Statistical Analysis of Compositional Data

Carles Barceló Vidal

J. Antoni Martín Fernández

Santiago Thió Fdez-Henestrosa

Dept. d'Informàtica i Matemàtica Aplicada

Universitat de Girona

Campus de Montilivi

E-17071 Girona

Catalunya – Spain.

What is compositional data?

Traditionally,

Composition

| |

positive vector $\mathbf{x} = (x_1, \ldots, x_D)'$

whose components are subject to

a *constant sum restriction*:

$x_1 + \ldots + x_D = \text{constant}.$

Compositional data $\equiv$ *Closed data*

What is compositional data?

A positive vector $\mathbf{w} = (w_1, \ldots, w_D)'$ is *compositional* when our interest lies on the relative magnitudes w_j/w_k of its parts and not on the absolute values

Scale-invariance property

If a positive vector $\mathbf{w} = (w_1, \ldots, w_D)'$ is *compositional*, the vectors $\mathbf{w}$ and $k\mathbf{w}$, with $k > 0$, give us the same information

USA - President election - 2000

States	Bush	Gore	Others	Total
Alabama	943799	696741	26270	1666810
Alaska	136068	64252	30347	230667
⋮	⋮	⋮	⋮	⋮
Wisconsin	1235035	1240431	114415	2589881
Wyoming	147674	60421	5331	213426
Alabama	56,6%	41,8%	1,6%	100%
Alaska	59,0%	27,9%	13,1%	100%
⋮	⋮	⋮	⋮	100%
Wisconsin	47,7%	47,9%	4,4%	100%
Wyoming	69,2%	28,3%	2,5%	100%

Activity patterns of a statistician

Daily time (hours) devoted by an academic statistician to different activities: **te** = teaching; **co** = consultation; **ad** = administration; **re** = research; **ot** = other wakeful activities; **sl** = sleep.

D	te	co	ad	re	ot	sl	Total
1	3,5	2,0	4,5	2,5	6,5	5,0	24
2	4,0	2,0	2,5	3,0	6,5	6,0	24
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
19	2,5	2,5	3,0	2,0	5,0	8,5	24
20	2,5	2,0	3,0	3,0	4,0	9,0	24
1	14,4%	8,3%	18,8%	10,4%	27,1%	20,8%	100%
2	16,7%	8,3%	10,4%	2,5%	27,1%	25,0%	100%
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	100%
19	10,4%	10,4%	12,5%	8,3%	20,8%	35,4%	100%
20	10,5%	8,3%	12,5%	12,5%	16,7%	37,5%	100%

Artic lake

Sand, silt, clay composition (% by weight) of 39 sediment samples from an Artic lake

Sample	Sand	Silt	Clay	Total
S01	77.5	19.5	3.0	100%
S02	71.9	24.9	3.2	100%
⋮	⋮	⋮	⋮	⋮
S39	2.0	47.8	50.2	100%

Volcano H

Percentage of Cl, K_2O, P_2O_5, TiO_2 and SiO_2 in 46 samples of volcanic rocks from a volcano H

Sample	Cl	K_2O	P_2O_5	TiO_2	SiO_2	Total
1	0.0638	1.83	1.01	3.70	44.99	51.59%
2	0.1116	1.36	0.81	3.83	43.45	49.56%
3	0.0611	1.36	1.09	4.10	42.98	49.59%
⋮	⋮	⋮	⋮	⋮	⋮	⋮
44	0.0477	1.67	0.96	3.64	45.14	51.46%
45	0.0015	1.70	0.95	3.64	45.15	51.44%
46	0.0986	3.22	0.50	1.87	54.41	60.10%

Halimba boreholes

Percentages of Al_2O_3, SiO_2, Fe_2O_3, TiO_2, H_2O CaO and MgO in some samples from different drills in Halimba region (Hungary)

Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	CaO	MgO	Total
52,5	6,7	23,6	2,6	12,0	0,2	0,1	97,7%
47,7	4,6	32,1	2,3	12,0	2,0	0,0	100,7%
50,6	8,9	25,4	2,5	11,9	1,1	0,0	100,4%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The space of compositions

Any $D \times 1$ real vector $\mathbf{w} = (w_1, \ldots, w_D)'$ with positive components $w_1, \ldots, w_D$ will be called a *D-observational vector*.

Therefore, the set of these vectors will be $\mathbb{R}_+^D$, the positive orthant of $\mathbb{R}^D$.

Definition Two D-observational vectors $\mathbf{w}$ and $\mathbf{w}^*$ are *compositionally equivalent*, $\mathbf{w} \sim \mathbf{w}^*$, when there exists a positive proportionality constant k such that $\mathbf{w} = k\mathbf{w}^*$.

This relation classifies the vectors of $\mathbb{R}_+^D$ in classes of equivalence, called *D-compositions*.

The composition generated by an observational vector $\mathbf{w}$ will be symbolized by $\underline{\mathbf{w}}$, i.e.,

$$\underline{\mathbf{w}} = \{k\mathbf{w} : k \in \mathbb{R}^+\}.$$

Scale invariance

Definition A function f defined on $\mathbb{R}_+^D$ is said to be *scale invariant* if

$$f(k\mathbf{w}) = f(\mathbf{w}) \quad \text{for every } \mathbf{w} \in \mathbb{R}_+^D \text{ and } k \in \mathbb{R}^+$$

or equivalently

$$f(\mathbf{w}) = f(\mathbf{w}^*) \quad \text{when} \quad \mathbf{w} \sim \mathbf{w}^*.$$

Property Any scale invariant function $f(\mathbf{w})$ defined on $\mathbb{R}_+^D$ can be expressed in terms of ratios of the components $w_1, \ldots, w_D$ of $\mathbf{w}$, such as

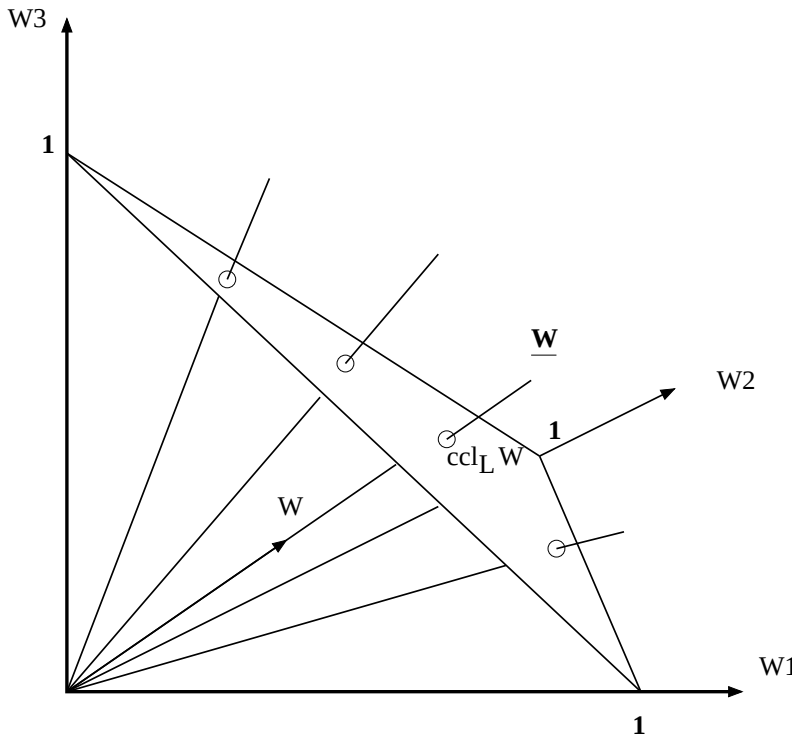
$$w_1/w_D, \ldots w_{D-1}/w_D \quad \text{or} \quad w_1/g(\mathbf{w}), \ldots, w_D/g(\mathbf{w}),$$

where $g(\mathbf{w}) = (w_1 w_2 \ \ldots \ w_D)^{1/D}$ is the geometric mean of the components of $\mathbf{w}$.

Property Any function defined on the compositional space $\mathcal{C}^{D-1}$ arises from a scale invariant function defined on the positive real space $\mathbb{R}_+^D$.

The space of compositions

A D-part composition can be geometrically interpreted as a ray from the origin in the positive orthant of $\mathbb{R}^D$:



The set $\mathcal{C}^{D-1}$ of all D-compositions will be called the $(D-1)$-dimensional *compositional space*.

The *compositional closure mapping* from $\mathbb{R}_+^D$ to $\mathcal{C}^{D-1}$ —denoted by ccl— is defined by

$$\mathrm{ccl}\,\mathbf{w} = \underline{\mathbf{w}} \qquad (\mathbf{w} \in \mathbb{R}_+^D).$$

Representation of a composition

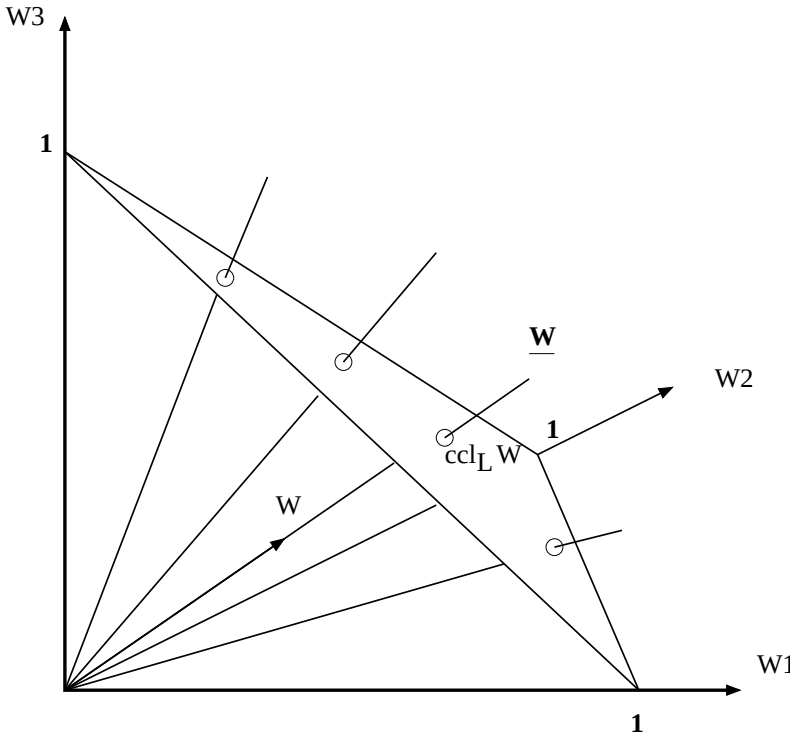
Linear criterion

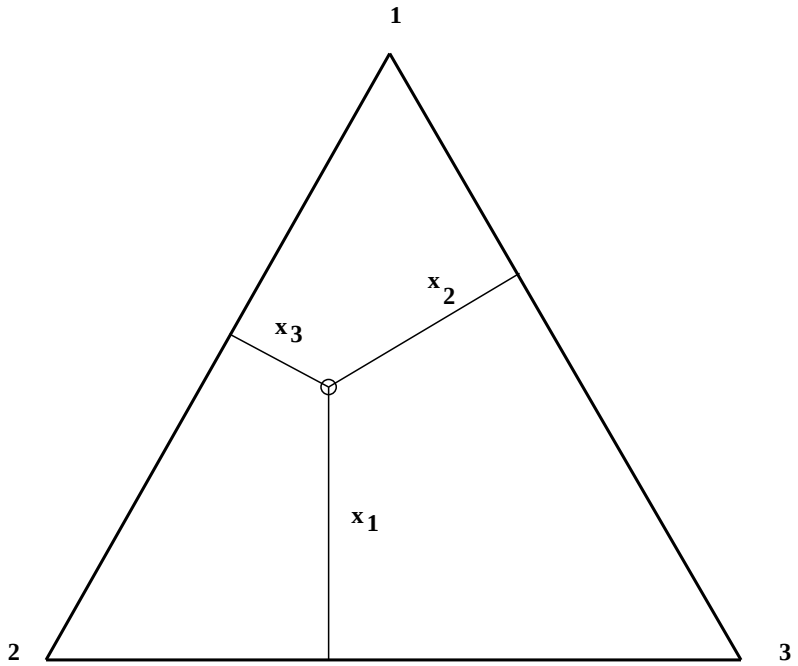
Definition The *linear criterion* selects from each D-composition $\underline{\mathbf{w}}$ the D-observational vector $\mathbf{w}^*$ —with components $w_1^*, \ldots, w_D^*$— whose sum is equal to 1. If this vector is symbolized by $\mathrm{ccl}_L\, \mathbf{w}$ —or by $\mathcal{C}\, \mathbf{w}$— then

$$\mathrm{ccl}_L\, \mathbf{w} = \mathcal{C}\, \mathbf{w} = \mathbf{w} / \sum_{j=1}^{D} w_j \qquad (\mathbf{w} \in \mathbb{R}_+^D).$$

The set of all the vectors $\mathbf{x} = \mathrm{ccl}_L\, \mathbf{w}$ $(\mathbf{w} \in \mathbb{R}_+^D)$ is the well-known $(D-1)$-dimensional *simplex* $\mathcal{S}^D$.

Linear criterion

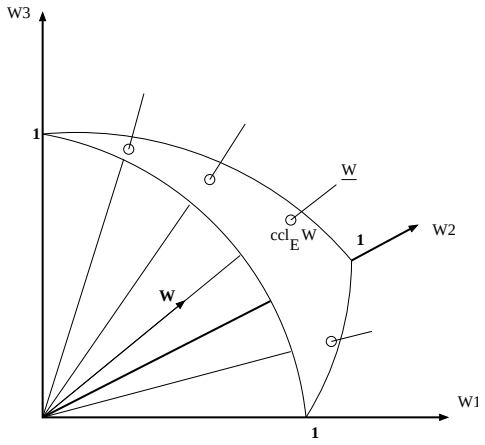




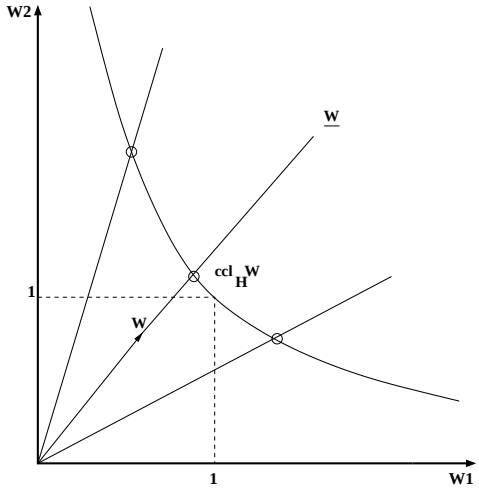
Representation of a composition

Other criteria

Spherical criterion



Hyperbolic criterion



Subcompositions

Sometimes, given a composition $\underline{\mathbf{w}}$ in $\mathcal{C}^{D-1}$, we may wish to focus attention on the relative magnitudes of a subset of components.

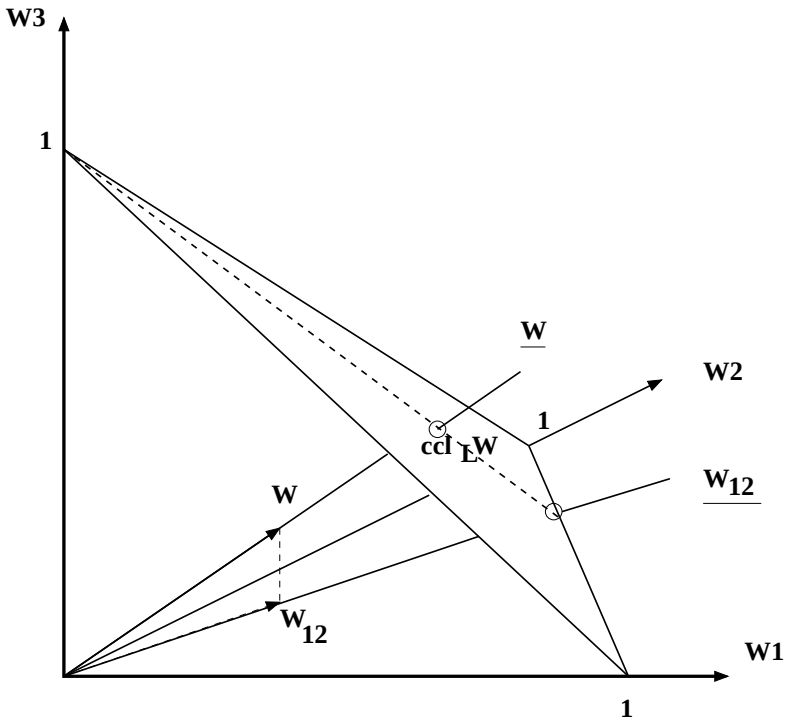
Definition If S is any subset of the indices $1, \ldots, D$ of a given a D-composition $\underline{\mathbf{w}} \in \mathcal{C}^{D-1}$, and $\mathbf{w}_S$ is the subvector formed from the corresponding components of $\mathbf{w}$, then $\underline{\mathbf{w}_S} = \mathrm{ccl}\, \mathbf{w}_S$ is termed a *subcomposition*.

