Background on 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)
UBIQUITY OF COMPOSITIONAL DATA COUNTS!

1200 Blue
1100 Orange
1300 Green

31% Blue
19% Orange
50% Green
Examples

• Abundance Quantification by High Throughout Sequencing
  • **Microbiome Composition (e.g., counts of 16s gene)**
  • Gene expression analysis (e.g., RNA-seq)
• Abundance Quantification by Flow Cytometry
• Proportion of observed mice that go on to develop a disease?
• Population of North Carolina that is pro-Trump? (e.g., polling results)
RESULTING COUNT TABLE

<table>
<thead>
<tr>
<th></th>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
<th>Species 4</th>
<th>Species 5</th>
<th>Species 6</th>
<th>Species 7</th>
<th>Species 8</th>
<th>Species 9</th>
<th>Species 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>23</td>
<td>53</td>
<td>2</td>
<td>44</td>
<td>10</td>
<td>88</td>
<td>94</td>
<td>66</td>
<td>73</td>
<td>67</td>
</tr>
<tr>
<td>Sample 2</td>
<td>69</td>
<td>64</td>
<td>70</td>
<td>47</td>
<td>8</td>
<td>97</td>
<td>47</td>
<td>6</td>
<td>64</td>
<td>19</td>
</tr>
<tr>
<td>Sample 3</td>
<td>33</td>
<td>100</td>
<td>68</td>
<td>78</td>
<td>59</td>
<td>87</td>
<td>71</td>
<td>31</td>
<td>67</td>
<td>24</td>
</tr>
<tr>
<td>Sample 4</td>
<td>5</td>
<td>63</td>
<td>57</td>
<td>27</td>
<td>86</td>
<td>81</td>
<td>83</td>
<td>92</td>
<td>46</td>
<td>62</td>
</tr>
<tr>
<td>Sample 5</td>
<td>76</td>
<td>80</td>
<td>46</td>
<td>70</td>
<td>92</td>
<td>92</td>
<td>6</td>
<td>46</td>
<td>37</td>
<td>68</td>
</tr>
<tr>
<td>Sample 6</td>
<td>58</td>
<td>7</td>
<td>37</td>
<td>45</td>
<td>25</td>
<td>62</td>
<td>78</td>
<td>44</td>
<td>89</td>
<td>30</td>
</tr>
<tr>
<td>Sample 7</td>
<td>10</td>
<td>87</td>
<td>32</td>
<td>80</td>
<td>9</td>
<td>91</td>
<td>59</td>
<td>90</td>
<td>67</td>
<td>77</td>
</tr>
<tr>
<td>Sample 8</td>
<td>21</td>
<td>89</td>
<td>73</td>
<td>39</td>
<td>44</td>
<td>80</td>
<td>97</td>
<td>83</td>
<td>80</td>
<td>4</td>
</tr>
<tr>
<td>Sample 9</td>
<td>85</td>
<td>77</td>
<td>82</td>
<td>72</td>
<td>15</td>
<td>19</td>
<td>44</td>
<td>4</td>
<td>83</td>
<td>76</td>
</tr>
<tr>
<td>Sample 10</td>
<td>67</td>
<td>87</td>
<td>68</td>
<td>58</td>
<td>73</td>
<td>29</td>
<td>87</td>
<td>4</td>
<td>48</td>
<td>79</td>
</tr>
<tr>
<td>Sample 11</td>
<td>90</td>
<td>5</td>
<td>28</td>
<td>49</td>
<td>39</td>
<td>20</td>
<td>78</td>
<td>92</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>Sample 12</td>
<td>98</td>
<td>93</td>
<td>55</td>
<td>12</td>
<td>54</td>
<td>75</td>
<td>27</td>
<td>95</td>
<td>83</td>
<td>98</td>
</tr>
<tr>
<td>Sample 13</td>
<td>31</td>
<td>97</td>
<td>52</td>
<td>9</td>
<td>93</td>
<td>84</td>
<td>45</td>
<td>97</td>
<td>81</td>
<td>27</td>
</tr>
<tr>
<td>Sample 14</td>
<td>12</td>
<td>77</td>
<td>22</td>
<td>17</td>
<td>71</td>
<td>12</td>
<td>56</td>
<td>86</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Sample 15</td>
<td>40</td>
<td>30</td>
<td>71</td>
<td>71</td>
<td>54</td>
<td>13</td>
<td>77</td>
<td>96</td>
<td>75</td>
<td>11</td>
</tr>
<tr>
<td>Sample 16</td>
<td>43</td>
<td>94</td>
<td>40</td>
<td>73</td>
<td>27</td>
<td>33</td>
<td>97</td>
<td>88</td>
<td>81</td>
<td>44</td>
</tr>
</tbody>
</table>
The Shape of Compositional Data (Microbiome Example)
compositional data: usual representation

