The statistical analysis of compositional data: Processes and regression

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

Instituto Colombiano del Petróleo
Piedecuesta, Santander, Colombia

March 20–23, 2007

Summary

Normal distribution in $\mathcal{S}^D$

Coordinates: $\mathbf{X}^*$ random variable in $\mathbb{R}^{D-1}$

$$\mathbf{X}^* \sim \mathrm{N}(\mu^*, \Sigma)$$

$$f_{\mathbf{X}^*}(\mathbf{x}^*) = \frac{1}{\sqrt{2\pi(\det\Sigma)^n}} \exp\left(-\frac{1}{2}(\mathbf{x}^* - \boldsymbol{\mu}^*)'\Sigma^{-1}(\mathbf{x}^* - \boldsymbol{\mu}^*)\right)$$

Simplex: given a basis in $\mathcal{S}^D$ and $\mathbf{X} = \mathrm{ilr}^{-1}(\mathbf{X}^*)$, then

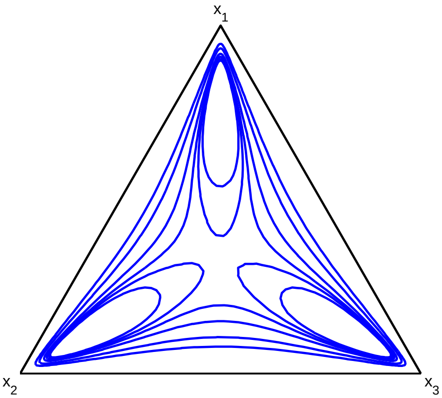
$$\mathbf{X} \sim \mathrm{N}_{\mathcal{S}^D}(\mu, \Sigma)$$

$$\mu = \mathrm{ilr}^{-1}(\mu^*)$$

The variance is represented in both cases by Σ

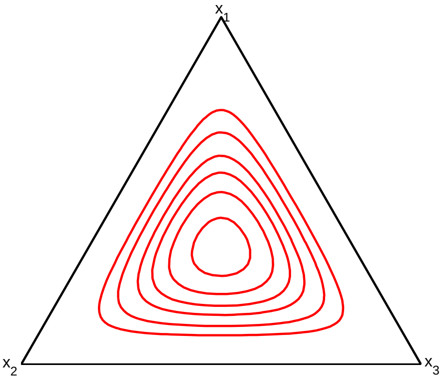
Normal on the simplex (logistic-normal)

$\mathcal{S}^3 \subset \mathbb{R}^2$, Lebesgue measure as reference:
Radon-Nikodym derivative: $f = \frac{dP}{d\lambda}$



Normal on the simplex (logistic-normal)

$\mathcal{S}^3$ as Euclidean space, Aitchison measure as reference:
Radon-Nikodym derivative: $f = \frac{dP}{d\lambda_{\mathcal{S}}} = \frac{dP}{d\lambda} \cdot \frac{d\lambda}{d\lambda_{\mathcal{S}}}$



Representation in the ternary diagram

- Compute the elliptical contours of the Normal distribution in coordinates, by points
- Select a basis, and take ilr^{-1} of each point in the contours
- plot the ilr^{-1} transformed points in the ternary diagram

This procedure is equivalent to represent the density with respect to the Aitchison measure

To obtain contours with respect Lebesgue measure in the ternary diagram, the Jacobian of the ilr transformation is taken into account

Exponential decay or growth

Bacteria growth: mass of 3 species: $x_1(t), x_2(t), x_3(t)$
Growth without interaction:

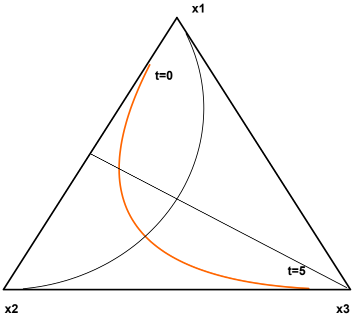
$$x_i(t) = x_i(0) \cdot \exp(\lambda_i t) \ , \ (\lambda_i > 0) \ , \ i = 1, 2, 3$$