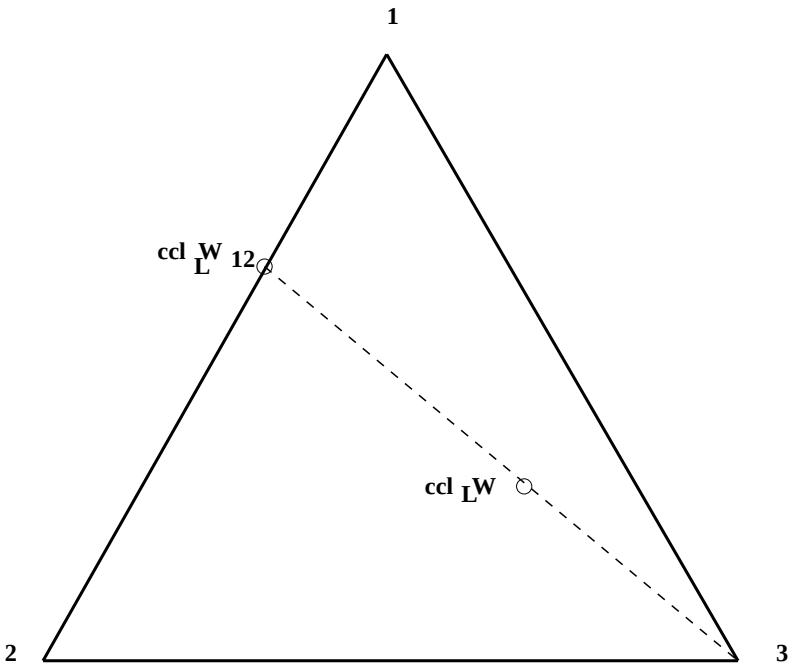
If the subset S is formed by C indices, with $2 \leq C < D$, the subcomposition $\underline{\mathbf{w}_S}$ belongs to the compositional space $\mathcal{C}^{C-1}$.

Definition The formation of a C-subcomposition $\underline{\mathbf{w}_S}$ from a D-composition $\underline{\mathbf{w}}$ may be considered as the mapping sub_S from $\mathcal{C}^{D-1}$ to $\mathcal{C}^{C-1}$:

$$\mathrm{sub}_S\, \underline{\mathbf{w}} = \underline{\mathbf{w}_S} \qquad (\underline{\mathbf{w}} \in \mathcal{C}^{D-1})$$

Subcompositions





Compositional Problems

1. Percentage of Cl, K_2O, P_2O_5, TiO_2 and SiO_2 in 46 samples of volcanic rocks from a volcano H:

Num.	Cl	K_2O	P_2O_5	TiO_2	SiO_2
1	0.0638	1.83	1.01	3.70	44.99
2	0.1116	1.36	0.81	3.83	43.45
3	0.0611	1.36	1.09	4.10	42.98
⋮	⋮	⋮	⋮	⋮	⋮
44	0.0477	1.67	0.96	3.64	45.14
45	0.0015	1.70	0.95	3.64	45.15
46	0.0986	3.22	0.50	1.87	54.41

Compositional Problems

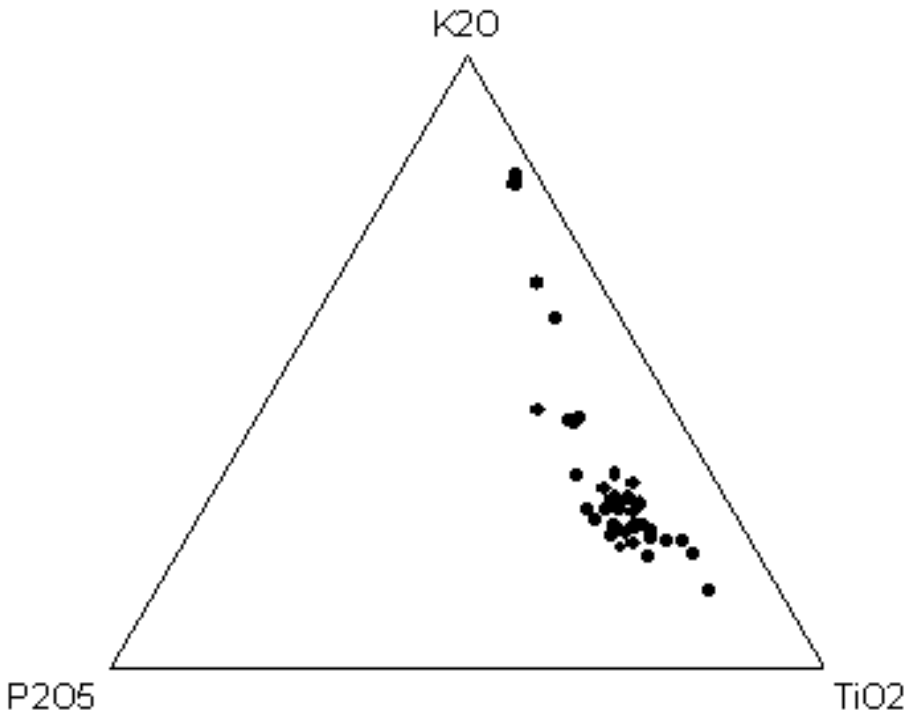
1.a It is possible to describe the pattern of variability of these volcanic rocks and to define a covariance or correlation structure?

1.b Is it possible to define a measure of total variability of this set of volcanic rocks?

1.c For a new volcanic rock specimen with known composition (Cl,K_2O,P_2O_5,TiO_2,SiO_2)' and claimed to be from the same volcano, can we say whether it is fairly typical from this volcano? If not, can we place some measure on its atypicality?

1.d To what extent, if any, do the subcomposition (Cl,K_2O,P_2O_5) explain the pattern of variability of the full composition?

Compositional Problems



1.e From this ternary diagram it seems that the pattern of (K_2O,P_2O_5,TiO_2) can be well adjusted by a curve. How can we confirm this?

Compositional Problems

2. Percentage of Cl, K_2O, P_2O_5, TiO_2 and SiO_2: 65 samples of volcanic rocks from a volcano *A*, and 19 samples from another volcano D.

Num.	Cl	K_2O	P_2O_5	TiO_2	SiO_2
1A	0.1776	0.64	0.34	1.57	49.26
2A	0.2050	1.55	0.43	1.61	48.22
⋮	⋮	⋮	⋮	⋮	⋮
65A	0.0391	1.70	0.55	1.63	50.91
1D	0.0181	0.64	0.31	2.55	49.33
2D	0.0053	1.10	0.59	2.81	46.89
⋮	⋮	⋮	⋮	⋮	⋮
19D	0.0200	0.99	0.37	3.16	45.85

Compositional Problems

2.a Can we detect any differences between the compositional pattern of volcano A and volcano D?

If so, how can we choose a 3-part subcomposition which somehow captures the essence of the two patterns individually and yet emphasizes the differences between the patterns?

2.b Is it possible to establish a classification rule for discriminating between volcanos A and D?

Compositional Problems

3. Sand, silt,clay composition (% by weight) of 39 sediment samples at different water depths in an Artic lake:

Num.	Sand	Silt	Clay	Depth (m)
S01	77.5	19.5	3.0	10.4
S02	71.9	24.9	3.2	11.7
⋮	⋮	⋮	⋮	⋮
S39	2.0	47.8	50.2	103.7

3.a Is sediment composition dependent on water depth?

3.b If so, how can we quantify the extent of the dependence?

How to analyze "closed" raw data?

Spurious correlations

Pearson (1897) "If $u = f(x, y)$ and $v = g(z, y)$ be two functions of three variables x, y, z, and these variables be selected at random so that there exists no correlation between x and y, y and z, or z and x, there will still be found to exist correlation between u and v. ... That is likely to occur when u and v are indices with the same denominator".

Consequence The standard covariance matrix $[s_{ij}]$ of a closed data set from $\mathcal{S}^D$ is always singular because

$$\sum_{j=1}^{D} s_{ij} = 0, \quad \text{for } i = 1, \ldots, D.$$

How to analyze "closed" raw data?

Subcompositional incoherence

Example

Scientist **A**	Scientist **B**
Full compositions from $\mathcal{S}^4$	Subcompositions from $\mathcal{S}^3$
(x_1, x_2, x_3, x_4)	(s_1, s_2, s_3)
$(0.1, 0.2, 0.1, 0.6)$	$(0.250, 0.500, 0.250)$
$(0.2, 0.1, 0.1, 0.6)$	$(0.500, 0.250, 0.250)$
$(0.3, 0.3, 0.2, 0.2)$	$(0.375, 0.375, 0.250)$
$\text{corr}\{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}\} = 0.5$	$\text{corr}\{\mathbf{s}_{(1)}, \mathbf{s}_{(2)}\} = -1$

Any statement that scientists **A** and **B** make about the common parts 1,2 and 3 must agree.

Statistics in $\mathbb{R}^D$

Translation In $\mathbb{R}^D$ the inner operation is *translation*. If $t \in \mathbb{R}^D$, the translation t moves the random vector X in $\mathbb{R}^D$ to a random vector $X + t$ in such a way that

$$\mathrm{E}\{X + t\} = \mathrm{E}\{X\} + t \quad \text{and} \quad \mathbf{\Sigma}\{X + t\} = \mathbf{\Sigma}\{X\}.$$

Scalar product For any random vector X on $\mathbb{R}^D$ and for any $\lambda \in R$,

$$\mathrm{E}\{\lambda X\} = \lambda \mathrm{E}\{X\} \quad \text{and} \quad \mathbf{\Sigma}\{\lambda X\} = \lambda^2 \mathbf{\Sigma}\{X\}.$$

Perturbations on $\mathcal{C}^{D-1}$

Scale invariance is the property which characterizes compositional data. Therefore, any "operation" involving compositions must be compatible with this property.

Definition We define an inner operation $\oplus$ in $\mathcal{C}^{D-1}$ as

$$\underline{\mathbf{w}} \oplus \underline{\mathbf{w}^*} = \mathrm{ccl}\,(w_1 w_1^*, \ldots, w_D w_D^*)'.$$

$(\mathcal{C}^{D-1}, \oplus)$ is a commutative group:

- Composition $\underline{\mathbf{1}_D} = \mathrm{ccl}\,(1, \ldots, 1)'$ is the neutral element.
- The inverse composition $\underline{\mathbf{w}}^{-1}$ of $\underline{\mathbf{w}} = \mathrm{ccl}\,(w_1, \ldots, w_D)'$ is the composition $\underline{\mathbf{w}^{-1}} = \mathrm{ccl}\,(1/w_1, \ldots, 1/w_D)'$.

The group of perturbations in $\mathcal{C}^{D-1}$

Definition Given a composition $\mathbf{p} \in \mathcal{C}^{D-1}$, the *perturbation* associated to $\mathbf{p}$ is the transformation from $\mathcal{C}^{D-1}$ to $\mathcal{C}^{D-1}$ defined by

$$\mathbf{c} \to \mathbf{p} \oplus \mathbf{c} \qquad (\mathbf{c} \in \mathcal{C}^{D-1}).$$

Then, we say that $\mathbf{p} \oplus \mathbf{c}$ is the composition which results when the *perturbation* $\mathbf{p}$ is applied to the composition $\mathbf{c}$.

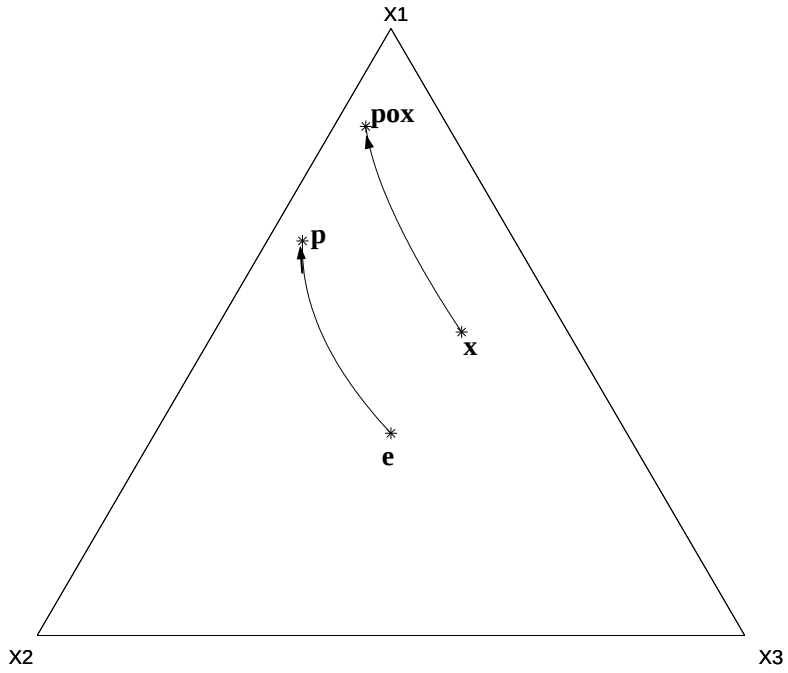
Moreover, given two compositions of $\mathcal{C}^{D-1}$

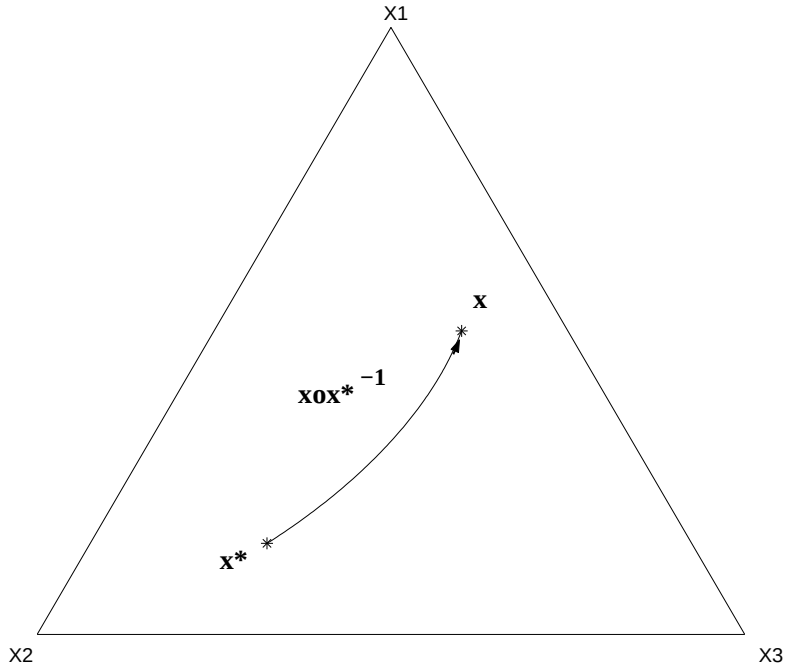
$$\underline{\mathbf{w}} = \mathrm{ccl}\,(w_1, \ldots, w_D)', \ \underline{\mathbf{w}^*} = \mathrm{ccl}\,(w_1^*, \ldots, w_D^*)',$$

there exists a unique perturbation $\mathbf{p}$ which transforms $\underline{\mathbf{w}}$ on $\underline{\mathbf{w}^*}$:

$$\mathbf{p} = \underline{\mathbf{w}^*} \oplus \underline{\mathbf{w}}^{-1} = \mathrm{ccl}\left(\frac{w_1^*}{w_1}, \ldots, \frac{w_D^*}{w_D}\right)'.$$

The group of perturbations in $\mathcal{C}^{D-1}$





The group of perturbations in $\mathcal{C}^{D-1}$

Perturbation in compositional space plays the same role as *translation* plays in real space. The set of all perturbations in $\mathcal{C}^{D-1}$ is a commutative group isomorphic to $(\mathcal{C}^{D-1}, \oplus)$. For this reason, we will also call *perturbation* the inner operation $\oplus$ defined on $\mathcal{C}^{D-1}$.

The assumption that the group of perturbations is the operating group on the compositional space is the *keystone* of the methodology introduced by Aitchison (1986). In fact, it implies to accept that the "difference" between two compositions $\underline{\mathbf{w}} = \mathrm{ccl}\,(w_1, \ldots, w_D)'$ and $\underline{\mathbf{w}^*} = \mathrm{ccl}\,(w_1^*, \ldots, w_D^*)'$ will be based on the *ratios* w_j^*/w_j between parts instead of on the arithmetic differences $w_j^* - w_j$.