**definition:** \( \mathbf{x} = [x_1, x_2, \ldots, x_D] \) is a \( D \)-part composition

\[
\begin{align*}
&x_i > 0, \quad \text{for all } i = 1, \ldots, D \\
&\sum_{i=1}^{D} x_i = \kappa \quad (\text{constant})
\end{align*}
\]

\( \kappa = 1 \iff \text{measurements in parts per unit} \)

\( \kappa = 100 \iff \text{measurements in percent} \)

other frequent units: ppm, ppb, ...

a composition is the representative in the simplex of equivalent vectors with strictly positive components

a **subcomposition** \( \mathbf{x}_s \) with \( s \) parts is obtained as the closure of a subvector \( [x_{i_1}, x_{i_2}, \ldots, x_{i_s}] \) of \( \mathbf{x} \)
composition data: definition

**Definition:** parts of some whole which carry only relative information

Proportional vectors with strictly positive components are compositionally equivalent if they are proportional: each equivalence class represents a composition.

usual representation: subject to a **constant sum constraint**
the simplex as sample space

\[ S^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D] \mid x_i > 0; \sum_{i=1}^{D} x_i = \kappa \right\} \]

standard representation for \( D = 3 \):
the ternary diagram
STATISTICAL COMFORT ZONE
\[ X + Y = k \]

And all Positive
COMPOSITIONAL
SIMPLEX

$\text{L+B+R}=k$

And all Positive
MODELING CHALLENGE

Lactobacillus

10%

40%

50%

Bacteroides

Ruminococcus

L+B+R=k
And all Positive
MODELING CHALLENGE

EVENT: DELIVERY OF BACTEROIDES PROBIOTIC
MODELING CHALLENGE

EVENT: BACTEROIDES
DELIVERY OF PROBIOTIC

CHALLENGE: CHANGES IN FRACTIONS OF LACTOBACILLUS
EVENT: BACTERIOIDES
PROBIOTIC

CHALLENGE: CHANGES IN FRACTIONS OF LACTOBACILLUS & RUMINOCOCCUS
MODELING CHALLENGE

- Add probiotic
- Measure distance
MODELING CHALLENGE

Isolate initial abundance [Iraction]

Community shift [euclidean distance]

- initial probiotic dosage [fraction]
  - add probiotic
  - measure distance

**Diagram:**
- X-axis: initial probiotic dosage [fraction]
- Y-axis: distance [euclidean]
MODELING CHALLENGE

Isolate initial abundance [Israction]

Community shift [euclidean distance]

• Probiotic addition alone shifts community composition

Challenges:
• Probiotic addition alone shifts community composition

initial probiotic dosage [fraction]

distance [euclidean]

add probiotic

measure distance

Challenges:
MODELING CHALLENGE

Challenges:

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

Isolate initial abundance

Community shift (euclidean distance)

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

Given: \( X + Y + Z = 100\% \)

