Statistical analysis of compositional data

G. Mateu-Figueras

Dep. d'Informàtica, Matemàtica Aplicada i Estadística
Universitat de Girona

February 26, 2014
Outline

1. compositional data
2. Aitchison geometry of the simplex
3. exploratory analysis
4. distributions on $S^D$
5. conclusions
compositional data

- Compositional data are parts of some whole which only carry relative information.
- The simplex (for $\kappa$ a constant)

$$S^D = \left\{ x = (x_1, \ldots, x_D) \in \mathbb{R}^D \mid x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}$$

- Standard representation for $D = 3$: ternary diagram
compositional data

- **compositional data** are parts of some whole which only carry relative information
- the **simplex** (for $\kappa$ a constant)

\[
S^D = \left\{ x = (x_1, \ldots, x_D) \in \mathbb{R}^D \mid x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}
\]

- standard representation
  for $D = 3$: ternary diagram
compositional data

- **compositional data** are parts of some whole which only carry **relative information**
- the **simplex** (for \( \kappa \) a constant)

\[
S^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D \mid x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}
\]

- standard representation for \( D = 3 \): **ternary diagram**
some compositional problems

- **MN blood system**: frequencies of MM, NN and MN blood types and the ethnic population. Despite the high variability, is there any stability in the data? Do they follow any genetic law?

- **Elections to the Parliament de Catalunya**: the total votes achieved by each party in each county. To characterize the regions.

- **Skye lavas**: relative proportions of $A$ ($Na_2O + K_2O$), $F$ ($Fe_2O_3$) and $M$ ($MgO$) of 23 basalt specimens from the Isle of Skye. To describe the variability of the geochemical composition.
some compositional problems

- **MN blood system**: frequencies of MM, NN and MN blood types and the ethnic population. Despite the high variability, is there any stability in the data? Do they follow any genetic law?

- **Elections to the Parlament de Catalunya**: the total votes achieved by each party in each county. To characterize the regions.

- **Skye lavas**: relative proportions of A ($Na_2O + K_2O$), F ($Fe_2O_3$) and M (MgO) of 23 basalt specimens from the Isle of Skye. To describe the variability of the geochemical composition.
some compositional problems

- **MN blood system:** frequencies of MM, NN and MN blood types and the ethnic population. Despite the high variability, is there any stability in the data? do they follow any genetic law?

- **elections to the Parlament de Catalunya:** the total votes achieved by each party in each counties. To characterize the regions.

- **skye lavas:** relative proportions of A \((Na_2O + K_2O)\), F \((Fe_2O_3)\) and M \((MgO)\) of 23 basalt specimens from the Isle of Skye. To describe the variability of the geochemical composition.
Spurious correlations (Pearson, 1897)

\[
x = (x_1, \ldots, x_D) \quad \sum_{i=1}^{D} x_i = \kappa \quad \text{cov}(x_i, x_1) + \cdots + \text{cov}(x_i, x_D) = 0
\]

<table>
<thead>
<tr>
<th>sample</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>0.2</td>
<td>0.3</td>
<td>0.3</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>\text{cov} ( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>0.007</td>
<td>0.003</td>
<td>0.000</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.003</td>
<td>0.002</td>
<td>-0.002</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0.000</td>
<td>-0.002</td>
<td>0.009</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>-0.010</td>
<td>-0.003</td>
<td>-0.007</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>\text{corr} ( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>1.000</td>
<td>0.866</td>
<td>0.000</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.866</td>
<td>1.000</td>
<td>-0.500</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0.000</td>
<td>-0.500</td>
<td>1.000</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>-0.866</td>
<td>-0.500</td>
<td>-0.500</td>
</tr>
</tbody>
</table>
difficulties

spurious correlations (Pearson, 1897)

\[ \mathbf{x} = (x_1, \ldots, x_D) \quad \sum_{i=1}^{D} x_i = \kappa \quad \text{cov}(x_i, x_1) + \cdots + \text{cov}(x_i, x_D) = 0 \]