Perturbations on $\mathcal{C}^{D-1}$

Interpretation

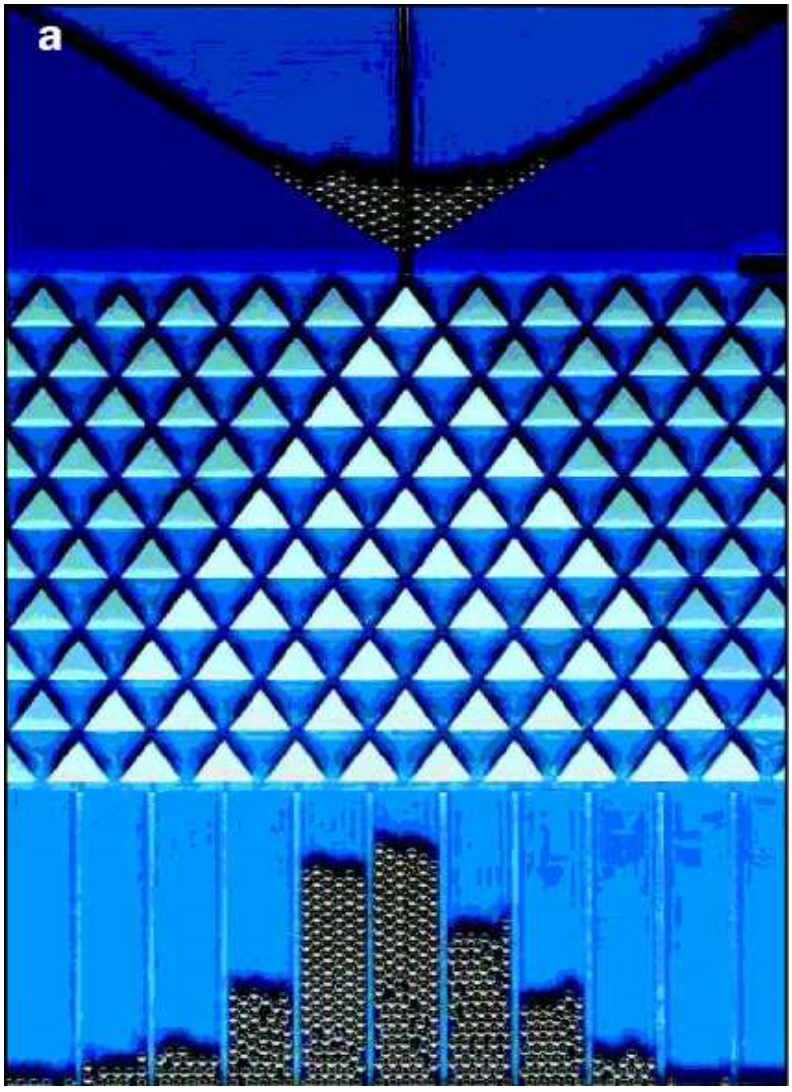
- Some natural processes in nature can be interpreted as a succession of changes from one initial composition $\underline{\mathbf{w}_0}$ to a final composition $\underline{\mathbf{w}_n}$ through the application of successive perturbations:

$$\begin{aligned} \underline{\mathbf{w}_0} &\longrightarrow \underline{\mathbf{p}_1} \oplus \underline{\mathbf{w}_0} = \underline{\mathbf{w}_1} \\ &\longrightarrow \underline{\mathbf{p}_2} \oplus \underline{\mathbf{w}_1} = \underline{\mathbf{w}_2} \\ &\dots \\ &\longrightarrow \underline{\mathbf{p}_n} \oplus \underline{\mathbf{w}_{n-1}} = \underline{\mathbf{w}_n}. \end{aligned}$$

  In this manner,

$$\underline{\mathbf{w}_n} = (\underline{\mathbf{p}_n} \oplus \underline{\mathbf{p}_{n-1}} \oplus \dots \underline{\mathbf{p}_1}) \oplus \underline{\mathbf{w}_0}.$$

Genesis of normal distribution



Particles fall from a funnel onto tips of triangles, where they are deviated to the left or to the right with equal probability (0.5) and finally fall into receptacles. If the tip of a triangle is at distance x from the left edge of the board, triangle tips to the right and to the left below it are placed at $x + k$ and $x - k$ (k constant).

Genesis of lognormal distribution

Particles fall from a funnel onto tips of triangles, where they are deviated to the left or to the right with equal probability (0.5) and finally fall into receptacles. If the tip of a triangle is at distance x from the left edge of the board, triangle tips to the right and to the left below it are placed at x/k and $x.k$ (k constant).

Perturbations on $\mathcal{C}^{D-1}$

Interpretation

- If $\underline{\mathbf{w}} = \text{ccl}\,(w_{SiO_2}, \ldots, w_{P_2O_5})'$ expresses the percentage composition of major oxides of a rock, its molecular composition will be

$$\underline{\mathbf{w}^*} = \text{ccl}\,(w_{SiO_2}/m_{SiO_2}, \ldots, w_{P_2O_5}/m_{P_2O_5})',$$

where m_j symbolizes the molecular weight of oxide j.

Therefore, composition $\underline{\mathbf{w}^*}$ can be obtained applying the perturbation

$$\underline{\mathbf{m}^{-1}} = (\text{ccl}\,(m_{SiO_2}, \ldots, m_{P_2O_5})')^{-1}$$

to composition $\underline{\mathbf{w}}$:

$$\underline{\mathbf{w}^*} = \underline{\mathbf{m}^{-1}} \oplus \underline{\mathbf{w}}.$$

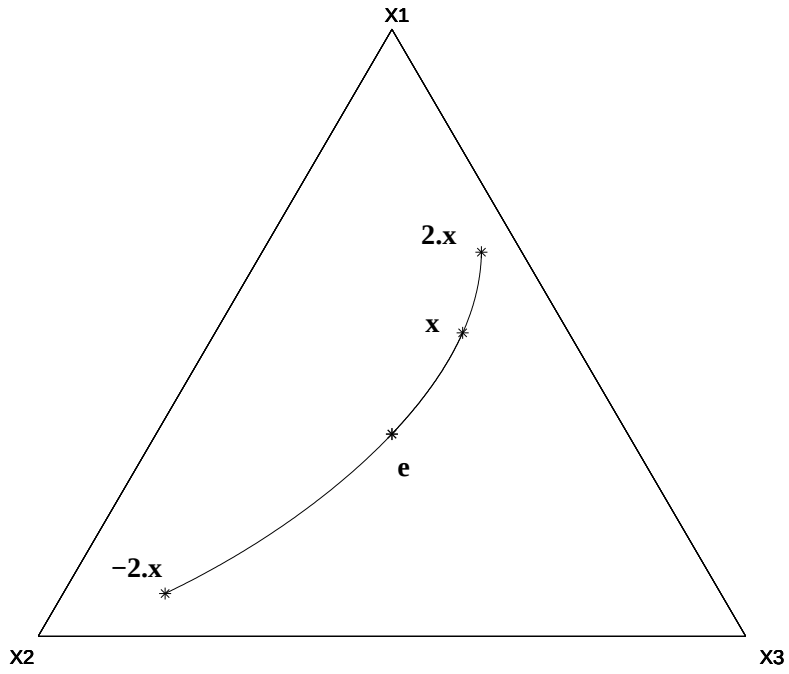
The vector space $(\mathcal{C}^{D-1}, \oplus, \otimes)$

Definition The external operation $\otimes$ in $\mathcal{C}^{D-1}$ is defined as

$$\lambda \otimes \underline{\mathbf{w}} = \mathrm{ccl}\,(w_1^\lambda, \ldots, w_D^\lambda)',$$

for each $\lambda \in \mathbb{R}$ and each $\underline{\mathbf{w}} \in \mathcal{C}^{D-1}$.

$(\mathcal{C}^{D-1}, \oplus, \otimes)$ is a vector space of dimension $D-1$.

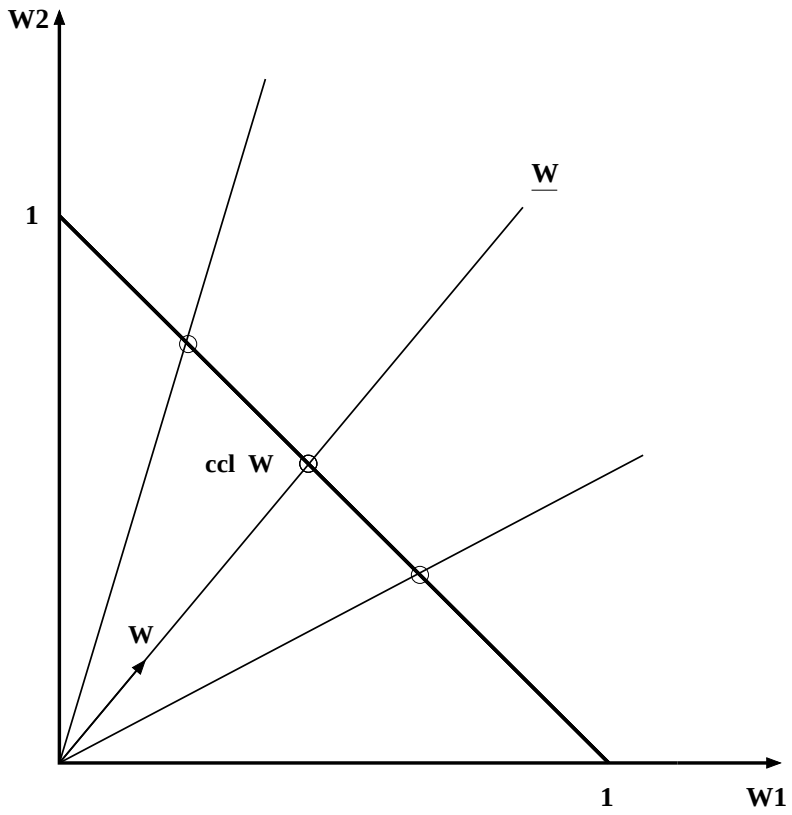


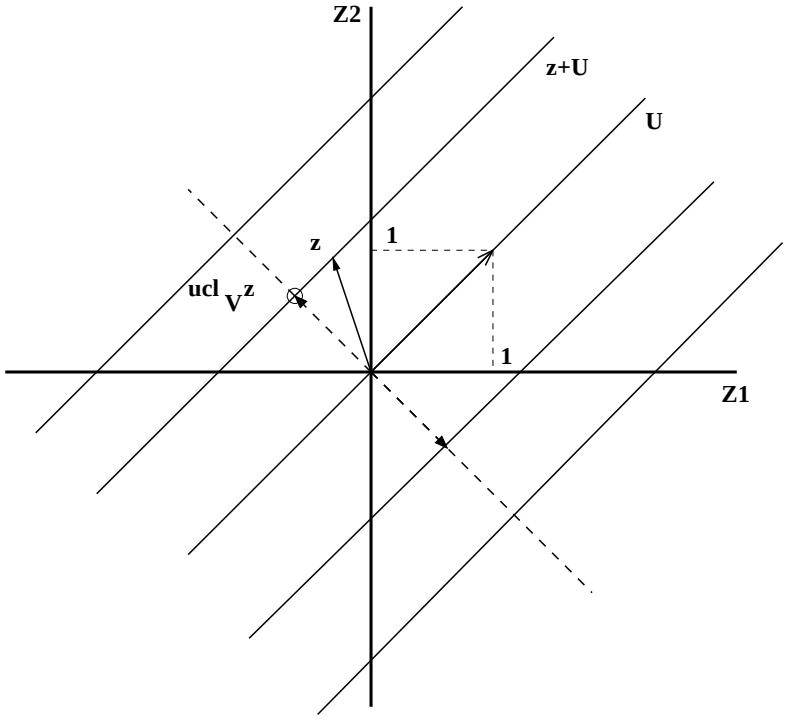
The *log* and the *exp* transformations

between $\mathbb{R}_+^D$ and $\mathbb{R}^D$

The logarithmic transformation on $\mathbb{R}_+^D$ transforms the rays from the origin —which represent the compositions of the space $\mathcal{C}^{D-1}$—, to straight lines of $\mathbb{R}^D$ parallel to vector $\mathbf{1}_D = (1, \overset{D}{\ldots}, 1)'$.

Inversely, the exponential transformation on $\mathbb{R}^D$ transforms these straight lines of $\mathbb{R}^D$ parallel to vector $\mathbf{1}_D$, to rays from the origin of $\mathbb{R}_+^D$.


W2
1
W
ccl W
W
1
W1


Z2
z+U
U
z
1
ucl V z
1
Z1

Centered logratio transformation

Definition The *centered logratio transformation* —denoted by clr — is the one-to-one function from the compositional space $\mathcal{C}^{D-1}$ to the subspace
$V = \{\mathbf{z} = (z_1, \ldots, z_D)' \in \mathbb{R}^D : z_1 + \ldots + z_D = 0\}$
of $\mathbb{R}^D$, defined by

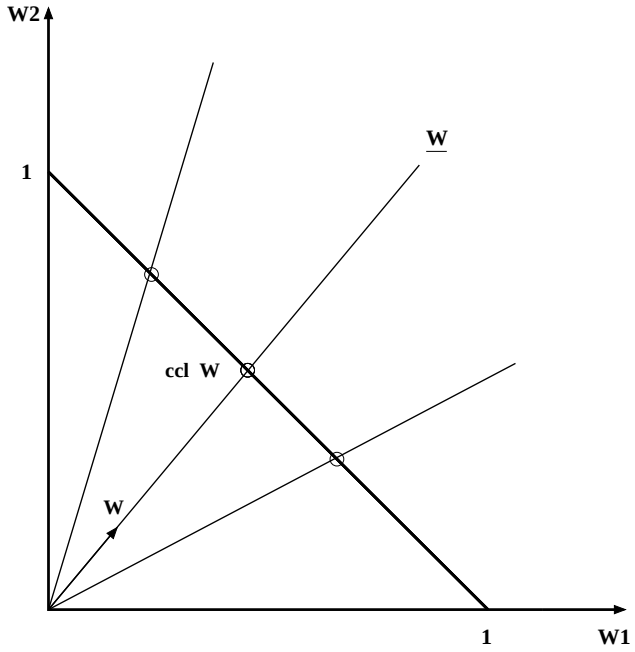
$$\operatorname{clr} \underline{\mathbf{w}} = \log \frac{\mathbf{w}}{g(\mathbf{w})} \qquad (\underline{\mathbf{w}} \in \mathcal{C}^{D-1}).$$

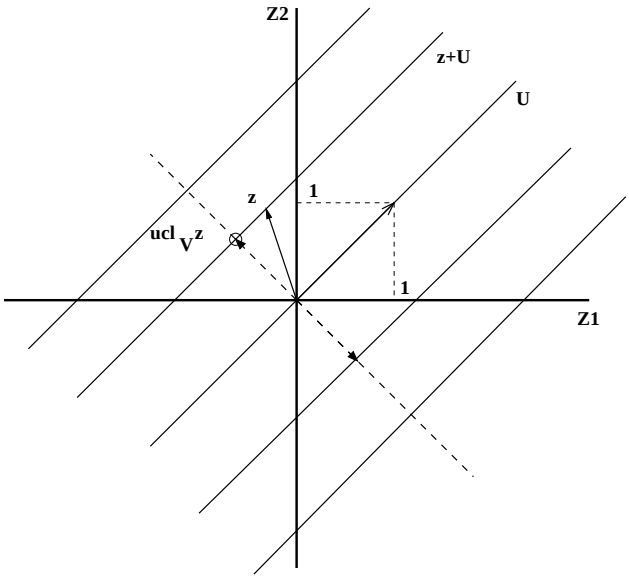
The inverse transformation, from V to $\mathcal{C}^{D-1}$, is given by

$$\operatorname{clr}^{-1} \mathbf{z} = \operatorname{ccl}(\exp \mathbf{z}) \quad (\mathbf{z} \in V).$$

The logarithmic and the exponential transformations establish a one-to-one correspondence between the simplex $\mathcal{S}^D$ and the hyperplan V in $\mathbb{R}^D$.

Centered logratio transformation





Centered logratio transformation

Property The *centered logratio transformation* is an isomorphism between the vector space $(\mathcal{C}^{D-1}, \oplus, \otimes)$ and the vector subspace

$$V = \{\mathbf{z} = (z_1, \ldots, z_D)' \in \mathbb{R}^D : z_1 + \ldots + z_D = 0\}$$

of $(\mathbb{R}^D, +, .)$. Therefore,

$$\operatorname{clr}(\underline{\mathbf{w}} \oplus \underline{\mathbf{w}^*}) = \operatorname{clr} \underline{\mathbf{w}} + \operatorname{clr} \underline{\mathbf{w}^*} \ ;$$

$$\operatorname{clr}(\lambda \otimes \underline{\mathbf{w}}) = \lambda \ \operatorname{clr} \underline{\mathbf{w}},$$

where $\underline{\mathbf{w}}, \underline{\mathbf{w}^*} \in \mathcal{C}^{D-1}$, and $\lambda \in \Re$.

Equally,

$$\operatorname{clr}^{-1}(\mathbf{z} + \mathbf{z}^*) = \operatorname{clr}^{-1} \mathbf{z} \oplus \operatorname{clr}^{-1} \mathbf{z}^* \ ;$$

$$\operatorname{clr}^{-1}(\lambda \ \mathbf{z}) = \lambda \otimes \operatorname{clr}^{-1} \mathbf{z},$$

where $\mathbf{z}, \mathbf{z}^* \in V$, and $\lambda \in \mathbb{R}$.

