evropský
sociální
fond v ČR

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání
pro konkurenceschopnost

INVESTICE
DO ROZVOJE
VZDĚLÁVÁNÍ

**Streamlining the Applied Mathematics Studies
at Faculty of Science of Palacký University in Olomouc
CZ.1.07/2.2.00/15.0243**

International Conference
Olomoucian Days of Applied Mathematics

ODAM 2013

Department of Mathematical analysis
and Applications of Mathematics

Faculty of Science

Palacký Univerzity Olomouc

The product space $\mathcal{T}$ (tools for compositional data with a total)

V. Pawlowsky-Glahn[1]; J. J. Egozcue[2]; D. Lovell[3]

[1]Dept. Informàtica i Matemàtica Aplicada, U. de Girona, España
(e-mail: vera.pawlowsky@udg.edu)
Past-President – International Association for Mathematical Geosciences

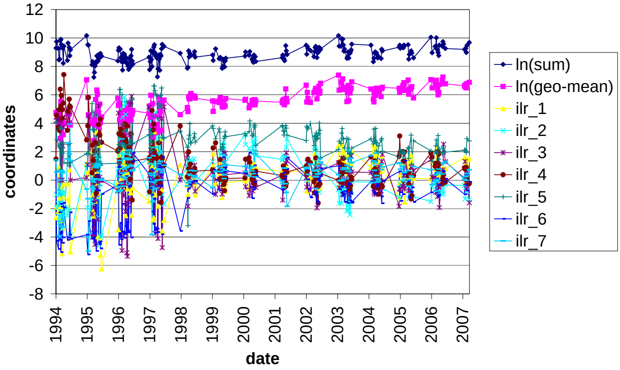
[2]Dept. Mat. Aplicada III, U. Politècnica de Catalunya, Barcelona, España

[3]CSIRO Mathematics, Informatics and Statistics, Canberra, Australia

Olomoucian Days of Applied Mathematics (ODAM) 2013
Olomouc, Česká republika, June 12–14, 2013

phytoplankton abundances: an example

- 173 samples, 8 taxa (cyanobacteria), period of 14 years



- groups: B = before 1998; A = after 1998; joint group = BA

background

- compositional data analysis deals with relative information between parts
- the total (abundances, mass, amount, ...) is in general not known or not informative
- when interest lies in analysing a composition for which the total is known and of interest, data are usually analysed as elements of $\mathbb{R}_+^D \subset \mathbb{R}^D$ or of $\mathbb{R}_+ \times \mathcal{S}^D = \mathcal{T} \subset \mathbb{R}^D$

questions

- is there a Euclidean space structure of $\mathbb{R}_+^D$ and of $\mathbb{R}_+ \times \mathcal{S}^D = \mathcal{T}$?
- if so, which is the relationship between the two spaces?
- which *total* is coherent with the representation in $\mathcal{S}^D$?

assumptions

- the Aitchison geometry is valid in the D-part simplex

$$\mathcal{S}^D = \left\{ [x_1, x_2, \ldots, x_D] \in \mathbb{R}^D_+ \;\middle|\; x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\} \subset \mathbb{R}^D_+$$

- the log-geometry is valid in $\mathbb{R}_+$, i.e. a log-transformed component of $\mathbb{R}_+$ corresponds to a Euclidean coordinate
- the total of $\mathbf{w} \in \mathbb{R}^D_+$ is a function, $t(\mathbf{w})$, of its components

procedure and goal

- induce a Euclidean space structure in both spaces through the product space
- study forms of total for inducing isomorphism and/or isometry between $\mathbb{R}^D_+$ and $\mathcal{T}$

Euclidean space structure of $\mathbb{R}_+^D$

notation: $\mathbf{w}, \mathbf{v} \in \mathbb{R}_+^D, \alpha \in \mathbb{R}$

plus-perturbation (Abelian inner group operation in $\mathbb{R}_+^D$)

$$\mathbf{w} \oplus_+ \mathbf{v} = [w_1 \cdot v_1, w_2 \cdot v_2, \ldots, w_D \cdot v_D]$$

plus-powering (external multiplication in $\mathbb{R}_+^D$)

$$\alpha \odot_+ \mathbf{w} = [w_1^\alpha, w_2^\alpha, \ldots, w_D^\alpha]$$

plus-inner-product

$$\langle \mathbf{w}, \mathbf{v} \rangle_+ = \langle \lg \mathbf{w}, \lg \mathbf{v} \rangle$$

plus-distance and plus-norm

$$d_+(\mathbf{w}, \mathbf{v}) = d(\lg \mathbf{w}, \lg \mathbf{v}) \; ; \quad \|\mathbf{w}\|_+ = \|\lg \mathbf{w}\|$$

properties of $(\mathbb{R}^D_+, \oplus_+, \odot_+, \langle\ ,\ \rangle_+)$

notation: $\mathbf{w}, \mathbf{v} \in \mathbb{R}^D_+$; $\alpha \in \mathbb{R}$; $\mathcal{C}\mathbf{w} = \mathbf{x}, \mathcal{C}\mathbf{v} = \mathbf{y} \in \mathcal{S}^D$; $\mathrm{g_m}\mathbf{w} = \left(\prod_{i=1}^D w_i\right)^{1/D}$