<table>
<thead>
<tr>
<th>sample</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>0.2</td>
<td>0.3</td>
<td>0.3</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>\text{cov}</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>0.007</td>
<td>0.003</td>
<td>0.000</td>
<td>-0.010</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.003</td>
<td>0.002</td>
<td>-0.002</td>
<td>-0.003</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0.000</td>
<td>-0.002</td>
<td>0.009</td>
<td>-0.007</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>-0.010</td>
<td>-0.003</td>
<td>-0.007</td>
<td>0.020</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>\text{corr}</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>1.000</td>
<td>0.866</td>
<td>0.000</td>
<td>-0.866</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.866</td>
<td>1.000</td>
<td>-0.500</td>
<td>-0.500</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0.000</td>
<td>-0.500</td>
<td>1.000</td>
<td>-0.500</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>-0.866</td>
<td>-0.500</td>
<td>-0.500</td>
<td>1.000</td>
</tr>
</tbody>
</table>
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></td>
<td>1.00</td>
<td>-0.87</td>
<td>-0.65</td>
</tr>
<tr>
<td>$x_3$</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.19</td>
</tr>
<tr>
<td>$x_4$</td>
<td></td>
<td></td>
<td></td>
<td>1.00</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></td>
<td>1.00</td>
<td>-0.79</td>
</tr>
<tr>
<td>$x_3^*$</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>
subcompositional incoherence (Aitchison, 1997)

**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>
subcompositional incoherence (Aitchison, 1997)

**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></td>
<td>-0.79</td>
</tr>
<tr>
<td>$x_3^*$</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>
principles

- **scale invariance**: the analysis should not depend on the closure constant $\kappa$

  $$f(\alpha x) = f(x), \quad \alpha > 0$$

- **subcompositional coherence**: studies performed on subcompositions should not stand in contradiction with those performed on the full composition
principles

- **scale invariance**: the analysis should not depend on the closure constant $\kappa$

  $$f(\alpha x) = f(x), \quad \alpha > 0$$

- **subcompositional coherence**: studies performed on subcompositions should not stand in contradiction with those performed on the full composition
Euclidean space structure of $S^D$

for $\mathbf{x}, \mathbf{y} \in S^D$, $\alpha \in \mathbb{R}$, and $C$ is the closure operation

- **perturbation**: $\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, \ldots, x_D y_D)$
- **powering**: $\alpha \odot \mathbf{x} = C(x_1^\alpha, \ldots, x_D^\alpha)$

- **inner product**:

  $$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

- **associated norm and distance**:

  $$\| \mathbf{x} \|^2_a = \frac{1}{D} \sum_{i<j} \left( \ln \frac{x_i}{x_j} \right)^2 ; \quad d_a^2(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{i<j} \left( \ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2$$
Euclidean space structure of $S^D$

for $\mathbf{x}, \mathbf{y} \in S^D$, $\alpha \in \mathbb{R}$, and $C$ is the closure operation

- **perturbation:** $\mathbf{x} \oplus \mathbf{y} = C(x_1y_1, \ldots, x_Dy_D)$
- **powering:** $\alpha \odot \mathbf{x} = C(x_1^\alpha, \ldots, x_D^\alpha)$
- **inner product:**

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

- **associated norm and distance:**

$$\|\mathbf{x}\|_a^2 = \frac{1}{D} \sum_{i<j} \left( \ln \frac{x_i}{x_j} \right)^2$$

$$d_a^2(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{i<j} \left( \ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2$$
orthonormal coordinates

- **orthonormal basis** on $S^D$: $\{e_1, e_2, \ldots, e_{D-1}\}$ (not unique)
- **coordinates** in this basis for $x \in S^D$ or **ilr** coordinates $x^* = (\langle x, e_1 \rangle a, \ldots, \langle x, e_{D-1} \rangle a)$

Example:

$e_1 = C(\exp(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}})), \quad e_2 = C(\exp(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0))$

$x^* = \begin{pmatrix} \sqrt{\frac{2}{3}} \ln \left(\frac{x_1 \cdot x_2}{x_3}\right)^{1/2}, & \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2} \end{pmatrix}$

Egozcue et al. (2003)

- Compositional operations are reduced to ordinary vector operations when representing compositions by their coordinates
- **The principle of working on coordinates**
orthonormal coordinates

- **orthonormal basis** on $S^D$: \{e_1, e_2, \ldots, e_{D-1}\} (not unique)
- **coordinates** in this basis for $x \in S^D$ or **ilr coordinates**
  \[ x^* = (\langle x, e_1 \rangle a, \ldots, \langle x, e_{D-1} \rangle a) \]
- **example:**
  \[ e_1 = C(\exp(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}})), \quad e_2 = C(\exp(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0)) \]
  \[ x^* = \left( \sqrt{\frac{2}{3}} \ln \frac{(x_1 \cdot x_2)^{1/2}}{x_3}, \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2} \right) \]
  Egozcue et al. (2003)