Isometric logratio transformation

Let $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}\}$ an orthonormal basis of the subspace

$$V = \{\mathbf{z} = (z_1, \ldots, z_D)' \in \mathbb{R}^D : z_1 + \ldots + z_D = 0\}.$$

Then, since $\operatorname{clr} \underline{\mathbf{w}} \in V$, it will be always possible to write

$$\operatorname{clr} \underline{\mathbf{w}} = u_1 \mathbf{v}_1 + \ldots + u_{D-1} \mathbf{v}_{D-1},$$

for any $\underline{\mathbf{w}} \in \mathcal{C}^{D-1}$.

Definition The *isometric logratio transformation* —denoted by $\operatorname{ilr}_{\mathcal{V}}$— is the one-to-one function from the compositional space $\mathcal{C}^{D-1}$ to $\mathbb{R}^{D-1}$ defined by

$$\operatorname{ilr}_{\mathcal{V}} \underline{\mathbf{w}} = (u_1, \ldots, u_{D-1})' \qquad (\underline{\mathbf{w}} \in \mathcal{C}^{D-1}).$$

Like clr , the transformation $\operatorname{ilr}_{\mathcal{V}}$ is an isomorphism between the vector spaces $(\mathcal{C}^{D-1}, \oplus, \otimes)$ and $(\mathbb{R}^{D-1}, +, .)$.

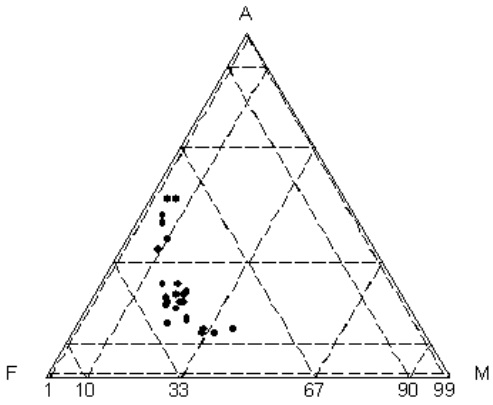
Skye Lavas

Sample	$Na_2O + K_2O$	Fe_2O_3	MgO
S1	52	42	6
S2	52	44	4
S3	47	48	5
	clr $(Na_2O + K_2O)$	clr (Fe_2O_3)	clr (MgO)
S1	0,7910	0,5775	-1,3685
S2	0,9107	0,7436	-1,6543
S3	0,7399	0,7609	-1,5008
	ilr $_1$ $[u_1]$	ilr $_2$ $[u_2]$	
S1	0,1510	1,6760	
S2	0,1181	2,0261	
S3	-0,0149	1,8381	

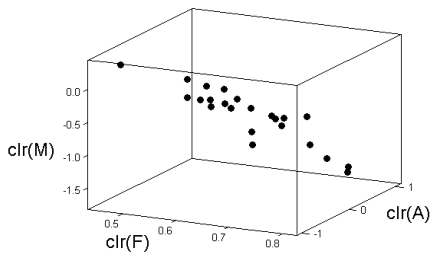
Hint The orthonormal basis $\mathcal{V}$ of the subspace $V \subset \mathbb{R}^3$ linked to the ilr coordinates is

$$\mathbf{v}_1 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0)', \quad \mathbf{v}_2 = (\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}})'$$

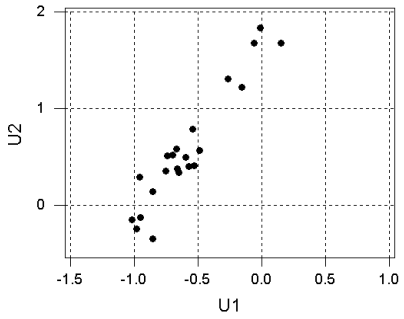
Skye Lavas



Ternary diagram



clr-coordinates



ilr-coordinates

$\mathcal{C}^{D-1}$ as an Euclidean space

The clr transformation between $\mathcal{C}^{D-1}$ and the subspace V of $\mathbb{R}^D$ allows to translate to $\mathcal{C}^{D-1}$ the real Euclidean structure defined on V:

$$\begin{aligned}
<\underline{\mathbf{w}}, \underline{\mathbf{w}^*}>_{\mathcal{C}} &= (\operatorname{clr} \underline{\mathbf{w}})' \operatorname{clr} \underline{\mathbf{w}^*} \\
&= (\log \mathbf{w})' \mathbf{H}_D \log \mathbf{w}^*, \\
\| \underline{\mathbf{w}} \|_{\mathcal{C}} &= \| \operatorname{clr} \underline{\mathbf{w}} \| = [(\log \mathbf{w})' \mathbf{H}_D \log \mathbf{w}]^{1/2}, \\
d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}^*}) &= d_{Euc}(\operatorname{clr} \underline{\mathbf{w}}, \operatorname{clr} \underline{\mathbf{w}^*}) \\
&= [(\log \mathbf{w}^* - \log \mathbf{w})' \mathbf{H}_D \\
&\quad (\log \mathbf{w}^* - \log \mathbf{w})]^{1/2},
\end{aligned}$$

where $\mathbf{H}_D$ is the $(D \times D)$-*centering matrix*. This matrix is equal $\mathbf{I}_D - D^{-1}\mathbf{J}_D$, where $\mathbf{I}_D$ is the identity matrix and $\mathbf{J}_D = \mathbf{1}_D \mathbf{1}_D'$.

Therefore, by construction, transformations clr and clr^{-1} —and also ilr and ilr^{-1}— preserve the distances defined in $\mathcal{C}^{D-1}$ and $\mathbb{R}^{D-1}$.

Compositional geometry in $\mathcal{C}^{D-1}$

- We can not analyze the simplex $\mathcal{S}^D$ as we analyze the Euclidean real space.

  Let

$$\underline{\mathbf{w}_1} = \mathrm{ccl}\,(1.000, 49.500, 39.500)'$$
$$\underline{\mathbf{w}_2} = \mathrm{ccl}\,(0.010, 49.995, 39.995)'$$
$$\underline{\mathbf{w}_3} = \mathrm{ccl}\,(25.0, 50.0, 25.0)'$$
$$\underline{\mathbf{w}_4} = \mathrm{ccl}\,(35.0, 30.0, 35.0)'$$

  be four compositions from $\mathcal{S}^3$.

  Then,

$$d_{Euc}(\underline{\mathbf{w}_1}, \underline{\mathbf{w}_2}) \approx 1.21 < 24.49 \approx d_{Euc}(\underline{\mathbf{w}_3}, \underline{\mathbf{w}_4}),$$

  whereas

$$d_{\mathcal{C}}(\underline{\mathbf{w}_1}, \underline{\mathbf{w}_2}) \approx 3.77 > 0.69 \approx d_{\mathcal{C}}(\underline{\mathbf{w}_3}, \underline{\mathbf{w}_4}).$$

Compositional geometry in $\mathcal{C}^{D-1}$

- Any linear variety on $\mathcal{C}^{D-1}$ —straight lines, planes, etc— can always be implicitly expressed by a system of linear equations in $\log w_1, \ldots, \log w_D$ in the form

$$\left.\begin{array}{ccc} a_{11}\log w_1 + & \ldots & + a_{1D}\log w_D = b_1 \\ \ldots & \ldots & \ldots \\ a_{m1}\log w_1 + & \ldots & + a_{mD}\log w_D = b_m \end{array}\right],$$

  with $a_{i1} + \ldots + a_{iD} = 0$, for each $i = 1, \ldots, m$.

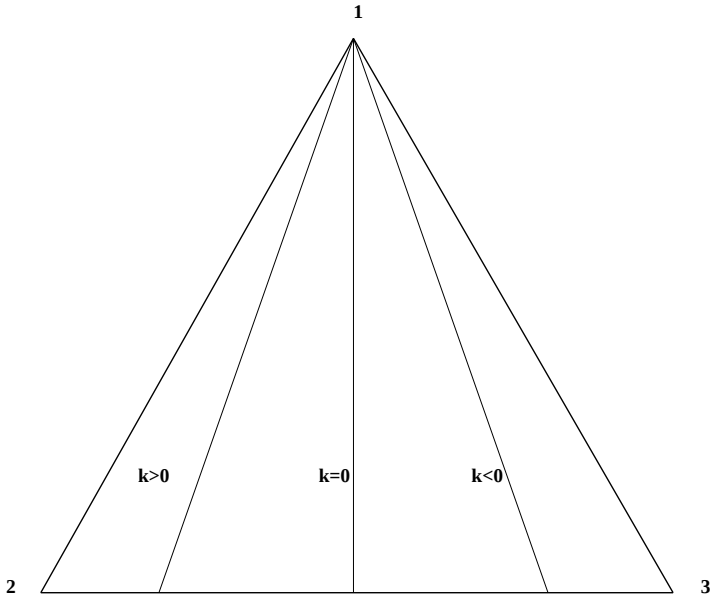
- In particular, the parametric equation —varying $t \in \mathbb{R}$— of a straight line on $\mathcal{C}^{D-1}$ is given by

$$\underline{\mathbf{w}(t)} = \mathrm{ccl}\,(\exp(\alpha_1 + \lambda_1 t), \ldots, \exp(\alpha_D + \lambda_D t))',$$

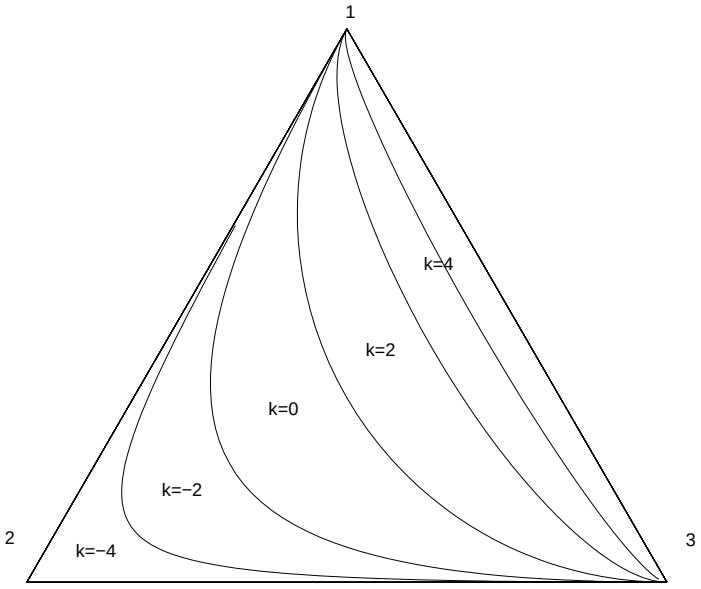
  where $\sum_{j=1}^{D} alpha_j = 0$ and $\sum_{j=1}^{D} \lambda_j = 0$.

- Similarly to real space, the concepts of parallelism and orthogonality can be introduced in $\mathcal{C}^{D-1}$.

Parallelism in $\mathcal{C}^2$

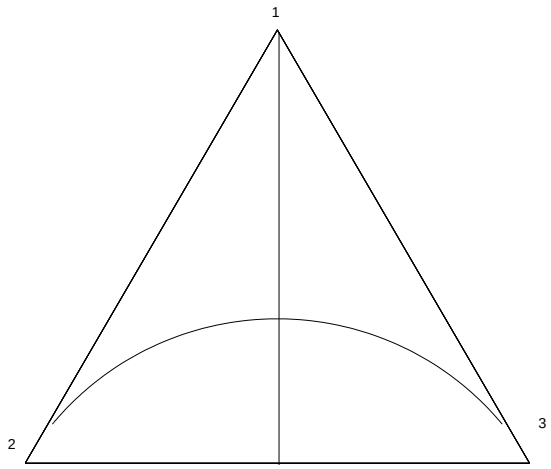


$$\log w_2 - \log w_3 = k$$



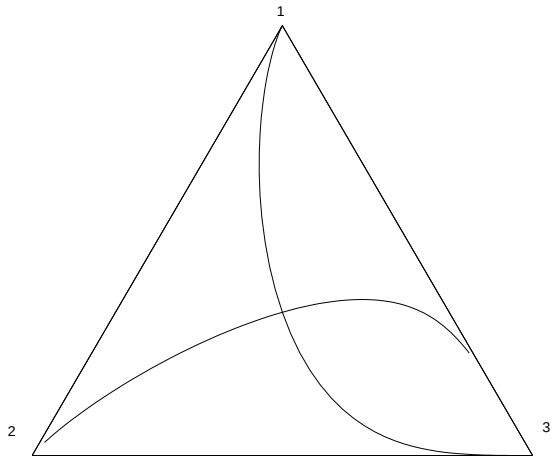
$$\log w_1 - 2\log w_2 + \log w_3 = k$$

Orthogonality in $\mathcal{C}^2$



$$w_2 - w_3 = 0$$

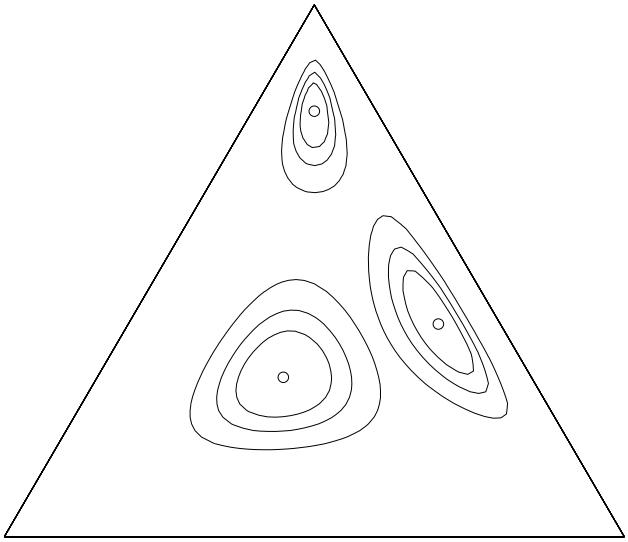
$$-2\log w_1 + \log w_2 + \log w_3 = 0$$



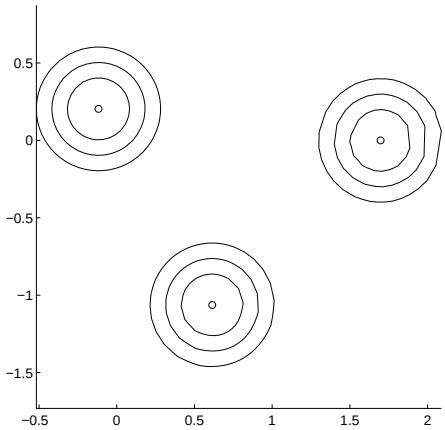
$$\log w_1 - 3\log w_2 + 2\log w_3 = 0$$

$$5\log w_1 - \log w_2 - 4\log w_3 = 0$$

Circles in $\mathcal{C}^2$



Simplex $\mathcal{S}^3$



clr -space

The alr transformation

Definition The *additive logratio transformation* of index j $(j = 1, \ldots, D)$ — denoted by alr_j — is the one-to-one transformation from $\mathcal{C}^{D-1}$ to $\mathbb{R}^{D-1}$ defined by

$$\underline{\mathbf{w}} \longrightarrow \mathbf{y} = \mathrm{alr}_j\, \underline{\mathbf{w}} = \log \frac{\mathbf{w}_{-j}}{w_j}.$$

where

$$\mathbf{w}_{-j} = (w_1, w_2, \ldots, w_{j-1}, w_{j+1}, \ldots, w_D)'.$$

The inverse transformation of alr_D, from $\mathbb{R}^{D-1}$ to $\mathcal{C}^{D-1}$, is given by

$$\mathrm{alr}_D^{-1}\, \mathbf{y} = \mathrm{ccl}\,(\exp y_1, \ldots, \exp y_{D-1}, 1)' \;\; (\mathbf{y} \in \mathbb{R}^{D-1}).$$

The alr transformation

Property The alr_j transformations $(j = 1, \ldots, D)$ are isomorphisms between the vector spaces $(\mathcal{C}^{D-1}, \oplus, \otimes)$ and $(\mathbb{R}^{D-1}, +, .)$, i.e.,

$$\text{alr}_j\,(\underline{\mathbf{w}} \oplus \underline{\mathbf{w}^*}) = \text{alr}_j\,\underline{\mathbf{w}} + \text{alr}_j\,\underline{\mathbf{w}^*}\ ;$$

$$\text{alr}_j\,(\lambda \otimes \underline{\mathbf{w}}) = \lambda\ \text{alr}_j\,\underline{\mathbf{w}},$$

$$\text{alr}_j^{-1}\,(\mathbf{y} + \mathbf{y}^*) = \text{alr}_j^{-1}\,\mathbf{y} \oplus \text{alr}_j^{-1}\,\mathbf{y}^*\ ;$$

$$\text{alr}_j^{-1}\,(\lambda\ \mathbf{y}) = \lambda \otimes \text{alr}_j^{-1}\,\mathbf{y},$$

where $\underline{\mathbf{w}}, \underline{\mathbf{w}^*} \in \mathcal{C}^{D-1}$, $\mathbf{y}, \mathbf{y}^* \in \mathbb{R}^{D-1}$ and $\lambda \in \mathbb{R}$.

