilr}^{-1}$

$$\mathbf{x}^* = \text{ilr}(\mathbf{x}) = \text{clr}(\mathbf{x}) \cdot \Psi' \quad , \quad \mathbf{x} = C(\exp(\mathbf{x}^* \Psi))$$

**Isometry:** $\text{ilr} : S^D \rightarrow \mathbb{R}^{D-1}$

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

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

$$\|\mathbf{x}\|_a = \|\text{ilr}(\mathbf{x})\| \quad , \quad d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\text{ilr}(\mathbf{x}_1), \text{ilr}(\mathbf{x}_2))$$
Building an orthonormal basis of balances

the intuitive approach

define a sequential binary partition and obtain the coordinates in the corresponding orthonormal basis

<table>
<thead>
<tr>
<th>order</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>coordinate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+1</td>
<td>−1</td>
<td>+1</td>
<td>+1</td>
<td>−1</td>
<td>$y_1 = \sqrt{\frac{3 \cdot 2}{3+2}} \ln \frac{(x_1 \cdot x_3 \cdot x_4)^{1/3}}{(x_2 \cdot x_5)^{1/2}}$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>−1</td>
<td>$y_2 = \sqrt{\frac{1 \cdot 1}{1+1}} \ln \frac{x_2}{x_5}$</td>
</tr>
<tr>
<td>3</td>
<td>+1</td>
<td>0</td>
<td>−1</td>
<td>−1</td>
<td>0</td>
<td>$y_3 = \sqrt{\frac{1 \cdot 2}{1+2}} \ln \frac{x_1}{(x_3 \cdot x_4)^{1/2}}$</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>−1</td>
<td>0</td>
<td>$y_4 = \sqrt{\frac{1 \cdot 1}{1+1}} \ln \frac{x_3}{x_4}$</td>
</tr>
</tbody>
</table>
Balances and balancing elements

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

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

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

\[ e_i = C(\exp[\psi_{i1}, \psi_{i2}, \ldots, \psi_{iD}]) \]

\[ \psi_{i+} = + \sqrt{\frac{s_i}{r_i(r_i + s_i)}} \quad , \quad \psi_{i-} = - \sqrt{\frac{r_i}{s_i(r_i + s_i)}} \quad , \quad \psi_{i0} = 0 \]
Processes of exponential growth or decay are straight-lines:

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

\[ x(t) = x(0) \oplus (t \circ \exp(\lambda)) \]
circles and ellipses

in $S^3$

coordinate representation
variability, centre and variance

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

variability with respect to $z$

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

centre and total variance

$$\text{Cen}[X] = \arg\min_z \text{Var}[X; z] \quad , \quad \text{TotVar}[X] = \min_z \text{Var}[X; z]$$

computation of centre

$$\text{Cen}[X] = \text{ilr}^{-1}(E[\text{ilr}(X)]) = C \exp(E[\ln(x)])$$
three decompositions of total variance

$\mathbf{X}$ random composition in $S^D$

$\mathbf{X}^* = \text{ilr}(\mathbf{X})$ random ilr-coordinates in $\mathbb{R}^{D-1}$

the decomposition of total variance

$$\text{TotVar}[\mathbf{X}] = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \text{Var} \left[ \ln \frac{x_i}{x_j} \right]$$

$$= \sum_{i=1}^{D} \text{Var}[\text{clr}_i(\mathbf{X})]$$

$$= \sum_{j=1}^{D-1} \text{Var}[\text{ilr}_j(\mathbf{X})]$$

basic in exploratory analysis and linear modelling
Balances Depicted on Phylogenetic Tree

Transform in Simplex

Data Embedded in PhILR Space
some specific references


