The sample space of compositional data.

J. J. Egozcue

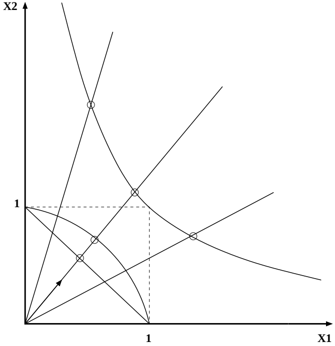
Dep. Matemática Aplicada III
UPC, Barcelona

Dep. EIO de la UPC
Barcelona
24 Noviembre, 2006

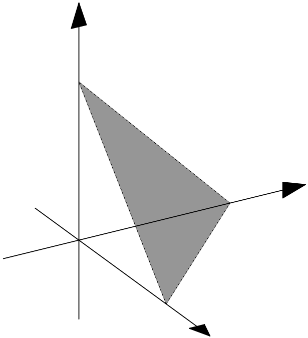
Summary

compositional data

- parts of some whole which only carry relative information
- typical units: parts per one, percentages, ppm, molar concentration...



compositional data in $\mathbb{R}^2$



simplex $\mathcal{S}^3 \subset \mathbb{R}^3$

sample space of compositional data

- the simplex (for κ a constant)

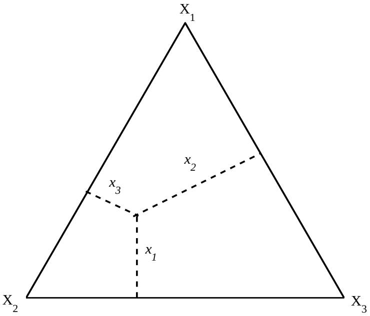
$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, \ldots, x_D] \in \mathbb{R}^D \;\middle|\; x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}$$

- compositional data are equivalence classes
  $\Rightarrow$ the value of κ is not important
- representation: ternary diagram

Closure operator: $\mathcal{C}\mathbf{x}$ normalizes to κ.

ternary diagram

For 3-part compositions,



milestone I: Karl Pearson, 1897

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

milestone II: Felix Chayes, 1960

- “On correlation between variables of constant sum”
- Chayes showed that correlations between closed data are induced by numerical constraints (negative bias or closure problem) and made attempts to separate the *spurious* part from the *real* correlation

example of *negative bias* and *spurious correlation*

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)

sample	x_1	x_2	x_3	x_4
1	0.1	0.2	0.1	0.6
2	0.2	0.1	0.2	0.5
3	0.3	0.3	0.1	0.3

sample	x'_1	x'_2	x'_3
1	0.25	0.50	0.25
2	0.40	0.20	0.40
3	0.43	0.43	0.14

correl	x_1	x_2	x_3	x_4
x_1	1.00	0.50	0.00	-0.98
x_2		1.00	-0.87	-0.65
x_3			1.00	0.19
x_4				1.00

correl	x'_1	x'_2	x'_3
x'_1	1.00	-0.57	-0.05
x'_2		1.00	-0.79
x'_3			1.00

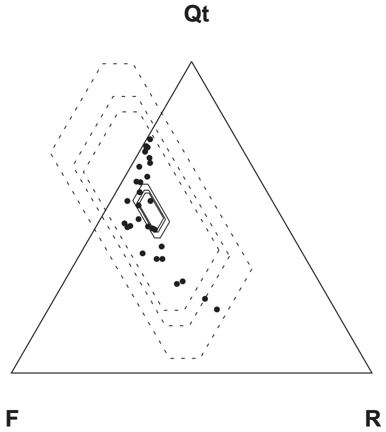
$$\mathbf{x} = [x_1, x_2, x_3, x_4]$$

$$\mathbf{x}' = \mathcal{C}[x_1, x_2, x_3]$$

attempts to model compositional uncertainty

hexagonal fields of variation employed in sedimentary petrology (error polygon)
limits 90%, 95%, 99%

illustration from Weltje (2006)