Property The alr_j transformations $(j = 1, \ldots, D)$ do not preserve the distances defined in the metric spaces $\mathcal{C}^{D-1}$ and $\mathbb{R}^{D-1}$, i.e.,

$$d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}^*}) \neq d_{Euc}(\text{alr}_j\,\underline{\mathbf{w}}, \text{alr}_j\,\underline{\mathbf{w}^*});$$

$$d_{Euc}(\mathbf{y}, \mathbf{y}^*) \neq d_{\mathcal{C}}(\text{alr}_j^{-1}\,\mathbf{y}, \text{alr}_j^{-1}\,\mathbf{y}^*).$$

Determination of a composition

A composition $\underline{\mathbf{w}} \in \mathcal{C}^{D-1}$ can be determined in several forms:

(i) Giving any D-observational vector belonging to $\underline{\mathbf{w}}$. Usually, we will choose the vector $\mathbf{x} = \mathcal{C}\mathbf{w} = \mathrm{ccl}_L\, \mathbf{w}$ belonging to $\mathcal{S}^D$.

(ii) Giving the components $(z_1, \ldots, z_D)' = \mathbf{z}$ of the centered logratio transformed vector $\mathrm{clr}\, \underline{\mathbf{w}}$. Since $\mathbf{z}$ belongs to the subspace V of $\mathbb{R}^D$, its components are related by the equality $z_1 + \ldots + z_D = 0$.

(iii) Giving the components $(y_1, \ldots, y_{D-1})' = \mathbf{y}$ of the additive logratio transformed vector $\mathrm{alr}_D\, \underline{\mathbf{w}}$. If it is needed, we can choose the components of any other logratio $\mathrm{alr}_j \underline{\mathbf{w}}$ $(j \neq D)$.

(iv) Giving the components $(u_1, \ldots, u_{D-1})' = \mathbf{u}$ of the isometric logratio transformed vector $\mathrm{ilr}_{\mathcal{V}}\, \underline{\mathbf{w}}$, where $\mathcal{V}$ is a known orthonormal basis of the subspace V of $\mathbb{R}^D$.

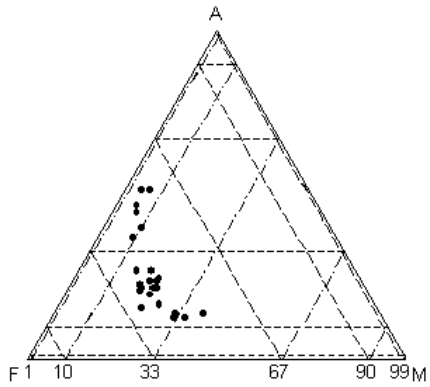
Determination of a composition

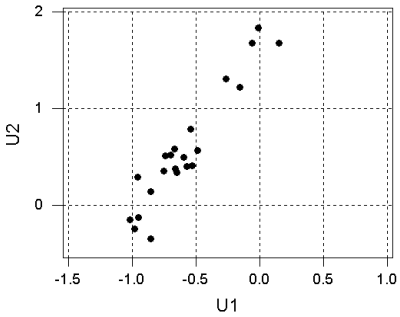
- Skye lavas: $\mathrm{A} = \mathrm{Na_2O} + \mathrm{K_2O},\ \mathrm{F} = \mathrm{Fe_2O_3},\ \mathrm{M} = \mathrm{MgO}$

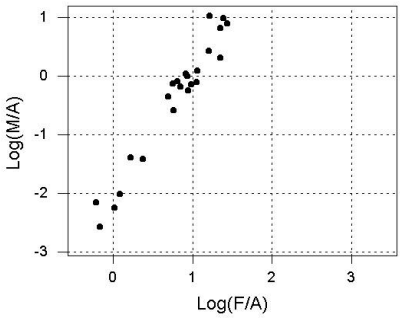
Sample	A	F	M
S1	52	42	6
S2	52	44	4
	clr (A)	clr (F)	clr (M)
S1	0.7910	0.5775	-1.3685
S2	0.9107	0.7436	-1.6543
	ilr $_1$ $[u_1]$	ilr $_2$ $[u_2]$	
S1	0,1510	1,6760	
S2	0,1181	2,0261	
	$\mathrm{alr_M A}$	$\mathrm{alr_M F}$	
S1	2.159	1.946	
S2	2.565	2.398	
	$\mathrm{alr_A F}$	$\mathrm{alr_A M}$	
S1	-0.214	-2.159	
S2	-0.167	-2.565	

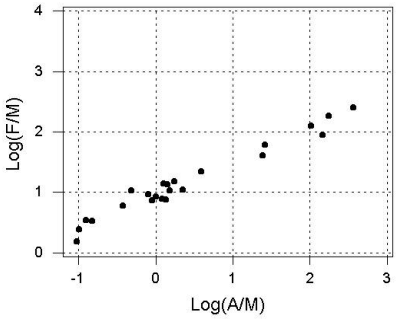
Hint The orthonormal basis $\mathcal{V}$ of the subspace $V \subset \mathbb{R}^3$ linked to the ilr coordinates is

$$\mathbf{v}_1 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0)', \quad \mathbf{v}_2 = (\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}})'.$$


A
F 1 10 33 67 90 99 M


2
1
0
U2
-1.5 -1.0 -0.5 0.0 0.5 1.0
U1


1
0
-1
-2
-3
Log(M/A)
0 1 2 3
Log(F/A)


4
3
2
1
0
Log(F/M)
-1 0 1 2 3
Log(A/M)

Compositional data set

- Raw data matrix

$$\mathbf{W} = [w_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, D],$$

or

$$\mathbf{X} = [x_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, D],$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})' \in \mathcal{S}^D$.

Example AFM composition of 23 aphyric Skye lavas [$\mathrm{A} = \mathrm{Na_2O} + \mathrm{K_2O},\ \mathrm{F} = \mathrm{Fe_2O_3},\ \mathrm{M} = \mathrm{MgO}$].

Obs.	A%	F%	M%
S1	52	42	6
⋮	⋮	⋮	⋮
S23	24	56	20

$$\mathbf{X} = \begin{bmatrix} 52 & 42 & 6 \\ \vdots & \vdots & \vdots \\ 24 & 56 & 20 \end{bmatrix}.$$

Compositional data set

- Centred logratio (clr) data matrix

$$\mathbf{Z} = [z_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, D],$$

where $z_{ij} = \log\left(w_{ij}/g(\mathbf{w}_i)\right)$, with $g(\mathbf{w}_i) = \left(\prod_{k=1}^{D} w_{ik}\right)^{1/D}$.

Example AFM composition of 23 aphyric Skye lavas [$\mathrm{A} = \mathrm{Na_2O} + \mathrm{K_2O},\ \mathrm{F} = \mathrm{Fe_2O_3},\ \mathrm{M} = \mathrm{MgO}$].

Obs.	A%	F%	M%
S1	52	42	6
⋮	⋮	⋮	⋮
S23	24	56	20

$$\begin{array}{c} \\ \mathbf{Z} = \end{array} \begin{array}{c} \begin{array}{ccc} \text{clr A} & \text{clr F} & \text{clr M} \end{array} \\ \begin{bmatrix} 0.791 & 0.577 & -1.368 \\ \vdots & \vdots & \vdots \\ -0.222 & 0.626 & -0.404 \end{bmatrix} \end{array}.$$

Compositional data set

- Additive logratio (alr) data matrix

$$\mathbf{Y} = [y_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, d],$$

where $y_{ij} = \log(w_{ij}/w_{iD})$.

Example AFM composition of 23 aphyric Skye lavas [$\mathrm{A} = \mathrm{Na_2O} + \mathrm{K_2O},\ \mathrm{F} = \mathrm{Fe_2O_3},\ \mathrm{M} = \mathrm{MgO}$].

Obs.	A%	F%	M%
S1	52	42	6
⋮	⋮	⋮	⋮
S23	24	56	20

$$\mathbf{Y} = \begin{array}{c} \begin{matrix} \log \frac{A}{M} & \log \frac{F}{M} \end{matrix} \\ \begin{bmatrix} 2.159 & 1.946 \\ \vdots & \vdots \\ 0.182 & 1.030 \end{bmatrix} \end{array}.$$

Center of a compositional data set

- The *center* of a set $\underline{\mathbf{W}}$ of n compositions $\underline{\mathbf{w}_1}$, $\ldots, \underline{\mathbf{w}_n}$ of $\mathcal{C}^{D-1}$, is the composition $\mathbf{g}$ defined by

$$\operatorname{cen} \underline{\mathbf{W}} = \mathbf{g} = (\frac{1}{n} \otimes \underline{\mathbf{w}_1}) \oplus \ldots \oplus (\frac{1}{n} \otimes \underline{\mathbf{w}_n}).$$

  This center is equal to

$$\mathbf{g} = \operatorname{ccl} \left(\left(\prod_{i=1}^{n} w_{i1} \right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} w_{iD} \right)^{1/n} \right)'.$$

- It verifies that

$$\operatorname{clr} \mathbf{g} = \overline{\mathbf{z}} = \sum_{i=1}^{n} \frac{1}{n} \mathbf{z}_i = \sum_{i=1}^{n} \frac{1}{n} \operatorname{clr} \underline{\mathbf{w}_i}.$$

  and

$$\operatorname{alr}_D \mathbf{g} = \overline{\mathbf{y}} = \sum_{i=1}^{n} \frac{1}{n} \mathbf{y}_i = \sum_{i=1}^{n} \frac{1}{n} \operatorname{alr}_D \underline{\mathbf{w}_i}.$$

Center of a compositional data set

Properties

$$\operatorname{cen} \underline{\mathbf{W}} = \operatorname{ccl} \left(\left(\prod_{i=1}^{n} w_{i1} \right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} w_{iD} \right)^{1/n} \right)'.$$

- $\operatorname{cen} \underline{\mathbf{W}} = \underbrace{\operatorname{argmin}}_{\xi \in \mathcal{C}^{D-1}} \left\{ \frac{d_{\mathcal{C}}(\underline{\mathbf{w}_1}, \xi) + \ldots + d_{\mathcal{C}}(\underline{\mathbf{w}_n}, \xi)}{n} \right\}$.
- $\operatorname{cen} \{\mathbf{p} \oplus \underline{\mathbf{W}}\} = \mathbf{p} \oplus \operatorname{cen} \underline{\mathbf{W}}$, where $\mathbf{p} \in \mathcal{C}^{D-1}$.
- $\operatorname{cen} \{t \otimes \underline{\mathbf{W}}\} = t \otimes \operatorname{cen} \underline{\mathbf{W}}$, where $t \in \mathbb{R}$.
- $\operatorname{cen} \{\underline{\mathbf{W}} \oplus \underline{\mathbf{W}^*}\} = \operatorname{cen} \underline{\mathbf{W}} \oplus \operatorname{cen} \underline{\mathbf{W}^*}$.

Center of a compositional data set

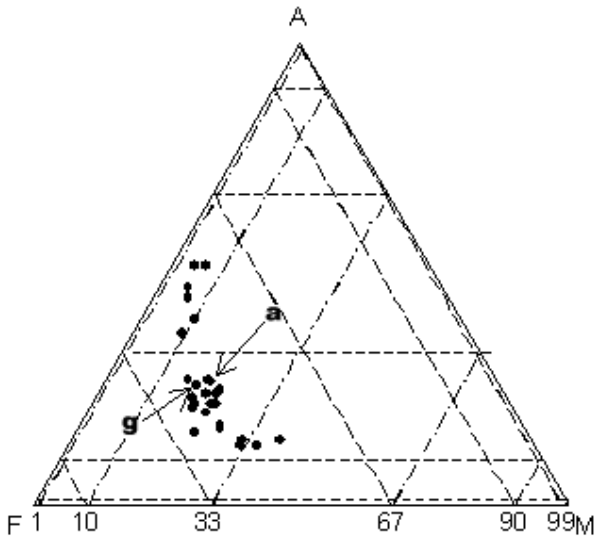
Example AFM composition of 23 aphyric Skye lavas

⋆ Compositional ("geometric") center:

$$\mathbf{g} = \mathrm{ccl}\,(25.85, 56.65, 17.50)'.$$

⋆ "Arithmetic" center:

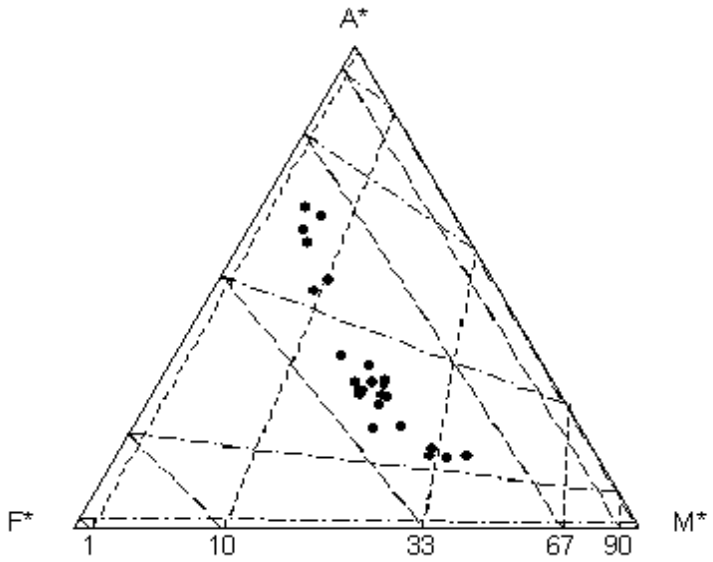
$$\mathbf{a} = \mathrm{ccl}\,(26.83, 53.74, 19.43)'.$$



Centering

To "centre" a compositional data set $\underline{\mathbf{w}_1}, \ldots \underline{\mathbf{w}_n}$ with centre $\mathbf{g}$, it suffices to consider the new data set $\underline{\mathbf{w}_1^*} = \mathbf{g}^{-1} \oplus \underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n^*} = \mathbf{g}^{-1} \oplus \underline{\mathbf{w}_n}$.

Obviously, the centre of the new "centered" data set $\underline{\mathbf{w}_1^*}, \ldots, \underline{\mathbf{w}_n^*}$ is $\mathrm{ccl}\,(1/D, \ldots, 1/D)'$.



Compositional covariance structure

- Variation matrix

$$\mathbf{T} = [\tau_{jk}] = \left[\text{var} \left\{ \log \frac{\mathbf{w}_{(j)}}{\mathbf{w}_{(k)}} \right\} \right].$$

  - ⋆ $\tau_{jk} = 0$ means a perfect relationship between $\mathbf{w}_{(j)}$ and $\mathbf{w}_{(k)}$ in the sense that the ratio $\mathbf{w}_{(j)}/\mathbf{w}_{(k)}$ is constant.
  - ⋆ The larger the value of τ_{jk} the more departure from proportionality between $\mathbf{w}_{(j)}$ and $\mathbf{w}_{(k)}$.
  - ⋆ A measure of degree of proportionality between two parts j and k is given by

$$\exp(-\sqrt{\tau_{jk}}).$$

    In this way, $\exp(-\sqrt{\tau_{jk}}) = 0$ means *zero propotionality*, and $\exp(-\sqrt{\tau_{jk}}) = 1$ means *perfect propotionality*.
  - ⋆ The variation matrix of any subcomposition is simply obtained by picking out on $\mathbf{T}$ all the logratio variances τ_{jk} associated with the parts j and k of the subcomposition.

Compositional covariance structure

- Logratio covariance matrix

$$\boldsymbol{\Sigma} = [\sigma_{jk}] \quad = \left[\text{cov}\left\{\mathbf{y}_{(j)}, \mathbf{y}_{(k)}\right\}\right]$$

$$= \left[\text{cov}\left\{\log \frac{\mathbf{w}_{(j)}}{\mathbf{w}_{(D)}}, \log \frac{\mathbf{w}_{(k)}}{\mathbf{w}_{(D)}}\right\}\right] .,$$

where

$$\mathbf{y}_j = \left(\log \frac{w_{1j}}{w_{1D}}, \ldots, \log \frac{w_{nj}}{w_{nD}}\right)',$$

for $j = 1, \ldots, D-1$.

Compositional covariance structure

- Centered covariance matrix

$$\mathbf{\Gamma} = [\gamma_{jk}] = \left[\text{cov}\left\{\mathbf{z}_{(j)}, \mathbf{z}_{(k)}\right\}\right].$$

where

$$\mathbf{z}_{(j)} = \left(\log\left(w_{1j}/g(\mathbf{w}_1)\right), \ldots, \log\left(w_{nj}/g(\mathbf{w}_n)\right)\right)',$$

for $j = 1, \ldots, D$.

Hint. Correlation $\text{corr}\left\{\mathbf{z}_{(j)}, \mathbf{z}_{(k)}\right\}$ is not a measure of a relationship between parts j and k because is subcompositionally incoherent.