$\oplus_+$

- associative and commutative
- neutral element (identity): $\mathbf{n}_+ = [1, 1, \ldots, 1]$
- inverse element: $\ominus_+\mathbf{w} = [1/w_1, 1/w_2, \ldots, 1/w_D]$

$\odot_+$

- distributive with respect to the vector group operation $\alpha \odot_+ (\mathbf{w} \oplus_+ \mathbf{v}) = (\alpha \odot_+ \mathbf{w}) \oplus_+ (\alpha \odot_+ \mathbf{v})$
- distributive with respect to field addition $(\alpha + \beta) \odot_+ \mathbf{w} = (\alpha \odot_+ \mathbf{w}) \oplus_+ (\beta \odot_+ \mathbf{w})$
- compatible with field multiplication $\alpha \odot_+ (\beta \odot_+ \mathbf{w}) = (\alpha \cdot \beta) \odot_+ \mathbf{w}$
- identity element: $1 \odot_+ \mathbf{w} = \mathbf{w}$

$\langle , \rangle_+$

- $\langle \mathbf{w}, \mathbf{v} \rangle_+ = \langle \mathbf{x}, \mathbf{y} \rangle_a + D \lg(\mathrm{g_m}\mathbf{w}) \cdot \lg(\mathrm{g_m}\mathbf{v})$

$d_+^2(,)$

- $d_+^2(\mathbf{w}, \mathbf{v}) = d_a^2(\mathbf{x}, \mathbf{y}) + D \lg^2\left(\frac{\mathrm{g_m}\mathbf{w}}{\mathrm{g_m}\mathbf{v}}\right)$

Euclidean space structure of $\mathbb{R}_+ \times \mathcal{S}^D = \mathcal{T}$

notation: $\mathbf{w} \in \mathbb{R}_+^D$, $t(\mathbf{w}) \in \mathbb{R}_+$, $\mathcal{C}\mathbf{w} = \mathbf{x} \in \mathcal{S}^D$, $\tilde{\mathbf{x}} = (t(\mathbf{w}), \mathbf{x}) \in \mathcal{T}$, $\alpha \in \mathbb{R}$

$t(\mathbf{w})$ is a total (abundance, mass); it can be the sum, product, arithmetic or geometric mean, a single component, ...

$\mathcal{T}$-perturbation (Abelian inner group operation in $\mathcal{T}$)

$$\begin{aligned} \tilde{\mathbf{x}} \oplus_T \tilde{\mathbf{y}} &= (t(\mathbf{w}) \oplus_+ t(\mathbf{v}), \mathbf{x} \oplus_a \mathbf{y}) \\ &= [t(\mathbf{w}) \cdot t(\mathbf{v}), \mathcal{C}[x_1 y_1, x_2 y_2, \ldots, x_D y_D]] \end{aligned}$$

$\mathcal{T}$-powering (external multiplication in $\mathcal{T}$)

$$\alpha \odot_T \tilde{\mathbf{x}} = (\alpha \odot_+ t(\mathbf{w}), \alpha \odot_a \mathbf{x}) = [t(\mathbf{w})^\alpha, \mathcal{C}[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha]]$$

$\mathcal{T}$-inner-product $\langle \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \rangle_T = \langle t(\mathbf{w}), t(\mathbf{v}) \rangle_+ + \langle \mathbf{x}, \mathbf{y} \rangle_a$

properties of $(\mathcal{T}, \oplus_T, \odot_T, \langle\ ,\ \rangle_T)$ $\forall \tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\mathbf{z}} \in \mathcal{T},\ \alpha \in \mathbb{R}$

(a) (associative) $(\tilde{\mathbf{x}} \oplus_T \tilde{\mathbf{y}}) \oplus_T \tilde{\mathbf{z}} = \tilde{\mathbf{x}} \oplus_T (\tilde{\mathbf{y}} \oplus_T \tilde{\mathbf{z}})$

(b) (commutative) $\tilde{\mathbf{x}} \oplus_T \tilde{\mathbf{y}} = \tilde{\mathbf{y}} \oplus_T \tilde{\mathbf{x}}$

(c) (neutral element) $\tilde{\mathbf{x}} \oplus_T \tilde{\mathbf{n}} = \tilde{\mathbf{x}}$, with $\tilde{\mathbf{n}} = [1, \mathcal{C}[1, 1, \ldots, 1]]$