naive modelling

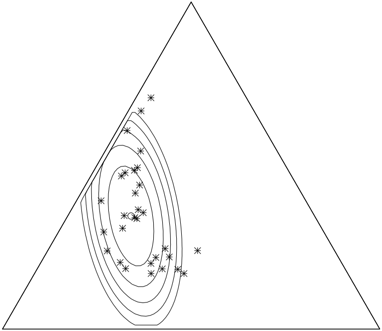


Figure: normal in $\mathbb{R}^2$

milestone III: John Aitchison, 1982

- "The statistical analysis of compositional data"
- as parts of a composition give only relative information, Aitchison suggested to use transformations based on log-ratios, e.g.
  - $\mathrm{alr}: \mathcal{S}^D \to \mathbb{R}^{D-1}, \quad \mathrm{alr}(\mathbf{x}) = \left[\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}\right]$
  - $\mathrm{clr}: \mathcal{S}^D \to \mathbb{R}^{D}, \quad \mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right]$

  where $g(\mathbf{x})$ stands for the geometric mean of the parts

Aitchison (1982, 1986)

Euclidean space structure of $\mathcal{S}^D$

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

- perturbation: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, \ldots, x_D y_D]$
- powering: $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$
- 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:

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

Aitchison (1982, 1986), operations and distance

Billheimer et al. (2001); Pawlowsky-Glahn and Egozcue (2001), Aitchison et al. (2002)

Euclidean space structure of $\mathcal{S}^D$

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

- perturbation: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, \ldots, x_D y_D]$
- powering: $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$
- 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:

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

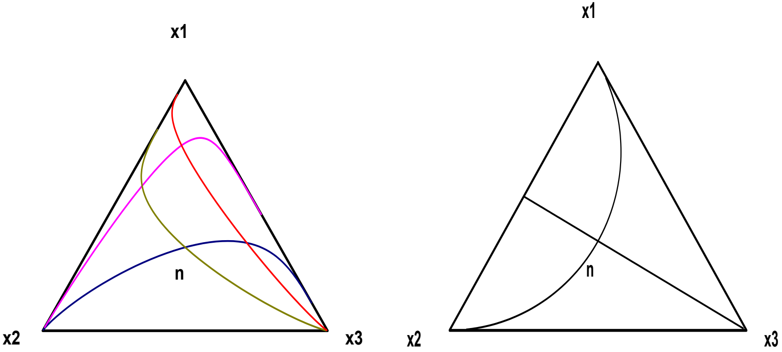
Aitchison (1982, 1986), operations and distance

Billheimer et al. (2001); Pawlowsky-Glahn and Egozcue (2001), Aitchison et al. (2002)

compositional lines

Correspond to exponential growth or decay of masses

$$\mathbf{y} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x}_1)$$



parallel lines orthogonal lines

illustration from Egozcue and Pawlowsky-Glahn (2006)

consequences

- in an Euclidean space an orthonormal basis always exists
- operations and metrics in the simplex are equivalent to ordinary operations and metrics in coordinates

$\mathbf{x} \in \mathcal{S}^D$, Coordinates: $\mathbf{y} = h(\mathbf{x}) \in \mathbb{R}^{D-1}$

- $\mathbf{x}_1 \oplus \mathbf{x}_2 \quad \Leftrightarrow \quad \mathbf{y}_1 + \mathbf{y}_2$
- $\alpha \odot \mathbf{x} \quad \Leftrightarrow \quad \alpha \cdot \mathbf{y}$
- $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathbf{y}_1, \mathbf{y}_2 \rangle$
- $d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_1, \mathbf{x}_2)$

consequences

- in an Euclidean space an orthonormal basis always exists
- operations and metrics in the simplex are equivalent to ordinary operations and metrics in coordinates

$\mathbf{x} \in \mathcal{S}^D$, Coordinates: $\mathbf{y} = h(\mathbf{x}) \in \mathbb{R}^{D-1}$