- Total (relative) variability

$$\begin{aligned}\text{totvar}_{\mathcal{C}}\{\underline{\mathbf{W}}\} &= \textstyle\sum_{i=1}^{n} \frac{1}{n} \text{d}_{\mathcal{C}}^2(\underline{\mathbf{w}_i}, \mathbf{g}) = \text{trace}\{\mathbf{\Gamma}\} \\ &= \tfrac{1}{2D}\mathbf{1}_D' \mathbf{T} \mathbf{1}_D = \tfrac{1}{D}\textstyle\sum_{i<j} \tau_{ij}.\end{aligned}$$

Compositional covariance structure

- The centered covariance matrix $\mathbf{\Gamma} = \left[\gamma_{jk}\right]$ is singular because

$$\sum_{k=1}^{D} \gamma_{jk} = 0 \quad (j = 1, \ldots, D).$$

- The relationships between the three covariance matrices $\mathbf{T}, \mathbf{\Sigma}$ and $\mathbf{\Gamma}$ are linear.

- The dimensionality of the covariance structure of a compositional raw data matrix from $\mathcal{C}^{D-1}$ is equal to $\frac{1}{2}D(D-1)$.

- The covariance matrix $\mathbf{T}$ —and also $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$— is coherent with the algebraic structure of $(\mathcal{C}^{D-1}, \oplus, \otimes)$, i.e.,

$$\mathbf{T}\{\mathbf{p} \oplus \underline{\mathbf{W}}\} = \mathbf{T}\{\underline{\mathbf{W}}\} \quad \text{and} \quad \mathbf{T}\{\lambda \otimes \underline{\mathbf{W}}\} = \lambda^2 \mathbf{T}\{\underline{\mathbf{W}}\},$$

where $\underline{\mathbf{W}}$ is a compositional raw data matrix from $\mathcal{C}^{D-1}$, $\mathbf{p} \in \mathcal{C}^{D-1}$ and $\lambda \in \mathbb{R}$.

Therefore,

$$\text{totvar}_{\mathcal{C}}\{\mathbf{p} \oplus \underline{\mathbf{W}}\} = \text{totvar}_{\mathcal{C}}\{\underline{\mathbf{W}}\};$$

$$\text{totvar}_{\mathcal{C}}\{\lambda \otimes \underline{\mathbf{W}}\} = \lambda^2 \text{totvar}_{\mathcal{C}}\{\underline{\mathbf{W}}\}.$$

Compositional covariance structure

Example AFM composition of 23 Skye lavas

⋆ Variation matrix

$$\mathbf{T} = \begin{bmatrix} 0 & 0.251 & 1.144 \\ 0.251 & 0 & 0.350 \\ 1.144 & 0.350 & 0 \end{bmatrix}.$$

⋆ Logratio covariance matrix

$$\mathbf{\Sigma}_A = \begin{bmatrix} 0.251 & 0.523 \\ 0.523 & 1.144 \end{bmatrix} \quad \mathbf{\Sigma}_F = \begin{bmatrix} 0.251 & -0.271 \\ -0.271 & 0.350 \end{bmatrix}$$

$$\mathbf{\Sigma}_M = \begin{bmatrix} 1.144 & 0.622 \\ 0.622 & 0.350 \end{bmatrix}.$$

⋆ Centered covariance matrix

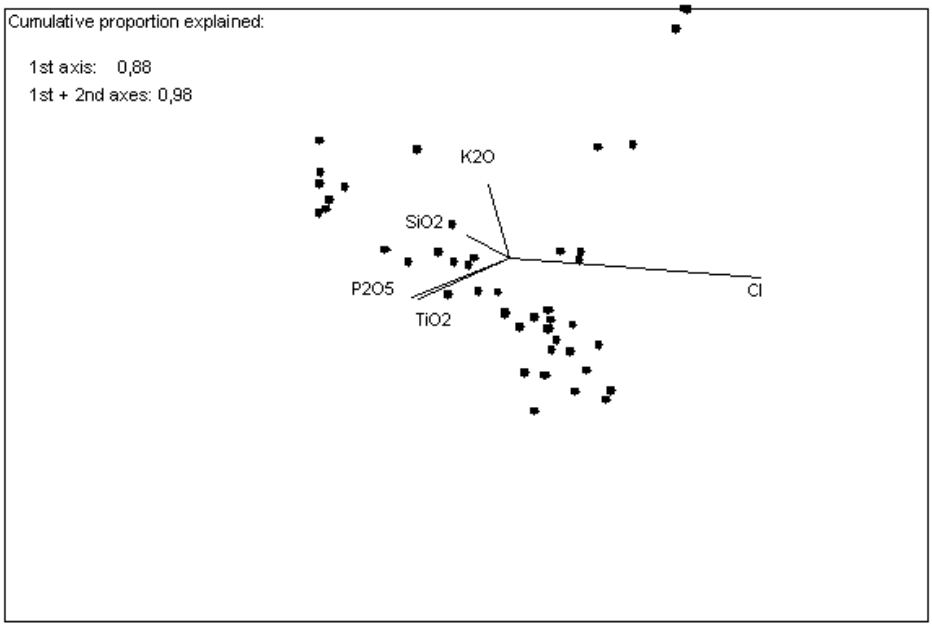
$$\mathbf{\Gamma} = \begin{bmatrix} 0.271 & 0.013 & -0.284 \\ 0.013 & 0.007 & -0.020 \\ -0.284 & -0.020 & 0.304 \end{bmatrix}.$$

⋆ Total variability: $\text{totvar}_{\mathcal{C}} = 0.582$.

Biplots

In general, the *biplot* is a simultaneous representation of the rows (observations) and columns (variables) of a $n \times p$ matrix $\mathbf{X}$ by means of a rank-2 approximation.

Usually, biplot analysis starts with performing some transformations on $\mathbf{X}$, depending on the nature of the data, to obtain a transformed matrix $\mathbf{Z}$ which is the one that is actually displayed.



Biplots

- The singular value decomposition (SVD) of $\mathbf{Z}$ provides a decomposition of this matrix:

  $$\mathbf{Z} = [\mathbf{u}_1 : \ldots : \mathbf{u}_r] \, \mathrm{diag}\{\lambda_1, \ldots, \lambda_r\} \, [\mathbf{v}_1 : \ldots : \mathbf{v}_r]',$$

  where

  - $\star$ r is the rank of $\mathbf{Z}$;
  - $\star$ $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are the standardized eigenvectors of $\mathbf{Z}'$;
  - $\star$ $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are the standardized eigenvectors of $\mathbf{Z}$;
  - $\star$ and $\lambda_1, \ldots, \lambda_r$ the corresponding positive eigenvalues in decreasing order.

- From this SVD of $\mathbf{Z}$, and using only the two first eigenvectors, a rank-2 approximation $\widehat{\mathbf{Z}}$ is obtained:

  $$\widehat{\mathbf{Z}} = [\mathbf{u}_1 : \mathbf{u}_2] \, \mathrm{diag}\{\lambda_1, \lambda_2\} \, [\mathbf{v}_1 : \mathbf{v}_2]'.$$

Biplots

- Then $\widehat{\mathbf{Z}}$ decomposes in

$$\hat{\mathbf{Z}} = \underbrace{[\lambda_1^{\alpha}\mathbf{u}_1 : \lambda_2^{\alpha}\mathbf{u}_2]}_{\mathbf{F}} \underbrace{[\lambda_1^{1-\alpha}\mathbf{v}_1 : \lambda_2^{1-\alpha}\mathbf{v}_2]'}_{\mathbf{G}'},$$

  where α is an arbitrary constant.

- The biplot represents simultaneously in $\mathbb{R}^2$ the rows of $\mathbf{F}$, which provides the coordinates of n points (in correspondence with the n rows/observations of $\mathbf{Z}$) , and the rows of $\mathbf{G}$, which provides the coordinates of p points (in correspondence with the columns/variables of $\mathbf{Z}$).

  Conventionally, the biplot depicts the variables by *rays* and the observations by points.

Depending on the constant α, the biplot favours the display of rows (observations) or columns (variables). For $\alpha = 0$, the biplot is called *covariance biplot*. In this case, the display of variables is favoured.

Biplots

- Singular value decomposition (SVD) of $\mathbf{Z}$:

  $$\mathbf{Z} = [\mathbf{u}_1 : \ldots : \mathbf{u}_r] \operatorname{diag}\{\lambda_1, \ldots, \lambda_r\} [\mathbf{v}_1 : \ldots : \mathbf{v}_r]',$$

- Rank-2 approximation $\hat{\mathbf{Z}}$:

  $$\hat{\mathbf{Z}} = \underbrace{[\lambda_1^{\alpha}\mathbf{u}_1 : \lambda_2^{\alpha}\mathbf{u}_2]}_{\mathbf{F}} \underbrace{[\lambda_1^{1-\alpha}\mathbf{v}_1 : \lambda_2^{1-\alpha}\mathbf{v}_2]'}_{\mathbf{G}'},$$

- The ratio

  $$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_r}$$

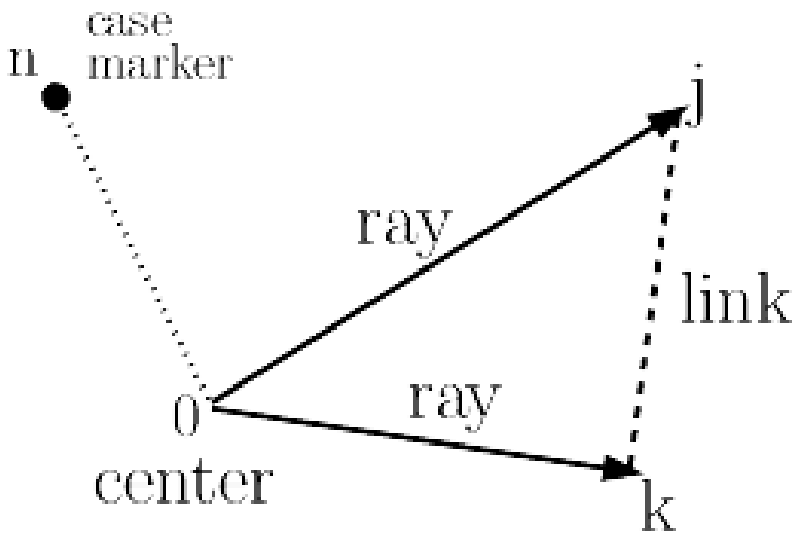
  is a measure of the proportion of the "variability" of $\mathbf{Z}$ captured by the biplot.

Relative variation diagrams

Definition The *relative variation diagram* of a compositional data set $\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}$ of $\mathcal{C}^{D-1}$ is the covariance biplot of the matrix $\mathbf{Z}_c$ which we obtain after centering the D columns of the centered logratio matrix $\mathbf{Z}$.

- Elements
  - ⋆ *Origin*, labeled O.
  - ⋆ *Vertices*, for each of the D parts (variables/columns) of compositions, labeled $1, \ldots, j, \ldots D$.
  - ⋆ *Case marker*, for each of the n observations (rows), labeled $1, \ldots, i, \ldots n$.
  - ⋆ *Ray*. Is the join Oj of origin O to a vertex j.
  - ⋆ *Link*. Is the join jk of two vertices j and k.

Relative variation diagrams



Relative variation diagrams

- The vertices and case markers are both centered at the origin O.
- Rays and inter-ray angles represent the centered logratio matrix $\mathbf{\Gamma}$:

$$|Oj|^2 = \hat{\gamma}_{jj} = \text{estimate of var}\left\{\mathbf{z}_{(j)}\right\},$$
$$|Oj| \cdot |Ok| = \hat{\gamma}_{jk} = \text{estimate of cov}\left\{\mathbf{z}_{(j)}, \mathbf{z}_{(k)}\right\},$$

so that

$$\cos \widehat{jOk} = \text{estimate of corr}\left\{\mathbf{z}_{(j)}, \mathbf{z}_{(k)}\right\}.$$

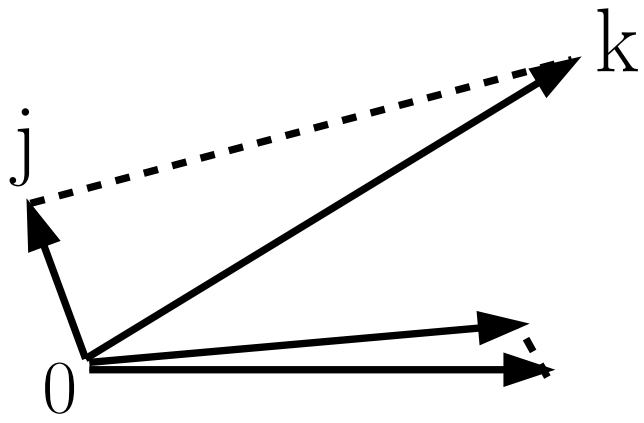
Hint. Remember that correlation $\text{corr}\left\{\mathbf{z}_{(j)}, \mathbf{z}_{(k)}\right\}$ is not a measure of a relationship between parts j and k because is subcompositionally incoherent.

Relative variation diagrams

- The squared lengths of the links represent the set of estimated relative variances:

$$|jk|^2 = \hat{\tau}_{jk} = \text{estimate of var} \left\{ \log \frac{\mathbf{w}_{(j)}}{\mathbf{w}_{(k)}} \right\}.$$

Therefore, if two vertices j and k coincide or are close together then components $\mathbf{w}_{(j)}$ and $\mathbf{w}_{(k)}$ are in constant proportion or nearly so.



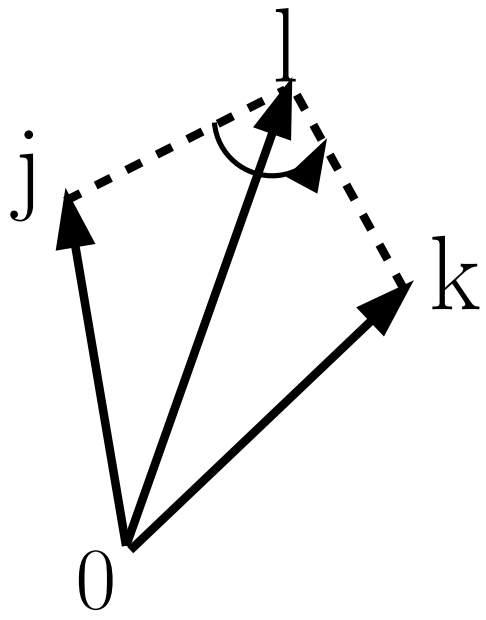
Relative variation diagrams

- Links jl and kl, with a common vertex l, represent the estimated logratio covariance matrix $\widehat{\mathbf{\Sigma}}_l$:

$$|jl| \cdot |kl| \approx \left| \text{cov} \left\{ \log \frac{\mathbf{w}_{(j)}}{\mathbf{w}_{(l)}}, \log \frac{\mathbf{w}_{(k)}}{\mathbf{w}_{(l)}} \right\} \right|,$$

so that

$$\cos \widehat{jlk} \approx \text{corr} \left\{ \log \frac{\mathbf{w}_{(j)}}{\mathbf{w}_{(l)}}, \log \frac{\mathbf{w}_{(k)}}{\mathbf{w}_{(l)}} \right\}.$$

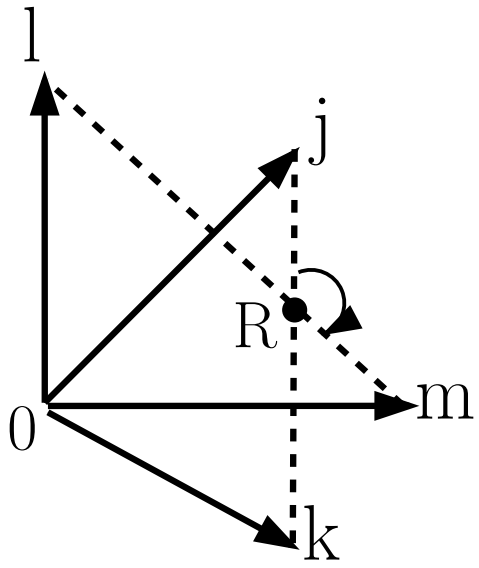


Relative variation diagrams

- If the links jk and lm intersect at R then

$$\left|\cos \widehat{jRm}\right| \approx \left|\text{corr}\left\{\log \frac{\mathbf{w}_{(j)}}{\mathbf{w}_{(k)}}, \log \frac{\mathbf{w}_{(l)}}{\mathbf{w}_{(m)}}\right\}\right|.$$

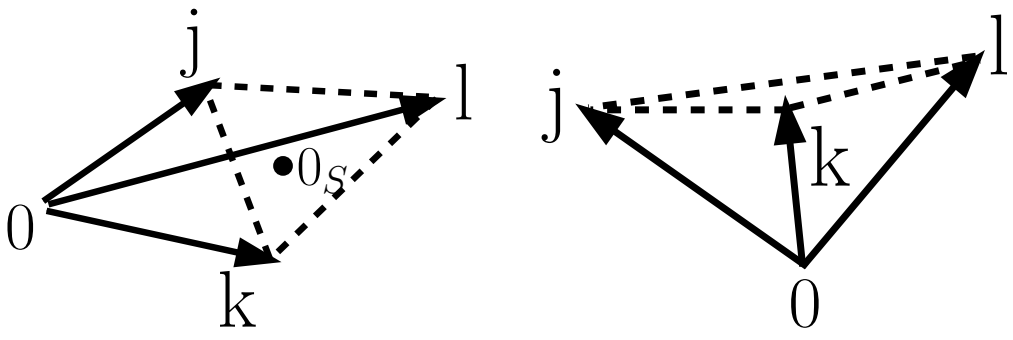
  Therefore, if two links jk and lm intersect at right angles then the logratios $\log(\mathbf{w}_{(j)}/\mathbf{w}_{(k)})$ and $\log(\mathbf{w}_{(l)}/\mathbf{w}_{(m)})$ will be uncorrelated and, within the context of logistic normality, independent, i.e., subcompositions (j, k) and (l, m) are independent.