Considered as compositional: straight-line in $\mathcal{S}^3$

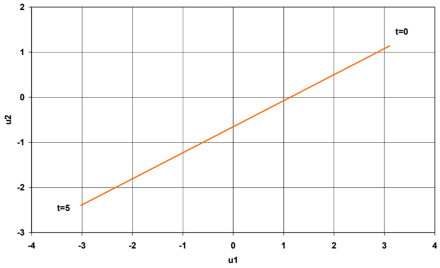
$$\mathbf{x}(t) = \mathbf{x}(0) \oplus (t \odot \exp(\boldsymbol{\lambda}))$$

Bacteria growth

$$\mathbf{x}(0) = [10.0, 2.0, 0.1], \ \boldsymbol{\lambda} = [1, 2, 3] \ , t = 0, \ldots, 5$$



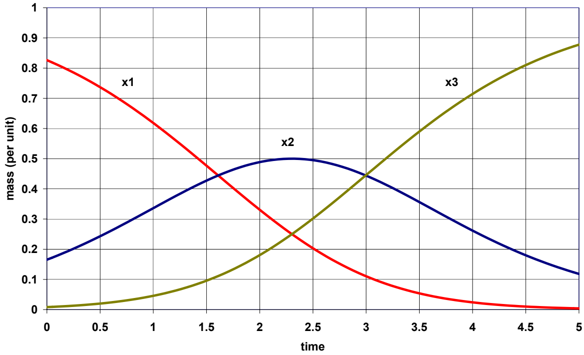
in $\mathcal{S}^3$



coordinate representation

Bacteria growth

$$\mathbf{x}(0) = [10.0, 2.0, 0.1], \ \boldsymbol{\lambda} = [1, 2, 3] \ , t = 0, \ldots, 5$$



Complementary process

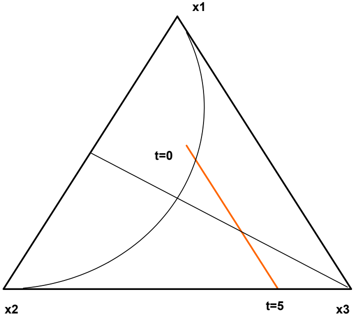
Three isotopes:
$x_1(t)$ radioactive; decays with rate λ_1
$x_2(t)$ inert; does neither grow nor decay $\lambda_2 = 0$
$x_3(t)$ residual of decomposition of x_1

$$x_1(t) = x_1(0) \cdot \exp(\lambda_1 t)\ ,\ x_2(t) = x_2(0)\ ,\ x_3(t) = x_3(0) + x_1(0) - x_1(t)$$

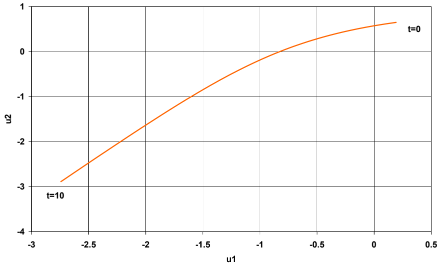
parameter	x_1	x_2	x_3
disintegration rate	0.5	0.0	0.0
initial mass	1.0	0.4	0.5
balance 1	$+1$	$+1$	-1
balance 2	$+1$	-1	0

Radioactive disintegration

The complementary part $x_3(t)$ makes the process non-linear



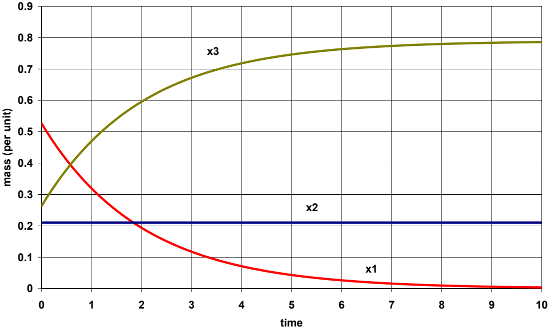
in $\mathcal{S}^3$



coordinate representation

Radioactive disintegration

$$\mathbf{x}(0) = [10.0, 2.0, 0.1], \ \boldsymbol{\lambda} = [1, 2, 3] \ , t = 0, \ldots, 5$$



Perturbation versus mixture

Consider a initial, final composition of a liquid, $\mathbf{z}_0$ and $\mathbf{z}_1$