- compositional operations are reduced to ordinary vector operations when representing compositions by their coordinates
- **the principle of working on coordinates**
orthonormal coordinates

- **orthonormal basis** on $S^D$: $\{e_1, e_2, \ldots, e_{D-1}\}$ (not unique)
- **coordinates** in this basis for $x \in S^D$ or **ilr coordinates**
  
  $x^* = (\langle x, e_1 \rangle, \ldots, \langle x, e_{D-1} \rangle)$

- example:
  
  $e_1 = C(\exp(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}})), \quad e_2 = C(\exp(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0))$

  $x^* = \left(\sqrt{2} \ln \left(\frac{x_1 \cdot x_2}{x_3}\right)^{1/2}, \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}\right)$

  Egozcue et al. (2003)

- compositional operations are reduced to ordinary vector operations when representing compositions by their coordinates
- **the principle of working on coordinates**
parallel lines

in $S^3$

coordinate representation
circles and ellipses

in $S^3$

coordinate representation
the MN blood system

\[ \sqrt{\frac{2}{3}} \ln \left( \frac{MM \cdot NN}{MN} \right)^{1/2} = -0.57 \]
the MN blood system

Hardy-Weinberg law: $MN^2 = 4MM \cdot NN$

$$\sqrt{\frac{2}{3}} \ln \left( \frac{MM \cdot NN}{MN} \right)^{1/2} = -0.57$$
building an orthonormal basis using sequential binary partitions (SBP)

**example:** sequential binary partition for $\mathbf{x} \in S^5$; 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>$x_1^* = \sqrt{\frac{3 \cdot 2}{3+2}} \ln \left( \frac{x_1 \cdot x_3 \cdot x_4}{x_2 \cdot x_5} \right)^{1/3}$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>−1</td>
<td>$x_2^* = \sqrt{\frac{1 \cdot 1}{1+1}} \ln \frac{x_2}{x_5}$</td>
</tr>
<tr>
<td>3</td>
<td>+1</td>
<td>0</td>
<td>−1</td>
<td>−1</td>
<td>0</td>
<td>$x_3^* = \sqrt{\frac{1 \cdot 2}{1+2}} \ln \left( \frac{x_1}{x_3 \cdot x_4} \right)^{1/2}$</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>−1</td>
<td>0</td>
<td>$x_4^* = \sqrt{\frac{1 \cdot 1}{1+1}} \ln \frac{x_3}{x_4}$</td>
</tr>
</tbody>
</table>
coordinates $\Rightarrow$ balances

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

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

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 $S_i$.

Egozcue, Pawlowsky-Glahn (2005)
Log-ratio approach (Aitchison, 1980-86)

**log-ratio** transformations introduced by J. Aitchison:

- **alr**: $S^D \rightarrow \mathbb{R}^{D-1}$, $\text{alr}(x) = \left( \ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right)$

  drawback: not an isometry

- **clr**: $S^D \rightarrow \mathbb{R}^D$, $\text{clr}(x) = \left( \ln \frac{x_1}{g(x)}, \ldots, \ln \frac{x_D}{g(x)} \right)$,

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

  drawback: a constrained transformed vector
log-ratio transformations introduced by J. Aitchison:

- **alr**: $S^D \rightarrow \mathbb{R}^{D-1}$, 
  $$\text{alr}(x) = \left( \ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right)$$

  drawback: not an isometry

- **clr**: $S^D \rightarrow \mathbb{R}^D$, 
  $$\text{clr}(x) = \left( \ln \frac{x_1}{g(x)}, \ldots, \ln \frac{x_D}{g(x)} \right)$$,

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

  drawback: a constrained transformed vector
the treatment of zeros

case 1: the part with zeros is not important for the study
⇒ the part should be omitted

case 2: the part is important, the zeros are essential
⇒ divide the sample into two or more populations, according to the presence/absence of zeros

case 3: the part is important, the zeros are rounded zeros
⇒ use imputation techniques

for a review, see Martín-Fernández et al. (2011)
center and variability

let \( X = \{x_i = (x_{i1}, \ldots, x_{iD}) \in S^D : i = 1, \ldots, n\} \)