Relative variation diagrams

- The relative variation diagram for any subcomposition S is simply the subdiagram formed by selecting the vertices corresponding to the parts of the subcomposition and taking the centroide O_S of these vertices as the center of the subcompositional biplot.

  Therefore, if a subset — say $1, \ldots, C$— of vertices is approximately collinear then the associated subcomposition has a "compositional" one-dimensional structure.



Volcano H

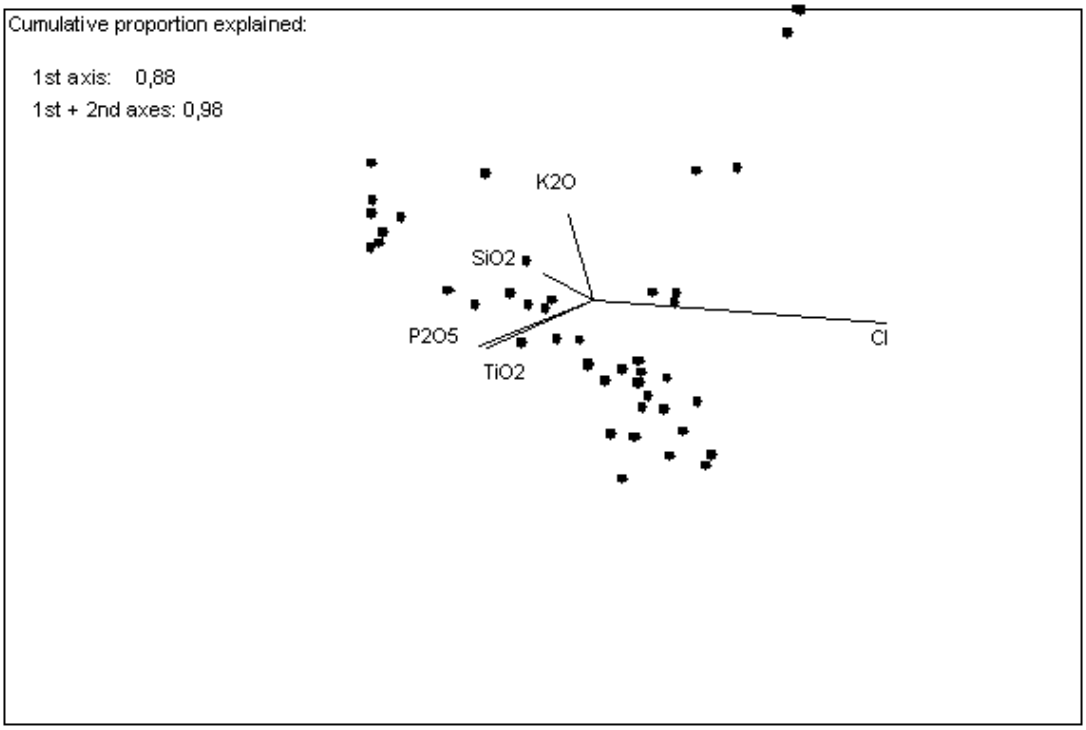
- Parts: 1=Cl; 2=K_2O; 3=P_2O_5; 4=TiO_2; 5=SiO_2.
- Variation matrix $\mathbf{T}$

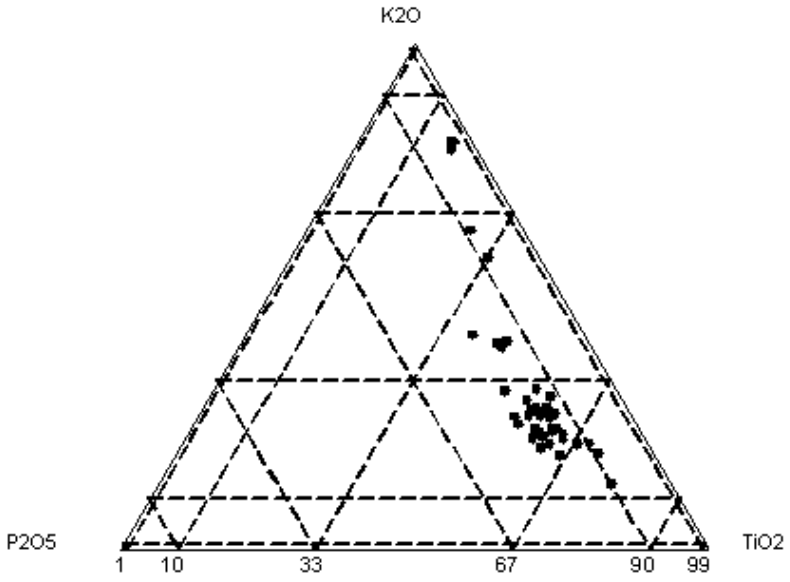
$$\begin{bmatrix} 0 & 2,784 & 4,134 & 3,970 & 2,966 \\ 2,784 & 0 & 0,647 & 0,645 & 0,146 \\ 4,134 & 0,647 & 0 & 0,071 & 0,304 \\ 3,970 & 0,645 & 0,071 & 0 & 0,249 \\ 2,966 & 0,146 & 0,304 & 0,249 & 0 \end{bmatrix}$$

- Centered covariance matrix $\mathbf{\Gamma}$

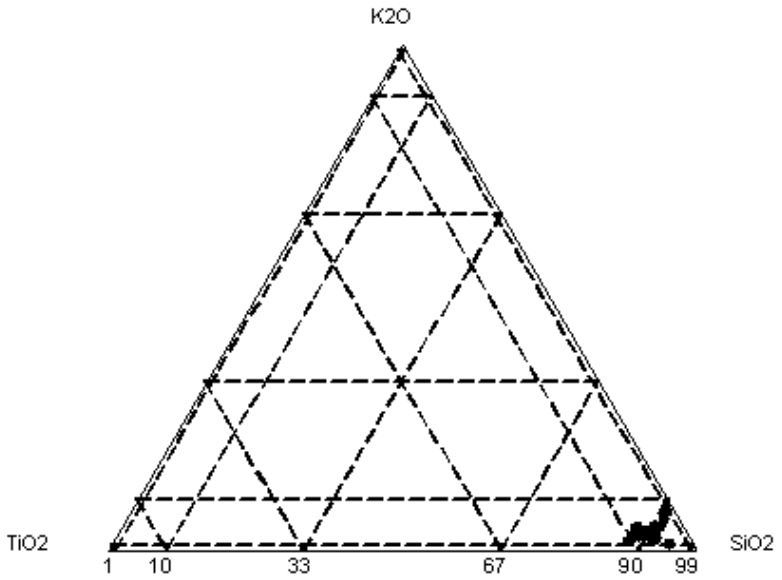
$$\begin{bmatrix} 2,134 & -0,221 & -0,803 & -0,743 & -0,368 \\ -0,221 & 0,208 & -0,022 & -0,043 & 0,079 \\ -0,803 & -0,022 & 0,394 & 0,337 & 0,094 \\ -0,743 & -0,043 & 0,337 & 0,350 & 0,099 \\ -0,368 & 0,079 & 0,094 & 0,099 & 0,096 \end{bmatrix}$$

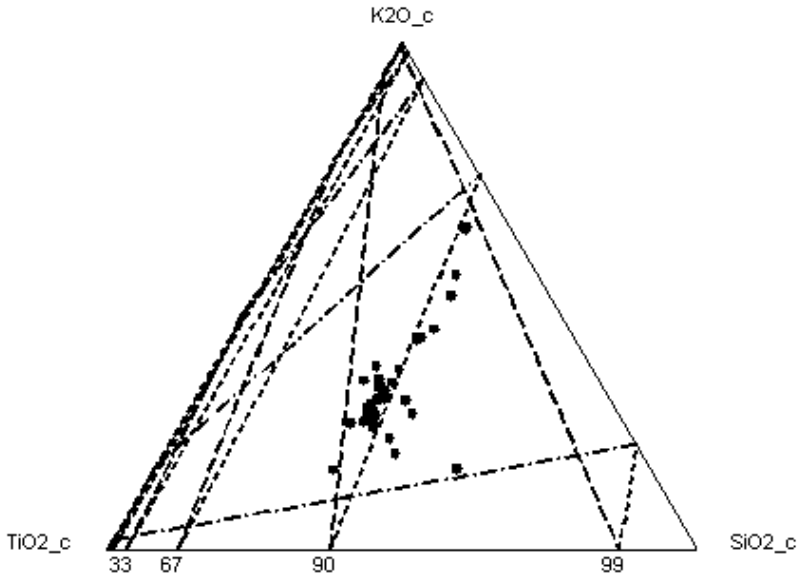
Volcano H


Cumulative proportion explained:
1st axis: 0,88
1st + 2nd axes: 0,98
K2O
SiO2
P2O5
TiO2
Cl


K2O
P2O5
TiO2
1
10
33
67
90
99

Volcano H





Dimension-reducing techniques

Compositional PCA

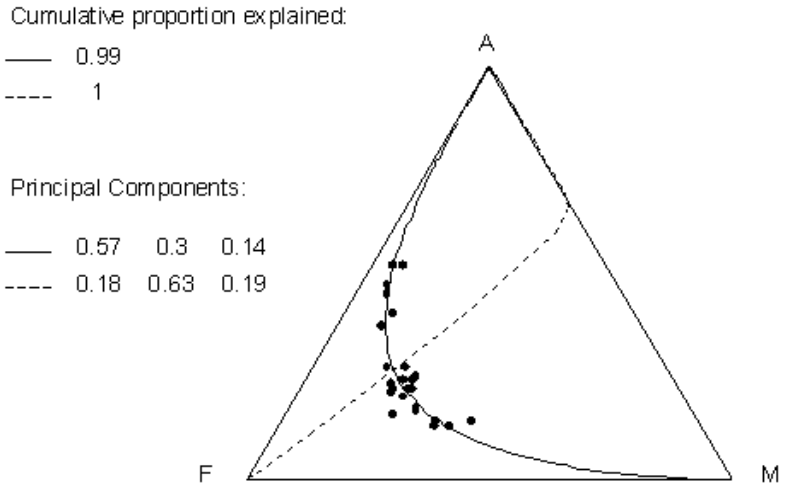
Given a set of compositions $\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}$ of $\mathcal{C}^{D-1}$ with center $\mathbf{g}$, the PCA will start looking for a direction —determined by a $\mathcal{C}$-unitary composition $\mathbf{c}_1$— such that the total variability of the $\mathcal{C}$-orthogonal projections of $\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}$ on the compositional straight line through $\mathbf{g}$ with direction $\mathbf{c}_1$ will be maximum. And so on.

Property The compositional principal components of a set of compositions $\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}$ of $\mathcal{C}^{D-1}$ can be determined from the standard principal components of the clr -transformed observations $\operatorname{clr} \underline{\mathbf{w}_1}, \ldots, \operatorname{clr} \underline{\mathbf{w}_n}$.

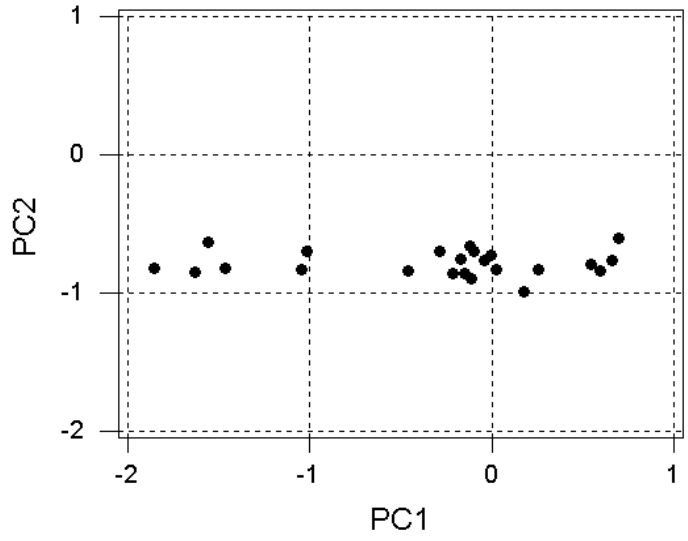
Compositional PCA

In this manner, the positive eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{D-1}$ of the centered logratio covariance matrix $\mathbf{\Gamma}$ give the decomposition of totvar$_{\mathcal{C}}$, and the corresponding unitary eigenvectors $\mathbf{z}_1^*, \ldots, \mathbf{z}_{D-1}^*$ determine the corresponding directions $\mathrm{clr}^{-1}\mathbf{z}_1^*, \ldots, \mathrm{clr}^{-1}\mathbf{z}_{D-1}^*$ of the principal axes.

Skye Lavas



PC1: $0.4436 \log A - 0.8154 \log F + 0.3719 \log M \approx -0.7849$

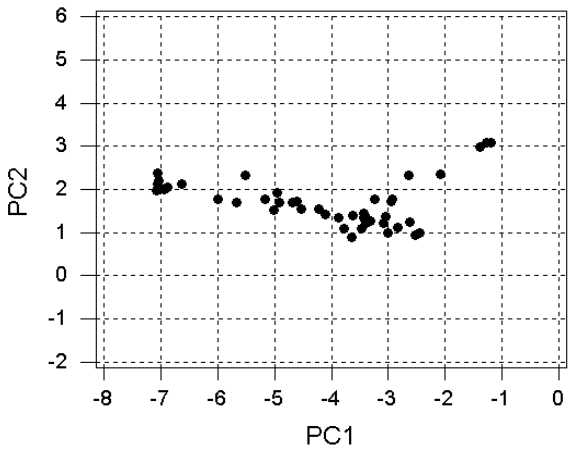


Volcano H

(Cl, K_2O, P_2O_5, TiO_2, SiO_2)

PC_1: (0.4246,0.1656,0.1264,0.1296,0.1538)' (87.8%)

PC_2: (0.1469,0.3836,0.1230,0.1182,0.2293)' (acum 98.1%)



Dimension-reducing techniques

Subcomposition analysis

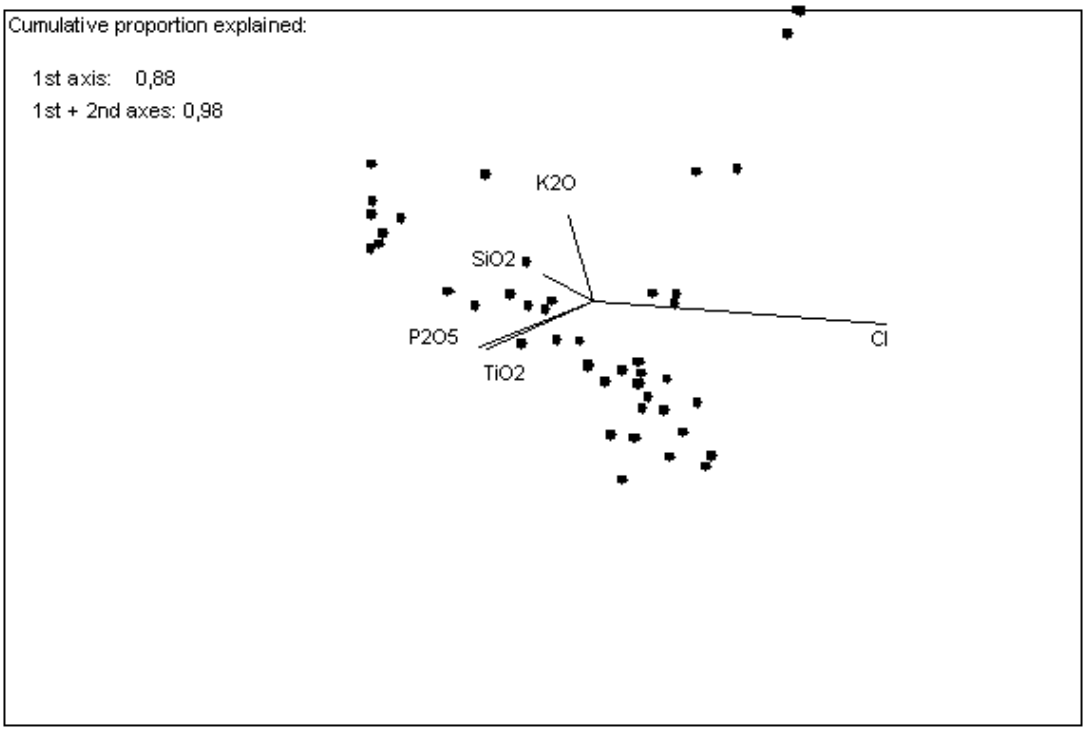
Let $\underline{\mathbf{w}}_1, \ldots, \underline{\mathbf{w}}_n$ be a compositional data set of $\mathcal{C}^{D-1}$, and let $\mathrm{sub}_S\, \underline{\mathbf{w}}_1, \ldots, \mathrm{sub}_S\, \underline{\mathbf{w}}_n$ be the set of the corresponding subcompositions of $\mathcal{C}^{C-1}$ associated to a subset S of parts $1, \ldots, D$. Then, the ratio

$$\frac{\mathrm{totvar}_{\mathcal{C}}\{\mathrm{sub}_S\, \underline{\mathbf{w}}_1, \ldots, \mathrm{sub}_S\, \underline{\mathbf{w}}_n\}}{\mathrm{totvar}_{\mathcal{C}}\{\underline{\mathbf{w}}_1, \ldots, \underline{\mathbf{w}}_n\}}$$

gives the proportion of total variabilitiy retained by the subcompositions.