- $\mathbf{x}_1 \oplus \mathbf{x}_2 \quad \Leftrightarrow \quad \mathbf{y}_1 + \mathbf{y}_2$
- $\alpha \odot \mathbf{x} \quad \Leftrightarrow \quad \alpha \cdot \mathbf{y}$
- $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathbf{y}_1, \mathbf{y}_2 \rangle$
- $d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_1, \mathbf{x}_2)$

example of orthogonal coordinates

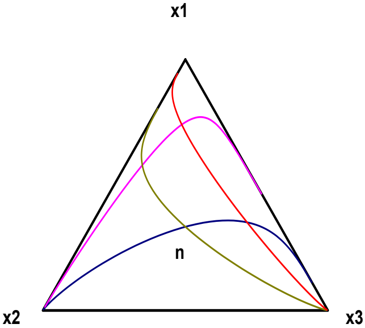
- example of orthonormal basis in $\mathcal{S}^3$:

$$\mathbf{e}_1 = \mathcal{C}\left[\exp\frac{1}{\sqrt{2}}, \exp\frac{-1}{\sqrt{2}}, 1\right], \quad \mathbf{e}_2 = \mathcal{C}\left[\exp\frac{1}{\sqrt{6}}, \exp\frac{1}{\sqrt{6}}, \exp\frac{-2}{\sqrt{6}}\right]$$

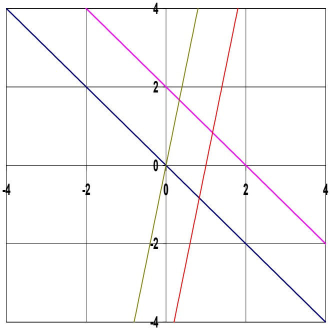
- coordinates for $\mathbf{x} = [x_1, x_2, x_3] \in \mathcal{S}^3$ in this basis:

$$y_1 = \frac{1}{\sqrt{2}} \ln\frac{x_1}{x_2}, \quad y_2 = \frac{1}{\sqrt{6}} \ln\frac{x_1 \cdot x_2}{x_3 \cdot x_3}$$

parallel lines

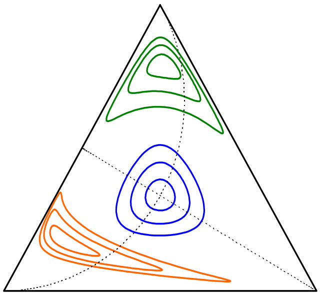


in $\mathcal{S}^3$

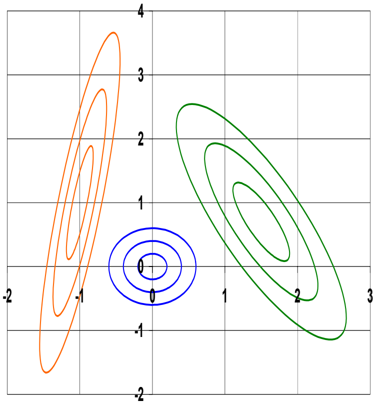


coordinate representation

circles and ellipses



in $\mathcal{S}^3$



coordinate representation

building an orthonormal basis

the intuitive approach

example: for $\mathbf{x} \in \mathcal{S}^5$ define a sequential binary partition and obtain the coordinates in the corresponding orthonormal basis

order	x_1	x_2	x_3	x_4	x_5	coordinate
1	$+1$	-1	$+1$	$+1$	-1	$y_1 = \sqrt{\frac{3\cdot 2}{3+2}} \ln \frac{(x_1 \cdot x_3 \cdot x_4)^{1/3}}{(x_2 \cdot x_5)^{1/2}}$
2	0	$+1$	0	0	-1	$y_2 = \sqrt{\frac{1\cdot 1}{1+1}} \ln \frac{x_2}{x_5}$
3	$+1$	0	-1	-1	0	$y_3 = \sqrt{\frac{1\cdot 2}{1+2}} \ln \frac{x_1}{(x_3 \cdot x_4)^{1/2}}$
4	0	0	$+1$	-1	0	$y_4 = \sqrt{\frac{1\cdot 1}{1+1}} \ln \frac{x_3}{x_4}$

balances

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

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

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 et al. (2003)