- **center** (closed geometric mean) of \( X \):
  \[
g = C(g_1, g_2, \ldots, g_D), \quad \text{with } g_j = \left( \prod_{i=1}^{n} x_{ij} \right)^{1/n}
  \]

- **total variance** of \( X \):
  \[\text{TotVar}[X] = \frac{1}{n} \sum_{i=1}^{n} d^2_a(x_i, g)\]

- **variation array** of \( X \):
  \[
  \begin{pmatrix}
    - & \text{var} \left[ \ln \frac{x_1}{x_2} \right] & \cdots & \text{var} \left[ \ln \frac{x_1}{x_D} \right] \\
    E \left[ \ln \frac{x_1}{x_2} \right] & - & \cdots & \vdots \\
    \vdots & \ddots & - & \text{var} \left[ \ln \frac{x_{D-1}}{x_D} \right] \\
    E \left[ \ln \frac{x_1}{x_D} \right] & \cdots & E \left[ \ln \frac{x_{D-1}}{x_D} \right] & -
  \end{pmatrix}
  \]
center and variability

let $X = \{x_i = (x_{i1}, \ldots, x_{iD}) \in S^D : i = 1, \ldots, n\}$

- **center** (closed geometric mean) of $X$:
  \[
g = C(g_1, g_2, \ldots, g_D), \text{ with } g_j = \left(\prod_{i=1}^{n} x_{ij}\right)^{1/n}
\]

- **total variance** of $X$:
  \[
  \text{TotVar}[X] = \frac{1}{n} \sum_{i=1}^{n} d_a^2(x_i, g)
\]

- **variation array** of $X$:

\[
\begin{pmatrix}
- & \text{var} [\ln \frac{x_1}{x_2}] & \cdots & \text{var} [\ln \frac{x_1}{x_D}] \\
E [\ln \frac{x_1}{x_2}] & - & \cdots & \vdots \\
\vdots & \ddots & \ddots & - \\
E [\ln \frac{x_1}{x_D}] & \cdots & E [\ln \frac{x_{D-1}}{x_D}] & -
\end{pmatrix}
\]
center and variability

let \( \mathbf{X} = \{ \mathbf{x}_i = (x_{i1}, \ldots, x_{iD}) \in \mathcal{S}^D : i = 1, \ldots, n \} \)

- **center** (closed geometric mean) of \( \mathbf{X} \):
  
  \[
  \mathbf{g} = C(g_1, g_2, \ldots, g_D), \quad \text{with } g_j = \left( \prod_{i=1}^{n} x_{ij} \right)^{1/n}
  \]

- **total variance** of \( \mathbf{X} \):
  \[
  \text{TotVar}[\mathbf{X}] = \frac{1}{n} \sum_{i=1}^{n} d^2_\mathcal{S}(\mathbf{x}_i, \mathbf{g})
  \]

- **variation array** of \( \mathbf{X} \):
  
  \[
  \begin{pmatrix}
  - & \text{var}\left[\ln \frac{x_1}{x_2}\right] & \cdots & \text{var}\left[\ln \frac{x_1}{x_D}\right] \\
  \text{E}\left[\ln \frac{x_1}{x_2}\right] & - & \cdots & \vdots \\
  \vdots & \ddots & - & \text{var}\left[\ln \frac{x_{D-1}}{x_D}\right] \\
  \text{E}\left[\ln \frac{x_1}{x_D}\right] & \cdots & \text{E}\left[\ln \frac{x_{D-1}}{x_D}\right] & -
  \end{pmatrix}
  \]}
example: ParlCat2010 data set

votes achieved by PP, CiU, SI, C’s, ERC, PSC, ICV

\[ \mathbf{g} = (0.097, 0.505, 0.044, 0.017, 0.102, 0.179, 0.056) \]
clr biplot

- graphical display of a multivariate data set (individuals and variables)
- clr-biplot
- particular rules of interpretation
  - $\|ray\|$ $\approx$ variance clr component
  - $\|link\|$ $\approx$ variance logratio
  - perpendicular links $\Rightarrow$ possible incorrelated logratios
  - parallel links $\Rightarrow$ possible high correlated logratios
  - coincident vertices $\Rightarrow$ two redundant parts
  - collinear vertices $\Rightarrow$ possible one-dimensional variability