If: \( X \) increases \( \rightarrow \) \( Y + Z \) must decrease

Not actually 3 independent variables
Compositional Effects

Fake Data
Compositional Effects

Fake Data (Normalized)
Compositional Effects

\[ y_t = \beta x_t + \alpha \]
Compositional Effects

Real Data
Compositional Effects

Real Data

Fake Data
the problem: negative bias & spurious correlation

**example:** scientists A and B record the composition of aliquots of soil samples; A records (animal, vegetable, mineral, water) compositions, B records (animal, vegetable, mineral) after drying the sample; both are absolutely accurate

(adapted from Aitchison, 2005)

<table>
<thead>
<tr>
<th>sample A</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sample B</th>
<th>$x'_1$</th>
<th>$x'_2$</th>
<th>$x'_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>0.40</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>3</td>
<td>0.43</td>
<td>0.43</td>
<td>0.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>corr A</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>1.00</td>
<td>0.50</td>
<td>0.00</td>
<td>-0.98</td>
</tr>
<tr>
<td>$x_2$</td>
<td>1.00</td>
<td>-0.87</td>
<td>-0.65</td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>1.00</td>
<td>0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_4$</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>corr B</th>
<th>$x'_1$</th>
<th>$x'_2$</th>
<th>$x'_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x'_1$</td>
<td>1.00</td>
<td>-0.57</td>
<td>-0.05</td>
</tr>
<tr>
<td>$x'_2$</td>
<td>1.00</td>
<td>-0.79</td>
<td></td>
</tr>
<tr>
<td>$x'_3$</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
why a new geometry on the simplex?

in real space $\mathbb{R}^D$ we **add** vectors, we **multiply** them by a constant, we look for **orthogonality** between vectors, we look for **distances** between points, ...

possible because $\mathbb{R}^D$ is a Euclidean vector space

**BUT** the usual Euclidean geometry in real space is not a proper geometry for compositional data because

- **results might not be in the simplex** when we **add** compositional vectors, **multiply** them by a constant, or compute **confidence regions**

- **Euclidean differences are not always reasonable:** from 0.05% to 0.10% the amount is doubled; from 50.05% to 50.10% the increase is negligible
**basic operations**

**closure** of \( z = [z_1, z_2, \ldots, z_D] \in \mathbb{R}_+^D \), with closure constant = \( \kappa \)

\[
C\left[ z \right] = \begin{bmatrix}
\frac{\kappa \cdot z_1}{\sum_{i=1}^{D} z_i}, & \frac{\kappa \cdot z_2}{\sum_{i=1}^{D} z_i}, & \cdots, & \frac{\kappa \cdot z_D}{\sum_{i=1}^{D} z_i}
\end{bmatrix}
\]

\( C\left[ z \right] \) is the representative of \( z \) in \( S^D \)

**perturbation** of \( x \in S^D \) by \( y \in S^D \)

\[
x \oplus y = C\left[ x_1 y_1, x_2 y_2, \ldots, x_D y_D \right]
\]

**powering** of \( x \in S^D \) by \( \alpha \in \mathbb{R} \)

\[
\alpha \odot x = C\left[ x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha \right]
\]
interpretation of perturbation and powering

**left:** perturbation of initial compositions (○) by \( \mathbf{p} = [0.1, 0.1, 0.8] \) resulting in compositions (★)

**right:** powering of compositions (★) by \( \alpha = 0.2 \) resulting in compositions (○)
comments

- closure = projection of a point in $\mathbb{R}_+^D$ on $S^D$

- points on a ray are projected onto the same point
  - a ray in $\mathbb{R}_+^D$ is an equivalence class
  - the point on $S^D$ is a representative of the class
  - a generalization to other representatives is possible

- for $\mathbf{z} \in \mathbb{R}_+^D$ and $\mathbf{x} \in S^D$, $\mathbf{x} \oplus (\alpha \odot \mathbf{z}) = \mathbf{x} \oplus (\alpha \odot C[\mathbf{z}])$
vector space structure of \((S^D, \oplus, \odot)\)