Egozcue, Pawlowsky-Glahn (2005,2006)

centre and total variance

Metric variability with respect to a point $\mathbf{z}$
$\mathbf{X}$ random composition with values in $\mathcal{S}^D$

$$\mathrm{Var}[\mathbf{X};\mathbf{z}] = \mathrm{E}[d_a^2(\mathbf{X},\mathbf{z})] = \mathrm{E}[d^2(h(\mathbf{X}),h(\mathbf{z}))]$$

Center or mean in the simplex: value of $\mathbf{z}$ minimizing $\mathrm{Var}[\mathbf{X};\mathbf{z}]$

$$\mathrm{Cen}[\mathbf{X}] = h^{-1}(\mathrm{E}[h(\mathbf{X})]) = \mathcal{C}\exp(\mathrm{E}[\ln \mathbf{X}])$$

Total or metric variance: minimum variability

$$\mathrm{Var}[\mathbf{X}] = \mathrm{Var}[\mathbf{X};\mathrm{Cen}[\mathbf{X}]] = \mathrm{E}[d^2(h(\mathbf{X}),\mathrm{E}[h(\mathbf{X})])]$$

Pawlowsky-Glahn, Egozcue (2001)

centre and total variance

Metric variability with respect to a point $\mathbf{z}$

$\mathbf{X}$ random composition with values in $\mathcal{S}^D$

$$\mathrm{Var}[\mathbf{X};\mathbf{z}] = \mathrm{E}[d_a^2(\mathbf{X},\mathbf{z})] = \mathrm{E}[d^2(h(\mathbf{X}),h(\mathbf{z}))]$$

Center or mean in the simplex: value of $\mathbf{z}$ minimizing $\mathrm{Var}[\mathbf{X};\mathbf{z}]$

$$\mathrm{Cen}[\mathbf{X}] = h^{-1}(\mathrm{E}[h(\mathbf{X})]) = \mathcal{C}\exp(\mathrm{E}[\ln\mathbf{X}])$$

Total or metric variance: minimum variability

$$\mathrm{Var}[\mathbf{X}] = \mathrm{Var}[\mathbf{X};\mathrm{Cen}[\mathbf{X}]] = \mathrm{E}[d^2(h(\mathbf{X}),\mathrm{E}[h(\mathbf{X})])]$$

Pawlowsky-Glahn, Egozcue (2001)

centre and total variance

Metric variability with respect to a point $\mathbf{z}$

$\mathbf{X}$ random composition with values in $\mathcal{S}^D$

$$\mathrm{Var}[\mathbf{X};\mathbf{z}] = \mathrm{E}[d_a^2(\mathbf{X},\mathbf{z})] = \mathrm{E}[d^2(h(\mathbf{X}),h(\mathbf{z}))]$$

Center or mean in the simplex: value of $\mathbf{z}$ minimizing $\mathrm{Var}[\mathbf{X};\mathbf{z}]$

$$\mathrm{Cen}[\mathbf{X}] = h^{-1}(\mathrm{E}[h(\mathbf{X})]) = \mathcal{C}\exp(\mathrm{E}[\ln\mathbf{X}])$$

Total or metric variance: minimum variability

$$\mathrm{Var}[\mathbf{X}] = \mathrm{Var}[\mathbf{X};\mathrm{Cen}[\mathbf{X}]] = \mathrm{E}[d^2(h(\mathbf{X}),\mathrm{E}[h(\mathbf{X})])]$$

Pawlowsky-Glahn, Egozcue (2001)

Estimators of centre and variance

Compositional sample: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$
sample centre: closed geometric mean

$$\widehat{\mathrm{Cen}}[\mathbf{X}] = \mathcal{C} \exp\left(\frac{1}{n}\sum_i \ln \mathbf{x}_i\right)$$

Sample total variance: trace of sample covariance matrix of coordinates.