Aitchison and Greenacre (2002)
**example: ParlCat2010 data set**

(explains 86% variance)

\[ Z_{dr} = U(VT)^T \]

\[ \text{var} \left( \ln \left( \frac{ICV}{g} \right) \right) = 0.0417 \]

\[ \text{var} \left( \ln \left( \frac{C's}{g} \right) \right) = 0.2898 \]
example: ParlCat2010 data set

(var \left( \ln \left( \frac{ICV}{g} \right) \right) = 0.0417 \quad var \left( \ln \left( \frac{C's}{g} \right) \right) = 0.2898
example: ParlCat2010 data set
(explains 86% variance)

\[ \text{var} \left( \ln \left( \frac{S_I}{C'} \right) \right) = 0.8915 \]
\[ \text{var} \left( \ln \left( \frac{CiU}{ERC} \right) \right) = 0.0732 \]
example: ParlCat2010 data set

(explains 86% variance)

\[ corr \left( \ln \left( \frac{C'}{ERC} \right), \ln \left( \frac{PSC}{ICV} \right) \right) = -0.041 \]
example: ParlCat2010 data set
(explains 86% variance)

$Z_{clr} = U(V)^T$
to visualize
- sequential binary **partition**
- **center** of each balance
- proportion of the sample total **variance** corresponding to each balance.
- **summary statistics** of each balance (box-plot of percentiles 5, 25, 50, 75, 95)
- adequate to represent different **groups**

Pawlowsky-Glahn and Egozcue (2011)
example: ParlCat2010 data set
example: ParlCat2010 data set
logistic normal (Aitchison 1980-86)

\[ x : \Omega \rightarrow S^D \]

- **transform** \( x \) to \( \mathbb{R}^{D-1} \) using a log-ratio transformation
- define the density of the **transformed vector** and go back to \( S^D \) using the **change of variable** theorem
- the result is a density function for \( x \) with respect to \( \lambda \) on \( S^D \)

\[ (\text{Aitchison, 1997}) \]

\( E[x] \) is not a meaningful measure of central location
\( \text{cen}[x] \) is the alternative which minimizes \( E[d_a^2(x, \text{cen}[x])] \)
logistic normal (Aitchison 1980-86)

\[ x : \Omega \rightarrow S^D \]

- **transform** \( x \) to \( \mathbb{R}^{D-1} \) using a log-ratio transformation
- define the density of the **transformed vector** and go back to \( S^D \) using the **change of variable** theorem
- the result is a density function for \( x \) with respect to \( \lambda \) on \( S^D \)

\[ (Aitchison, 1997) \]

\( E[x] \) is not a meaningful measure of central location
\( cen[x] \) is the alternative which minimizes \( E[d_a^2(x, cen[x])] \)
densities and measures

- on $S^D$: density functions expressed with respect to the Aitchison measure $\lambda_a$
- density functions of the vector of coordinates with respect to $\lambda$.

$$d\lambda/d\lambda_a = \sqrt{D} x_1 x_2 \cdots x_D, \quad \lambda_a(A) = \lambda(A^*)$$
densities and measures

- on $S^D$: density functions expressed with respect to the Aitchison measure $\lambda_a$
- density functions of the vector of coordinates with respect to $\lambda$.

$$\frac{d\lambda}{d\lambda_a} = \sqrt{D} \ x_1 x_2 \cdots x_D, \quad \lambda_a(A) = \lambda(A^*)$$
normal on $S^D$

\[ x : \Omega \rightarrow S^D \]

a random composition $x$ is \textbf{normally distributed on} $S^D$ with parameters $\mu$ and $\Sigma$ if its density function is

\[
f_x(x) = (2\pi)^{-(D-1)/2}|\Sigma|^{-1/2} \exp \left[-\frac{1}{2} (x^* - \mu^*)' \Sigma^{-1} (x^* - \mu^*) \right]
\]

\[
\text{usual normal density applied to coordinates } x^* \text{ and } f_x = \frac{dP}{d\lambda_a}
\]

\[
\mu = E_a[x] = \text{cen}[x]
\]

Mateu-Figueras et al (2013)
normal on $S^D$

A random composition $\mathbf{x}$ is **normally distributed on $S^D$** with parameters $\mu$ and $\Sigma$ if its density function is