- **commutative group structure** of \((S^D, \oplus)\)
  1. commutativity: \(x \oplus y = y \oplus x\)
  2. associativity: \((x \oplus y) \oplus z = x \oplus (y \oplus z)\)
  3. neutral element: \(e = C[1, 1, \ldots, 1] = \text{barycentre of } S^D\)
  4. inverse of \(x\): \(x^{-1} = C[x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}]\)

  \[\Rightarrow \quad x \oplus x^{-1} = e \quad \text{and} \quad x \oplus y^{-1} = x \odot y\]

- **properties of powering**
  1. associativity: \(\alpha \odot (\beta \odot x) = (\alpha \cdot \beta) \odot x\)
  2. distributivity 1: \(\alpha \odot (x \oplus y) = (\alpha \odot x) \oplus (\alpha \odot y)\)
  3. distributivity 2: \((\alpha + \beta) \odot x = (\alpha \odot x) \oplus (\beta \odot x)\)
  4. neutral element: \(1 \odot x = x\)
complete inner product space structure

inner product:  $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$,  \(\mathbf{x}, \mathbf{y} \in S^D\)

norm:  $\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left( \ln \frac{x_i}{x_j} \right)^2}$,  \(\mathbf{x} \in S^D\)

distance:  $d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left( \ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}$,  \(\mathbf{x}, \mathbf{y} \in S^D\)

Aitchison geometry on the simplex

\((S^D, \oplus, \odot)\) is a \((D - 1)\)-dim. Euclidean space
properties of the Aitchison geometry

distance and perturbation: \( d_a(p \oplus x, p \oplus y) = d_a(x, y) \)

distance and powering: \( d_a(\alpha \odot x, \alpha \odot y) = |\alpha|d_a(x, y) \)

compositional lines: \( y = x_0 \oplus (\alpha \odot x) \)
\( (x_0 = \text{starting point}, \ x = \text{leading vector}) \)

orthogonal lines: \( y_1 = x_0 \oplus (\alpha_1 \odot x_1), y_2 = x_0 \oplus (\alpha_2 \odot x_2), \)
\[ y_1 \perp y_2 \iff \langle x_1, x_2 \rangle_a = 0 \]

(the inner product of the leading vectors is zero)

parallel lines: \( y_1 = x_0 \oplus (\alpha \odot x) \parallel y_2 = p \oplus x_0 \oplus (\alpha \odot x) \)
orthogonal compositional lines

orthogonal grids in $S^3$, equally spaced, 1 unit in Aitchison distance; the right grid is rotated $45^\circ$ with respect to the left grid
ellipses and shifted segments
advantages of complete inner product spaces

- **orthonormal basis** can be constructed: \( \{e_1, \ldots, e_{D-1}\} \)
- **coordinates obey the rules** of real Euclidean space:
  \[ x \in S^D \Rightarrow y = [y_1, \ldots, y_{D-1}] \in \mathbb{R}^{D-1}, \text{ with } y_i = \langle x, e_i \rangle_a \]
- **standard methods** can be directly applied to coordinates
- **expressing results as compositions is easy**:
  if \( h : S^D \mapsto \mathbb{R}^{D-1} \) assigns to each \( x \in S^D \) its coordinates, i.e. \( h(x) = y \), then
  \[ h^{-1}(y) = x = \bigoplus_{i=1}^{D-1} y_i \odot e_i \]

**PRINCIPLE OF WORKING ON COORDINATES**
conclusions

- the Aitchison geometry of the simplex offers a powerful tool to analyse CoDa

- the geometry is apparently complex, but it is completely equivalent to standard Euclidean geometry in real space

- the key is to use a proper representation in coordinates


Aitchison, J. and J. J. Egozcue: Compositional data analysis: where are we and where should we be heading?, Mathematical Geology, 37, 7, 833-854, 2005.

