The statistical analysis of compositional data:
The Aitchison geometry

Prof. Dr. Vera Pawlowsky-Glahn
Prof. Dr. Juan José Egozcue
Ass. Prof. Dr. René Meziat

Instituto Colombiano del Petróleo
Piedecuesta, Santander, Colombia
March 20–23, 2007
IAMG Distinguished Lecturer – 2007

Prof. Dr. Vera Pawlowsky-Glahn

Department of Computer Science and Applied Mathematics
University of Girona, Spain
compositional data are parts of some whole which only carry relative information

usual units of measurement: parts per unit, percentages, ppm, ppb, concentrations, ...

historically: data subject to a constant sum constraint

examples: geochemical analysis; (sand, silt, clay) composition; proportions of minerals in a rock; ...
historical remarks: end of the XIXth century

Karl Pearson, 1897: “On a form of spurious correlation which may arise when indices are used in the measurement of organs”

- he was the first to point out dangers that may befall the analyst who attempts to interpret correlations between ratios whose numerators and denominators contain common parts

- the closure problem was stated within the framework of classical statistics, and thus within the framework of Euclidean geometry in real space
**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></td>
<td></td>
<td>1.00</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>
historical remarks: from 1897 to 1980 (and beyond)

- The fact that correlations between closed data are induced by numerical constraints caused Felix Chayes to attempt to separate the *spurious* part from the *real* correlation (“On correlation between variables of constant sum”, 1960)

- Many studied the effects of closure on methods related to correlation and covariance analysis (principal component analysis, partial and canonical correlation analysis) or distances (cluster analysis)

- An exhaustive search was initiated within the framework of classical (applied) statistics
historical remarks: end of the XXth century


- **key idea:** compositional data represent parts of some whole; they only carry *relative information*

- by analogy with the log-normal approach, Aitchison projected the sample space of compositional data, the $D$-part simplex $S^D$, to real space $\mathbb{R}^{D-1}$ or $\mathbb{R}^D$, using log-ratio transformations

- the *log-ratio approach* was born ...
compositional data: definition

definition: parts of some whole which carry only relative information \iff compositional data are equivalence classes

usual representation: subject to a constant sum constraint
**compositional data: usual representation**

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

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

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

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

other frequent units: ppm, ppb, ...

a **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 \( \mathbf{x} \)
the simplex as sample space

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

standard representation for \( D = 3 \):

the ternary diagram
**example 1: genetic hypothesis**

**data:** genotyps in the MN system of blood groups; **code:** Ab = Aborigines; Ch = Chinese; In = Indian; AmIn = American Indian; Es = Eskimo;

**question:** despite the high variability which can be observed, is there any inherent stability in the data? do they follow any genetic law?
requirements for a proper analysis

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

- **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 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 linear vector space

BUT Euclidean geometry 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 \( z = [z_1, z_2, \ldots, z_D] \in \mathbb{R}_+^D \)

\[
C [z] = \left[ \begin{array}{c} \frac{\kappa \cdot z_1}{\sum_{i=1}^{D} z_i} \\ \frac{\kappa \cdot z_2}{\sum_{i=1}^{D} z_i} \\ \vdots \\ \frac{\kappa \cdot z_D}{\sum_{i=1}^{D} z_i} \end{array} \right]
\]

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

\[
x \oplus y = C [x_1y_1, x_2y_2, \ldots, x_Dy_D]
\]

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

\[
\alpha \odot x = C [x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha]
\]
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 representant of the class
  - A generalization to other representants is possible

- For $z \in \mathbb{R}^D_+$ and $x \in S^D$, $x \oplus (\alpha \odot z) = x \oplus (\alpha \odot C[z])$
**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:** \(e = C[1, 1, \ldots, 1]\) = barycentre of \(S^D\)
4. **inverse of** \(x\): \(x^{-1} = C [x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}]\)

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

**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:** \(1 \odot x = x\)
inner product space structure of \((S^D, \oplus, \odot)\)

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

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

distance: \[
d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left( \ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}, \quad \mathbf{x}, \mathbf{y} \in S^D
\]
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
circles and other geometric figures
advantages of Euclidean 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
  \[
  h^{-1}(y) = x = \bigoplus_{i=1}^{D-1} y_i \odot e_i
  \]
conclusions

- The Aitchison geometry of the simplex offers a new 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.