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-(D-1)/2} |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2} (\mathbf{x}^* - \mu^*)' \Sigma^{-1} (\mathbf{x}^* - \mu^*) \right]$$

usual **normal density** applied to coordinates $\mathbf{x}^*$ and $f_{\mathbf{x}} = \frac{dP}{d\lambda_a}$

$$\mu = \mathbb{E}_a[\mathbf{x}] = \text{cen}[\mathbf{x}]$$

Mateu-Figueras et al (2013)
normal on $S^D$

A random composition $\mathbf{x}$ is **normally distributed on** $S^D$ with parameters $\mu$ and $\Sigma$ if its density function is

$$f_\mathbf{x}(\mathbf{x}) = (2\pi)^{-(D-1)/2} |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2} (\mathbf{x}^* - \mu^*)' \Sigma^{-1} (\mathbf{x}^* - \mu^*) \right]$$

usual **normal density** applied to coordinates $\mathbf{x}^*$ and $f_\mathbf{x} = \frac{dP}{d\lambda_a}$

$$\mu = E_a[\mathbf{x}] = cen[\mathbf{x}]$$

Mateu-Figueras et al (2013)
comparison

\[ \mu^* = (0, 0), \Sigma = \text{Id} \]

\[ S^D \subset \mathbb{R}^D \]

\[ S^D \text{ as Euclidian space} \]

logistic normal
Lebesgue measure \( \lambda \)

normal on \( S^D \)
Aitchison measure \( \lambda_a \)
Invariance under perturbation

\[ p = (0.93, 0.05, 0.02) \]

\[ x^* = \left( \frac{1}{\sqrt{2}} \ln \left( \frac{x_1}{x_2} \right), \frac{1}{\sqrt{6}} \ln \left( \frac{x_1 x_2}{x_3} \right) \right) \]

\[ \mu^* = (-0.5, -0.5), \quad \mu^* = (1.5, 1.5), \quad \Sigma = I_d \]
tests of normality on $S^D$

$H_0$: the sample of coordinates comes from a multivariate normal distribution

- based on empirical distribution function (EDF) tests
- Anderson-Darling, Cramer-von Mises and Watson statistics
- three possible cases
  - all $(D - 1)$ marginal, univariate distributions
  - all $(D - 1)(D - 2)/2$ bivariate angle distributions
  - the $(D - 1)$-dimensional radius distribution
- problem: dependence of the orthonormal basis
tests of normality on $S^D$

$H_0$: the sample of coordinates comes from a multivariate normal distribution

- based on empirical distribution function (EDF) tests
- Anderson-Darling, Cramer-von Mises and Watson statistics
- three possible cases
  - all $(D - 1)$ marginal, univariate distributions
  - all $(D - 1)(D - 2)/2$ bivariate angle distributions
  - the $(D - 1)$-dimensional radius distribution

problem: dependence of the orthonormal basis
tests of normality on $S^D$

$H_0$: the sample of coordinates comes from a multivariate normal distribution

- based on empirical distribution function (EDF) tests
- Anderson-Darling, Cramer-von Mises and Watson statistics
- three possible cases
  - all $(D - 1)$ marginal, univariate distributions
  - all $(D - 1)(D - 2)/2$ bivariate angle distributions
  - the $(D - 1)$-dimensional radius distribution
- problem: dependence of the orthonormal basis
example: aphyric Skye lavas

\( X=(A,F,M) \) composition of 23 basalt specimens from the Isle of Skye (Aitchison, 1986)

\[
\hat{\mu}^* = (0.555, 0.639) \quad \hat{\Sigma} = \begin{pmatrix} 0.126 & -0.229 \\ -0.229 & 0.456 \end{pmatrix}
\]
kernel density estimation

- the normal on $S^D$ for the kernel in the density estimator
- invariance with respect to the orthonormal basis

Chacón et al (2010)
other distributions on $S^D$

- the skew-normal distribution on $S^D$
- the Dirichlet distribution
- the shifted-scaled Dirichlet distribution
- ...
conclusions

- treat compositional data (CoDa) in the **simplex**, with its specific geometry
- do **not** apply ordinary multivariate statistics **directly** to CoDa
- the simplex has an **Euclidean** structure: **orthonormal coordinates** are available
- multivariate statistical models and methods **work properly** on coordinates of CoDa
- problem (or advantage): **interpretation** of coordinates
references