If the purpose of subcompositional analysis is to retain as much variability as possible for a given number C of parts, then we have to search for subcompositions of this size which maximize this ratio.

Volcano H


Cumulative proportion explained:
1st axis: 0,88
1st + 2nd axes: 0,98
K2O
SiO2
P2O5
TiO2
Cl

Subcomposition analysis

Example Percentage of Cl, K_2O, P_2O_5, TiO_2 and SiO_2 in 46 samples of volcanic rocks from an anonymous volcano H

- ⋆ Total variability: $\text{totvar}_{\mathcal{C}} = 3.1829$.
- ⋆ Total variability of 3-parts subcompositions:

Subcomposition	Percentage
P_2O_5, TiO_2, SiO_2	6.53%
K_2O, TiO_2, SiO_2	10.90%
K_2O, P_2O_5, SiO_2	11.48%
K_2O, P_2O_5, TiO_2	14.27%
Cl, K_2O, SiO_2	61.74%
Cl, TiO_2, SiO_2	75.25%
Cl, K_2O, TiO_2	77.48%
Cl, P_2O_5, SiO_2	77.53%
Cl, K_2O, P_2O_5	79.21%
Cl, P_2O_5, TiO_2	**85.61%**

Zeros in compositional data

- Logratio methodology is incompatible with composition with zeros in one or more parts.
- Two kinds of zeros:
  - ⋆ *Essential zeros*: part completely absent.
  - ⋆ *Rounded zeros*: no quantifiable proportion has been recorded.
- Treatment of essential zeros:
  - ⋆ Is it suitable to *amalgamate* some parts?
  - ⋆ Pre-classification: create initial groups according to the number and location of zeros, and analyze each group separately.
- Treatment of rounded zeros:
  - ⋆ Consider the zero values as missing values.
  - ⋆ Imputation: replace zero values by a small amount using non-parametric or parametric techniques.
  - ⋆ Apply log-ratio methodology to replaced observations of resulting data set.

Rounded zeros

Multiplicative replacement

Let be $\underline{\mathbf{w}} = \text{ccl}\,(w_1, \ldots, w_D)' \in \mathcal{C}^{D-1}$ any composition with some $w_j = 0$ (rounded zero).

The **multiplicative replacement** replaces $\underline{\mathbf{w}}$ by the composition $\underline{\mathbf{w}^{(r)}} = \text{ccl}\,(w_1^{(r)}, \ldots, w_D^{(r)})'$ defined by

$$\text{if } w_j = 0 \rightarrow w_j^{(r)} = \delta_j;$$

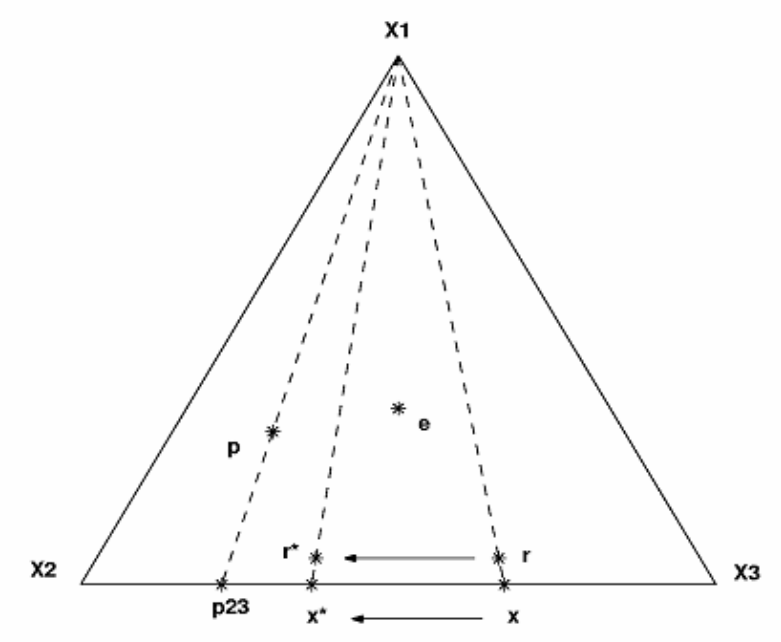
$$\text{if } w_j \neq 0 \rightarrow w_j^{(r)} = w_j \left(1 - \sum_{w_l=0} \delta_l\right).$$

where δ_j are the "small" values replacing zeros parts.

Rounded zeros

Multiplicative replacement

- It is a "natural" replacement.
- Ratio between two non-zero parts is preserved.
- It is compatible with subcompositions, perturbation and power transformation.
- Covariance structure of subcompositions with no zeros is preserved.



Modeling compositional data

In practice, many of the probability density functions (pdf) on the compositional space $\mathcal{C}^{D-1}$ will be defined from a pdf on the real space $\mathbb{R}^{D-1}$. Then the $\mathrm{alr}_j{}^{-1}$ transformations will allow to induce on the simplex $\mathcal{S}^D$ the corresponding pdf.

The most important pdf on $\mathcal{C}^{D-1}$ are:

- The Dirichlet class.
- The (additive) logistic normal class.
- The (additive) logistic skewnormal class.

Definition A random composition $\underline{\mathbf{w}}$ on $\mathcal{C}^{D-1}$ is said to have an *additive logistic normal* distribution (aln) of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ —written $\underline{\mathbf{w}} \sim \mathcal{L}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$— if the random vector $\mathbf{y} = \mathrm{alr}_D \underline{\mathbf{w}} = \log(\mathbf{w}_{-D}/w_D)$ has a $\mathcal{N}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on $\mathbb{R}^{D-1}$.

Logistic normal distributions on $\mathcal{C}^{D-1}$

Property Let $\underline{\mathbf{w}}$ be a random vector on $\mathcal{C}^{D-1}$. If $\mathrm{alr}_D\,\underline{\mathbf{w}} \sim \mathcal{N}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then all the other logratio random vectors $\mathrm{alr}_j\underline{\mathbf{w}}$ $(j = 1, \ldots, d)$ are normally distributed.

Property Let $\underline{\mathbf{w}}$ be a random composition on $\mathcal{C}^{D-1}$. Let $\underline{\mathbf{w}_S}$ be the random subcomposition on $\mathcal{C}^{C-1}$ corresponding to a subset S of C parts of $\underline{\mathbf{w}}$. If $\underline{\mathbf{w}} \sim \mathcal{L}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\underline{\mathbf{w}_S} \sim \mathcal{L}^{C-1}(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$, where $\boldsymbol{\mu}_S$ and $\boldsymbol{\Sigma}_S$ can be easily calculated from $\boldsymbol{\mu}$ and Σ.

Property Let $\underline{\mathbf{w}}$ be a random vector on $\mathcal{C}^{D-1}$, which $\underline{\mathbf{w}} \sim \mathcal{L}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If we perturb $\underline{\mathbf{w}}$ by a constant composition $\underline{\mathbf{p}} \in \mathcal{C}^{D-1}$, then the perturbed random vector $\underline{\mathbf{p}} \oplus \underline{\mathbf{w}} \sim \mathcal{L}^{D-1}(\boldsymbol{\mu} + \mathrm{alr}_D\,\underline{\mathbf{p}}, \boldsymbol{\Sigma})$.

Logistic normal distributions on $\mathcal{C}^{D-1}$

Estimation of parameters

To estimate the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a random composition $\underline{\mathbf{w}} \sim \mathcal{L}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from a random sample $\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}$ of $\underline{\mathbf{w}}$, we estimate by standard procedures the vector mean and the covariance matrix of a multivariate normal distribution from the alr_D-transformed random sample

$$\mathbf{y}_1 = \mathrm{alr}_D \underline{\mathbf{w}_1}, \ldots, \mathbf{y}_n = \mathrm{alr}_D \underline{\mathbf{w}_n}.$$

The maximum likelihood estimations of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by

$$\widehat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ij},$$

$$\widehat{\sigma}_{jk} = \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \widehat{\mu}_j)(y_{ik} - \widehat{\mu}_k),$$

for $j, k = 1, \ldots, D-1$.

Predictive regions

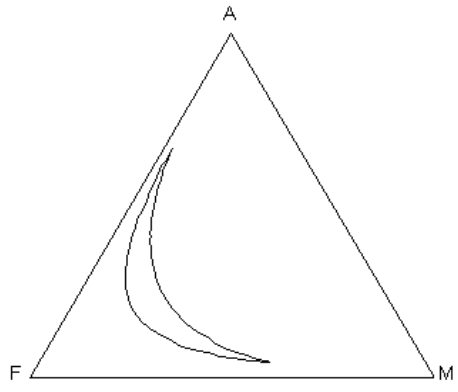
Definition Let $\underline{\mathbf{w}}$ be a random composition *aln* distributed on $\mathcal{C}^{D-1}$. If $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the estimates of the unknown parameters of $\underline{\mathbf{w}}$ from a random sample of size n, the $1-\alpha$ *predictive region* is defined as

$$\left\{\underline{\mathbf{w}^*} \in \mathcal{C}^{D-1} : \ (\mathrm{alr}_D\, \underline{\mathbf{w}^*} - \widehat{\boldsymbol{\mu}})' \widehat{\boldsymbol{\Sigma}}^{-1} (\mathrm{alr}_D\, \underline{\mathbf{w}^*} - \widehat{\boldsymbol{\mu}}) \leq r^2 \right\},$$

where r^2 is a real number such that

$$\mathrm{Prob}\left[\mathcal{F}_{D-1,n-(D-1)} \leq \frac{n(n-(D-1))}{(n^2-1)(D-1)} r^2\right] = 1-\alpha.$$

99% Predictive region - Skye lavas



ALN Predictive Region. Alpha = 0.99

Atypicality index

Definition If a random composition $\underline{\mathbf{w}}$ on $\mathcal{C}^{D-1}$ is $\mathcal{L}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributed, the *atypicality index* of a composition $\underline{\mathbf{w}^*} \in \mathcal{C}^{D-1}$ in relation to the random composition $\underline{\mathbf{w}}$ is defined as

$$\text{Prob}\left[\chi^2_{D-1} \leq (\text{alr}_D\, \underline{\mathbf{w}^*} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\text{alr}_D\, \underline{\mathbf{w}^*} - \boldsymbol{\mu})\right].$$

Definition Let $\underline{\mathbf{w}}$ be a random composition *aln* distributed on $\mathcal{C}^{D-1}$. If $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the estimates of the unknown parameters of $\underline{\mathbf{w}}$ from a random sample $\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}$ of size n, the *atypicality index* of a composition $\underline{\mathbf{w}^*} \in \mathcal{C}^{D-1}$ in relation to the compositional data set $\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}$ is defined as

$$\text{Prob}\left[\mathcal{F}_{D-1,n-(D-1)} \leq k(\text{alr}_D\, \underline{\mathbf{w}^*} - \widehat{\boldsymbol{\mu}})' \widehat{\boldsymbol{\Sigma}}^{-1} (\text{alr}_D\, \underline{\mathbf{w}^*} - \widehat{\boldsymbol{\mu}})\right],$$

where $k = \frac{n(n-(D-1))}{(n^2-1)(D-1)}$.

Compositional Regression

Artic lake

Sand, silt,clay composition of 39 sediment samples at different water depths in an Artic lake:

Num.	Sand	Silt	Clay	Depth (m)
S01	77.5	19.5	3.0	10.4
S02	71.9	24.9	3.2	11.7
⋮	⋮	⋮	⋮	⋮
S39	2.0	47.8	50.2	103.7

a. Is sediment composition dependent on water depth?

b. If so, how can we quantify the extent of the dependence?

Compositional Regression

- Compositions $\underline{\mathbf{w}_i} \in \mathcal{C}^{D-1}$ regressing on a real concomitant t_i $(i = 1, \ldots, n)$:

$$\underline{\mathbf{w}_i} = \underline{\boldsymbol{\beta}_0} \oplus (t_i \otimes \underline{\boldsymbol{\beta}_1}) \otimes \underline{\varepsilon_i} \quad (i = 1, \ldots, n),$$

  where

  - $\star$ $\underline{\boldsymbol{\beta}_0}$: constant;
  - $\star$ $\underline{\boldsymbol{\beta}_1}$: regression coefficient;
  - $\star$ $\underline{\varepsilon_i}$ $(i = 1, \ldots, n)$: errors.

- alr version of the regression model

$$\operatorname{alr} \underline{\mathbf{w}_i} = \operatorname{alr} \underline{\boldsymbol{\beta}_0} + t_i \operatorname{alr} \underline{\boldsymbol{\beta}_1} + \operatorname{alr} \underline{\varepsilon_i} \quad (i = 1, \ldots, n).$$

  Can be reparametrized as

$$\operatorname{alr} \underline{\mathbf{w}_i} = \boldsymbol{\alpha}_0 + t_i \boldsymbol{\alpha}_1 + \boldsymbol{\epsilon}_i \quad (i = 1, \ldots, n).$$

Compositional Regression

$$\operatorname{alr} \underline{\mathbf{w}_i} = \boldsymbol{\alpha}_0 + t_i \boldsymbol{\alpha}_1 + \boldsymbol{\epsilon}_i \quad (i = 1, \ldots, n).$$

- Estimations $\widehat{\boldsymbol{\alpha}_0}$ and $\widehat{\boldsymbol{\alpha}_1}$ are obtained by application of the least squares method. Then

$$\widehat{\boldsymbol{\beta}_0} = \operatorname{alr}^{-1} \widehat{\boldsymbol{\alpha}_0} \quad , \quad \widehat{\boldsymbol{\beta}_1} = \operatorname{alr}^{-1} \widehat{\boldsymbol{\alpha}_1}$$

- The error (residual) of $\underline{\mathbf{w}_i}$ $(i = 1, \ldots, n)$ will be

$$\mathrm{e}_i = \underline{\mathbf{w}_i} \ominus \underline{\widehat{\mathbf{w}_i}},$$

  where $\underline{\widehat{\mathbf{w}_i}} = \widehat{\boldsymbol{\beta}_0} \oplus (t_i \otimes \widehat{\boldsymbol{\beta}_1})$.

- Sum of squares of errors:

$$SSError = \sum_{i=1}^{n} \|\mathrm{e}_i\|_{\mathcal{C}}^2 = \sum_{i=1}^{n} \left(d_{\mathcal{C}}(\underline{\mathbf{w}_i}, \underline{\widehat{\mathbf{w}_i}}) \right)^2 .$$

- Proportion of variability explained by the fitted linear regression model:

$$1 - \frac{SSError}{totvar_{\mathcal{C}}\{\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}\}}.$$

Artic lake

Num.	Sand	Silt	Clay	Depth (m)
S01	77.5	19.5	3.0	10.4
⋮	⋮	⋮	⋮	⋮
S39	2.0	47.8	50.2	103.7

- alr fitted simple linear regression model:

$$\log(\text{sand/clay}) = 9.697 - 2.743\log(\text{depth}) + \epsilon_1;$$

$$\log(\text{silt/clay}) = 4.805 - 1.096\log(\text{depth}) + \epsilon_2.$$

- Fitted regression model in $\mathcal{S}^3$:

$$\text{ccl}_L\,(\text{sand}, \text{silt}, \text{clay})' = (0.992849, 0.007145, 0.00006)' \oplus$$

$$\log(\text{depth}) \otimes (0.04604, 0.238291, 0.71505)'.$$

- Proportion of variability explained by the fitted simple linear regression model:

$$1 - \frac{0.7006}{2.4692} = 0.716 \equiv 71.6\%.$$

Artic lake

Num.	Sand	Silt	Clay	Depth (m)
S01	77.5	19.5	3.0	10.4
⋮	⋮	⋮	⋮	⋮
S39	2.0	47.8	50.2	103.7