(d) (opposite element) opposite of $\tilde{\mathbf{x}} = [t(\mathbf{w}), x_1, x_2, \ldots, x_D]$,

$$\ominus_T \tilde{\mathbf{x}} = \left[\frac{1}{t(\mathbf{w})}, \mathcal{C}\left[\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_D}\right]\right] = (-1) \odot_T \tilde{\mathbf{x}}$$

(e) (distributive) $\alpha \odot_T (\tilde{\mathbf{x}} \oplus_T \tilde{\mathbf{y}}) = (\alpha \odot_T \tilde{\mathbf{x}}) \oplus_T (\alpha \odot_T \tilde{\mathbf{y}})$

(f) (unit) $1 \odot_T \tilde{\mathbf{x}} = \tilde{\mathbf{x}}$

inner product and squared distance

$$\langle \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \rangle_T = \lg(t(\mathbf{w})) \cdot \lg(t(\mathbf{v})) + \langle \mathbf{x}, \mathbf{y} \rangle_a$$

$$d_T^2(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = d_+^2(t(\mathbf{w}), t(\mathbf{v})) + d_a^2(\mathbf{x}, \mathbf{y}) = \lg^2 \frac{t(\mathbf{w})}{t(\mathbf{v})} + d_a^2(\mathbf{x}, \mathbf{y})$$

the *total* as a power of the product and as the sum of components

$h : \mathbb{R}^D_+ \to \mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$ such that

$h_p(\mathbf{w}) = (t_p(\mathbf{w}), \mathcal{C}\mathbf{w}) = \left(\left(\prod_{i=1}^D w_i\right)^\delta, \mathcal{C}\mathbf{w}\right)$	$h_s(\mathbf{w}) = (t_s(\mathbf{w}), \mathcal{C}\mathbf{w}) = \left(\sum_{i=1}^D w_i, \mathcal{C}\mathbf{w}\right)$
h_p is one-to-one (key for the proof: $\mathrm{g_m}\mathbf{x} = \mathrm{g_m}\mathbf{w}/(\sum_i w_i)$)	h_s is one-to-one
$h_p(\mathbf{w} \oplus_+ \mathbf{v}) = h_p(\mathbf{w}) \oplus_T h_p(\mathbf{v})$ $h_p(\alpha \odot_+ \mathbf{w}) = \alpha \odot_T h_p(\mathbf{w})$	$h_s(\mathbf{w} \oplus_+ \mathbf{v}) \neq h_s(\mathbf{w}) \oplus_T h_s(\mathbf{v})$ $h_s(\alpha \odot_+ \mathbf{w}) \neq \alpha \odot_T h_s(\mathbf{w})$
h_p is an isomorphism	**h_s is not an isomorphism**

distance properties

inequalities:

$$d_+^2(\mathbf{w},\mathbf{v}) = d_a^2(\mathbf{x},\mathbf{y}) + D\,\lg^2\left[\frac{\mathrm{g_m}\mathbf{w}}{\mathrm{g_m}\mathbf{v}}\right] \qquad \Rightarrow d_+ \geq d_a$$

$$d_{\mathcal{T}}^2(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = d_a^2(\mathbf{x},\mathbf{y}) + \lg^2\frac{t(\mathbf{w})}{t(\mathbf{v})} = d_+^2(\mathbf{w},\mathbf{v}) + \lg^2\frac{t(\mathbf{w})}{t(\mathbf{v})} - D\,\lg^2\left[\frac{\mathrm{g_m}\mathbf{w}}{\mathrm{g_m}\mathbf{v}}\right]$$

consequence: for $t = t_p$

- $\delta = 1/\sqrt{D} \;\Rightarrow\; d_{\mathcal{T}}^2 = d_+^2$
- $\delta > 1/\sqrt{D} \;\Rightarrow\; d_{\mathcal{T}}^2 > d_+^2$
- $\delta < 1/\sqrt{D} \;\Rightarrow\; d_{\mathcal{T}}^2 < d_+^2$

general result:

$h : \mathbb{R}^D_+ \to \mathcal{T}$, with $t_p(\mathbf{w}) = (\prod_i w_i)^{1/\sqrt{D}}$, $\mathbf{x} = \mathcal{C}\mathbf{w}$, is an isometry

induced structure in $\mathbb{R}_+^D$ by $\mathcal{T}_s = \mathbb{R}_+ \times \mathcal{S}^D$ (with the total = sum)

consider $h :\to \mathbb{R}_+^D \to \mathcal{T}_s = \mathbb{R}_+ \times \mathcal{S}^D$ such that
$h(\mathbf{w}) = (t_s(\mathbf{w}), \mathcal{C}\mathbf{w}) = (t_s(\mathbf{w}), \mathbf{x}) = \tilde{\mathbf{x}}$,
$h(\mathbf{v}) = (t_s(\mathbf{v}), \mathcal{C}\mathbf{v}) = (t_s(\mathbf{v}), \mathbf{y}) = \tilde{\mathbf{y}}$