Bias and mean-squared-error, when $\theta \in \mathcal{S}^D$

$\mathbf{Bias}(\widehat{\theta}) = \mathrm{Cen}[\widehat{\theta} \ominus \theta] = h^{-1}(\mathrm{E}[h(\widehat{\theta}) - h(\theta)])$

$\mathbf{MSE}(\widehat{\theta}) = \mathrm{E}[d_a^2(\widehat{\theta}, \theta)] = \mathrm{E}[\|h(\widehat{\theta}) - h(\theta)\|^2]$

Pawlowsky-Glahn, Egozcue (2002)

Estimators of centre and variance

Compositional sample: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$
sample centre: closed geometric mean

$$\widehat{\mathrm{Cen}}[\mathbf{X}] = \mathcal{C} \exp\left(\frac{1}{n}\sum_i \ln \mathbf{x}_i\right)$$

Sample total variance: trace of sample covariance matrix of coordinates.

Bias and mean-squared-error, when $\theta \in \mathcal{S}^D$

$$\mathbf{Bias}(\widehat{\theta}) = \mathrm{Cen}[\widehat{\theta} \ominus \theta] = h^{-1}(\mathrm{E}[h(\widehat{\theta}) - h(\theta)])$$

$$\mathbf{MSE}(\widehat{\theta}) = \mathrm{E}[d_a^2(\widehat{\theta}, \theta)] = \mathrm{E}[\|h(\widehat{\theta}) - h(\theta)\|^2]$$

Pawlowsky-Glahn, Egozcue (2002)

normal in the simplex

Simple idea: Model the distribution of coordinates!

$$h(\mathbf{X}) \sim N(\mu, \Sigma) \quad \Leftrightarrow \quad \mathbf{X} \sim N_{\mathcal{S}}(h^{-1}(\mu), \Sigma)$$

Central limit theorem: Independent $\mathbf{X}_i$ with $\mathrm{Cen}[\mathbf{X}_i] = h^{-1}(\boldsymbol{\mu})$, $\mathrm{Cov}[h(\mathbf{X}_i)] = \Sigma$,

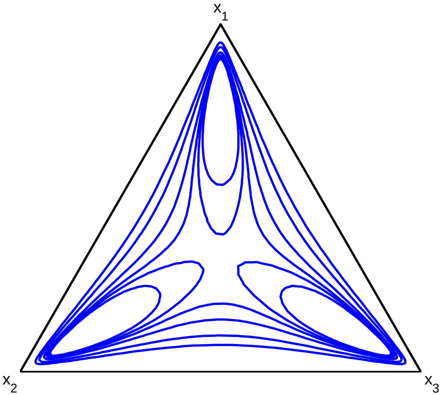
$$\frac{1}{n} \odot \bigoplus_{i=1}^{n} \mathbf{X}_i \approx N_{\mathcal{S}}(h^{-1}(\mu), n^{-1}\Sigma)$$

for large n.

Aitchison (1982,1986), Mateu-Figueras (2003)

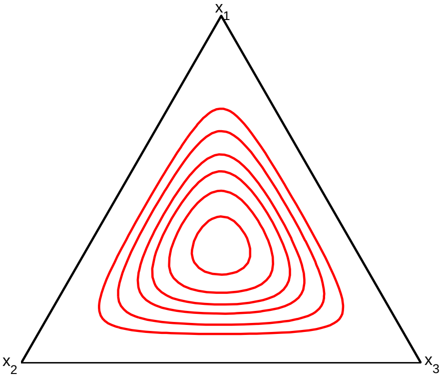
normal on the simplex (logistic-normal)

$\mathcal{S}^D \subset \mathbb{R}^D$, Lebesgue measure as reference



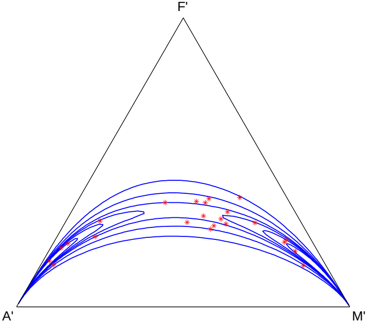
normal on the simplex (logistic-normal)

$\mathcal{S}^D$ as Euclidean space, Aitchison measure as reference



example: aphyric Skye lavas (modified)

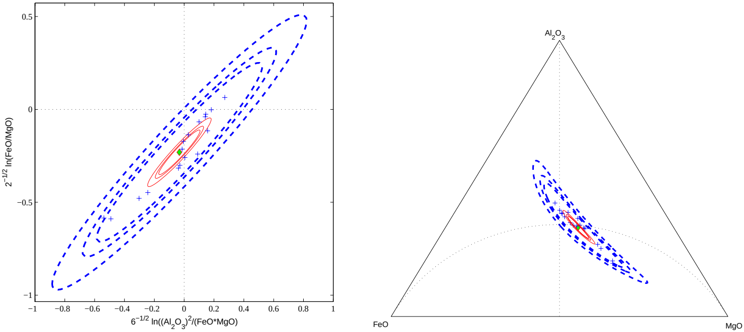
$\mathcal{S}^D \subset \mathbb{R}^D$ $\quad$ $\mathcal{S}^D$ as Euclidean space



logistic normal
Lebesgue measure λ

normal on $\mathcal{S}^D$
Aitchison measure λ_a

predictive regions for data and confidence regions for the mean (limits 90%, 95%, 99%)



data from Kilauea Iki lava lake, Hawaii, cited in Rollinson (1995)

regression model

Data: for $i = 1, 2, \ldots, n$
compositional response, $\mathbf{x}_i \in \mathcal{S}^D$,
real covariates, $\mathbf{t}_i = [t_0, t_1, t_2, \ldots, t_r]$, $t_0 = 1$

Statement: find compositional coefficients $\boldsymbol{\beta}_j \in \mathcal{S}^D$, minimizing

$$\mathrm{SSE} = \sum_{i=1}^{n} \|\hat{\mathbf{x}}(\mathbf{t}_i) \ominus \mathbf{x}_i\|_a^2 \ ,$$

$$\hat{\mathbf{x}}(\mathbf{t}) = \boldsymbol{\beta}_0 \oplus (t_1 \odot \boldsymbol{\beta}_1) \oplus \cdots \oplus (t_r \odot \boldsymbol{\beta}_r) = \bigoplus_{j=0}^{r} (t_j \odot \boldsymbol{\beta}_j) \ ,$$

regression model in coordinates

- Select a basis in $\mathcal{S}^D$, e.g. using sbp;
- Represent responses in coordinates: $\mathbf{x}_i^* = h(\mathbf{x_i}) \in \mathbb{R}^{D-1}$;
- Solve $D-1$ ordinary regression problems in coordinates to obtain coordinates of coefficients;
- Back-transform results into $\mathcal{S}^D$

For $k = 1, 2, \ldots, D$, find $\boldsymbol{\beta}^*$ minimizing

$$\mathrm{SSE}_k = \sum_{i=1}^{n} |\hat{x}_k^*(\mathbf{t}_i) - x_{ik}^*|^2 \ , \ k = 1, 2, \ldots, D-1 \ ,$$

$$\hat{x}_k^*(\mathbf{t}) = \beta_{0k}^* + \beta_{1k}^* \, t_1 + \cdots + \beta_{rk}^* \, t_r$$

Back-transform: $\boldsymbol{\beta}_j = h^{-1}(\boldsymbol{\beta}_j^*)$

example: statement

Vulnerability of a dike:

- Safety level or design d (wave-height-design)
- External actions h (wave-height of a storm)
- Outputs after an action θ_k, $k = 0, 1, \ldots, 4$
- Vulnerability description: $\mathbf{x}(d, h) = \mathrm{P}[\theta_k | d, h]$

Available data (from Monte Carlo simulations):

$$\mathbf{x}(d_i, h_i) = \mathrm{P}[\theta_k | d_i, h_i] \ , \ i = 1, 2, \ldots, n$$

affected by errors, especially, for low probabilities.