- Abelian group operation:
$$\mathbf{w} \oplus_{+s} \mathbf{v} = h^{-1}(h(\mathbf{w}) \oplus_T h(\mathbf{v})) = \left[\ldots, \frac{t_s(\mathbf{w})t_s(\mathbf{v})}{\sum_{j=1}^D x_j y_j} x_i y_i, \ldots\right]$$
- external multiplication:
$$\alpha \odot_{+s} \mathbf{w} = h^{-1}(\alpha \odot_T h(\mathbf{w})) = \left[\ldots, \frac{t_s(\mathbf{w})^\alpha}{\sum_{j=1}^D x_j^\alpha} x_i^\alpha, \ldots\right]$$
- squared distance:
$$d_{+s}^2(\mathbf{w}, \mathbf{v}) = d_T^2(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \ln^2(t_s(\mathbf{w})/t_s(\mathbf{v})) + d_a^2(\mathbf{x}, \mathbf{y})$$

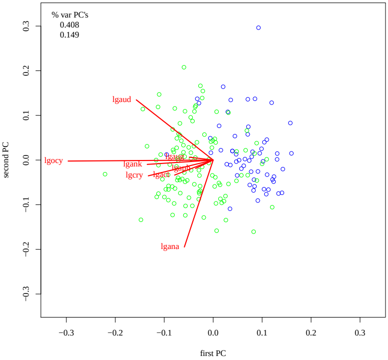
phytoplankton abundances in a river

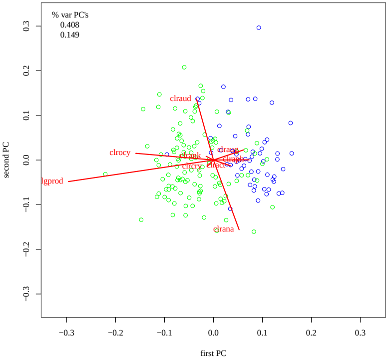
	Mvar			centre in $\mathbb{R}_+^D$							
	$\mathbb{R}_+^D, \mathcal{T}_p$	t_s	t_p	ana	aph	ocy	aug	aud	act	ank	cry
B	7.268	4244	7573866	627	491	94	2373	67	193	287	111
A	6.867	7530	69151799	957	671	856	3597	108	320	715	305
BA	8.705	6079	32945104	830	604	408	3129	92	270	526	218
	Mvar			**centres in $\mathcal{T}$**							
	$\mathcal{T}_s$	t_s	t_p	ana	aph	ocy	aug	aud	act	ank	cry
B	6.039	5456	7573866	.148	.116	.022	.559	.016	.046	.068	.026
A	5.311	9654	69151799	.127	.089	.114	.478	.014	.043	.095	.041
BA	6.241	7973	32945104	.137	.099	.067	.515	.015	.045	.087	.036

centres and metric variances for A, B, and BA samples;

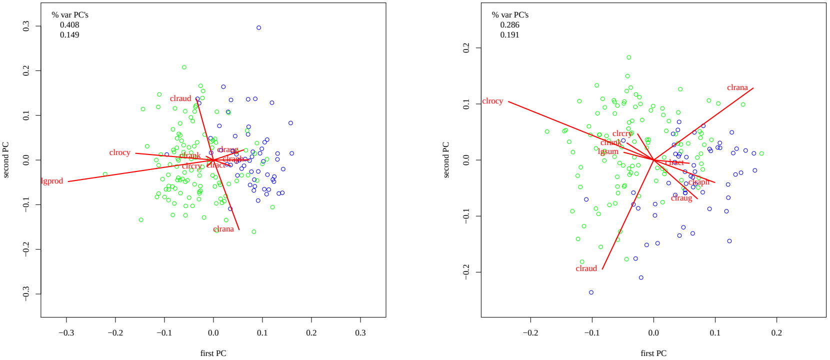
$$t_s = \sum_{i=1}^{D} w_i \; ; \qquad t_p = \left(\prod_{i=1}^{D} w_i \right)^{1/\sqrt{D}}$$

biplots in $\mathbb{R}^D_+$ and $\mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$





biplots in $\mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$



working in $\mathbb{R}^D_+$ and $\mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$

- allows easy comparison of structure and metrics
- the total as product of components leads to an isomorphism; the total as the product powered to $1/\sqrt{D}$ leads to an isometry
- the total as sum of components does not lead to an isomorphism; if the relevant total is the sum, it is not reasonable to work in $\mathbb{R}^D_+$ taking logarithms of the components