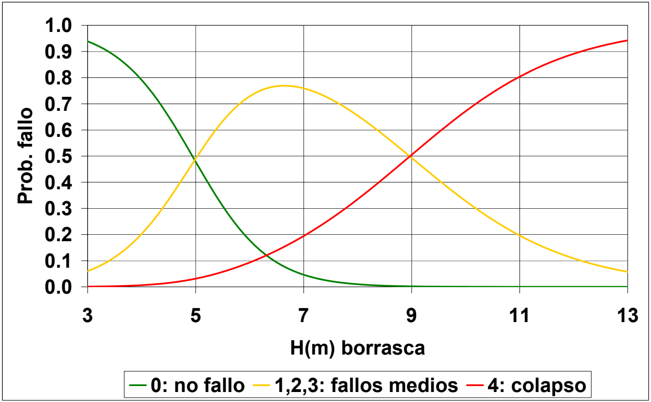
example: data set

Number of data: $n = 11$
Number of parts: $D = 4$
Number of covariates: $r = 2$

d	h	theta0	theta1	theta2	theta3
!diseño	Hs(m)	servicio	daño moderado	daño consid	colapso
3.0	3.0	8.9206E-01	8.9206E-02	1.7841E-02	8.9206E-04
3.0	18.0	6.6529E-05	1.9959E-03	3.3265E-01	6.6529E-01
15.0	3.0	9.9889E-01	9.9889E-04	9.9889E-05	9.9889E-06
15.0	18.0	7.4074E-02	3.7037E-01	5.1852E-01	3.7037E-02
5.0	5.0	8.4602E-01	1.2690E-01	2.5381E-02	1.6920E-03
6.0	10.0	8.2645E-03	1.2397E-01	8.2645E-01	4.1322E-02
9.0	4.0	9.7551E-01	1.9510E-02	4.8776E-03	9.7551E-05
10.0	7.0	9.1058E-01	8.1952E-02	7.2846E-03	1.8212E-04
11.0	18.0	1.0988E-04	1.0988E-02	4.3951E-01	5.4939E-01
12.0	7.0	9.7838E-01	1.9568E-02	1.9568E-03	9.7838E-05
7.0	14.0	4.8757E-04	2.4378E-02	4.8757E-01	4.8757E-01

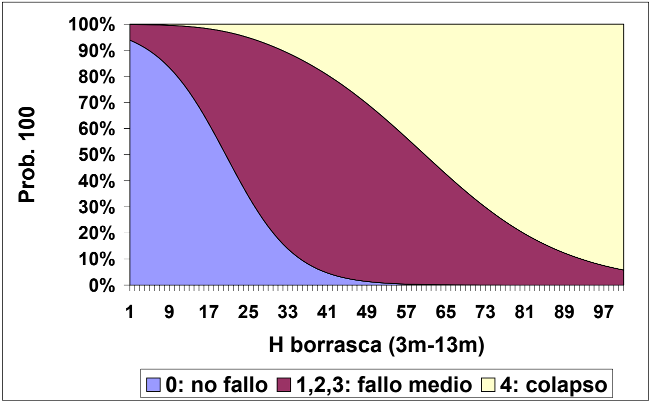
motivation for a simplicial linear model

M. Jiménez study of the Bastarreche-dike (Cartagena-Spain):



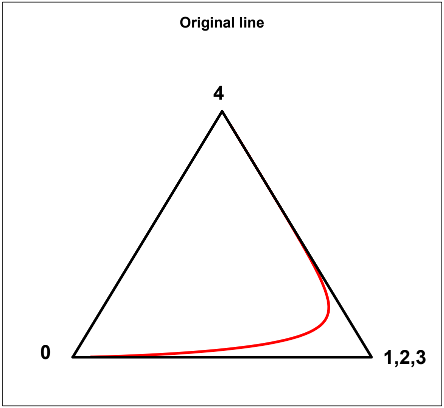
motivation for a simplicial linear model

M. Jiménez study of the Bastarreche-dike (Cartagena-Spain):



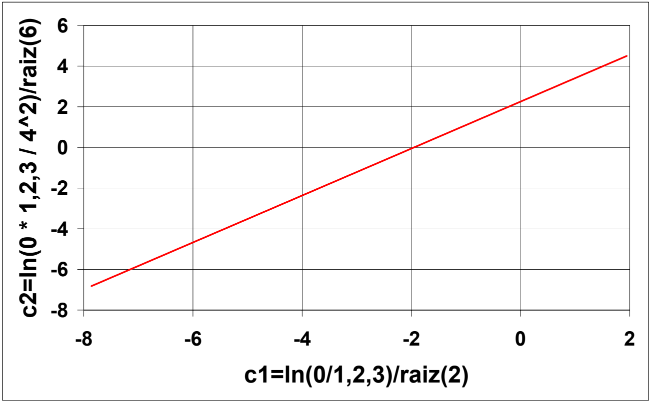
motivation for a simplicial linear model

M. Jiménez study of the Bastarreche-dike (Cartagena-Spain):

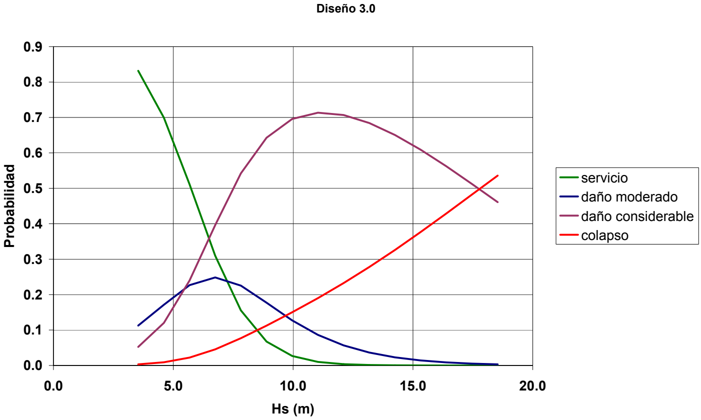


motivation for a simplicial linear model

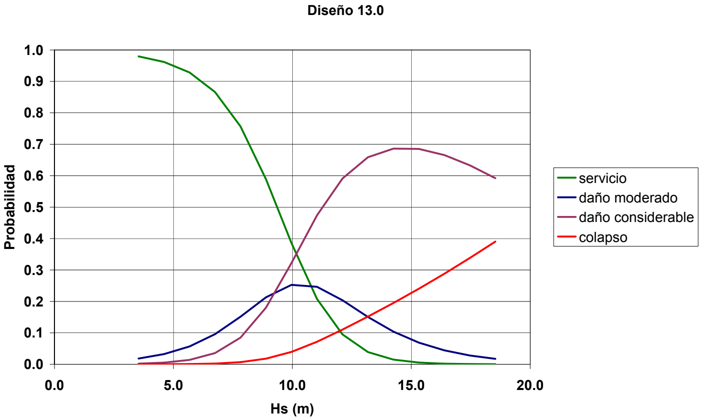
M. Jiménez study of the Bastarreche-dike (Cartagena-Spain):



example: linear model of vulnerability of a dike



example: linear model of vulnerability of a dike



conclusion

- Compositional data (CoDa) should be treated in the simplex with its specific geometry
- Ordinary multivariate statistics should not be directly applied to CoDa
- The simplex has its own Euclidean structure: cartesian coordinates are available
- Multivariate statistical models and methods work properly on coordinates of CoDa
- Problem (or advantage): interpretation of coordinates

CoDa analysis is easy!

Just transform CoDa into coordinates;
analyze whatsoever;
back-transform and interpret the results!

conclusion

- Compositional data (CoDa) should be treated in the simplex with its specific geometry
- Ordinary multivariate statistics should not be directly applied to CoDa
- The simplex has its own Euclidean structure: cartesian coordinates are available
- Multivariate statistical models and methods work properly on coordinates of CoDa
- Problem (or advantage): interpretation of coordinates

CoDa analysis is easy!

Just transform CoDa into coordinates;
analyze whatsoever;
back-transform and interpret the results!

further reading and activities

- Mathematical Geology Vol. 37 Nr. 7 (2005) – special issue on compositional data analysis
- Compositional data analysis in the Geosciences: From theory to practice (October 2006) — special publication of the Geological Society (SPE 264)
- CoDaWork'08, Girona (Spain), May 2008 (http://ima.udg.es/Activitats/CoDaWork08/)