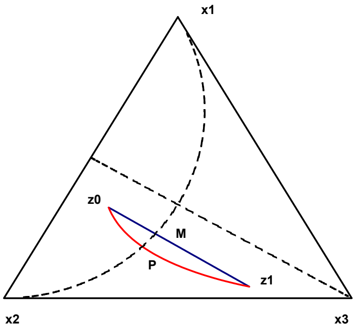
Mixing: change volume composition $[\alpha, 1-\alpha]$, $0 \leq \alpha \leq 1$

$$\mathbf{z}(\alpha) = ((1-\alpha)\cdot\mathbf{z}_0) + (\alpha\cdot\mathbf{z}_1)$$

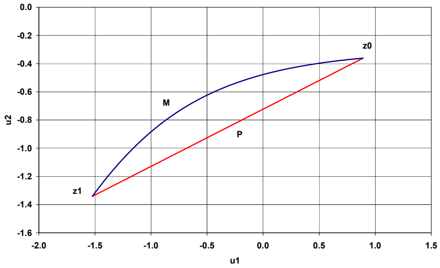
Perturbing:

$$\mathbf{z}(\tau) = \mathbf{z}_0 \oplus \{\tau \odot (\mathbf{z}_1 \ominus \mathbf{z}_0)\} = ((1-\tau)\odot\mathbf{z}_0) \oplus (\tau \odot \mathbf{z}_1)$$

Perturbation versus mixture, example



in $\mathcal{S}^3$



coordinate representation

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):

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

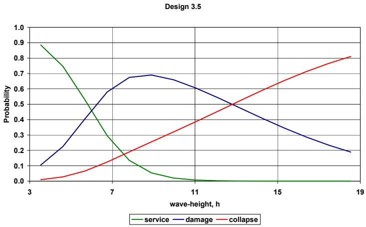
affected by errors, especially, for low probabilities.

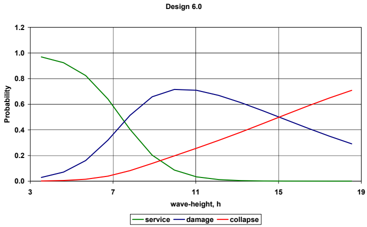
example: data set

Number of data: $n = 11$
Number of parts: $D = 3$
Number of covariates: $r = 2$

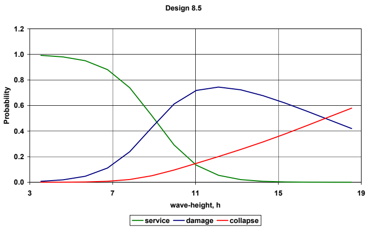
Design	Wave-height	p. service	p. damage	p.collapse
3.0	3.0	0.50	0.49	0.01
3.0	10.0	0.02	0.10	0.88
10.0	3.0	0.999	0.0009	0.0001
10.0	10.0	0.30	0.65	0.05
5.0	4.0	0.95	0.049	0.001
6.0	9.0	0.08	0.85	0.07
7.0	5.0	0.97	0.027	0.003
8.0	3.0	0.997	0.0028	0.0002
9.0	9.0	0.35	0.55	0.01

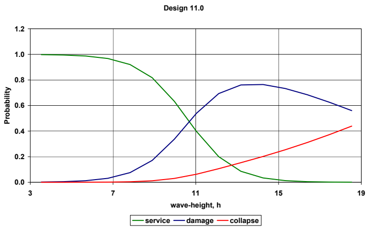
Example: linear model of vulnerability of a dike



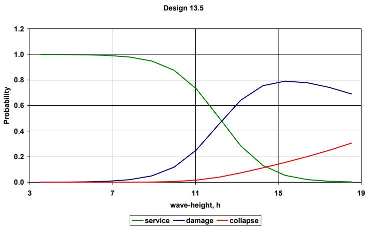


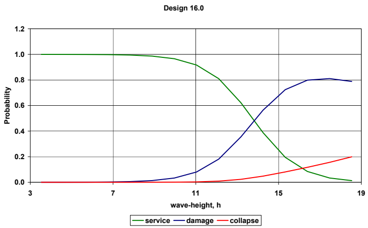
Example: linear model of vulnerability of a dike



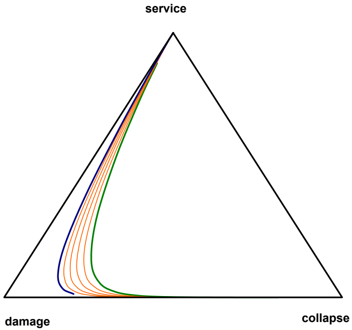


Example: linear model of vulnerability of a dike

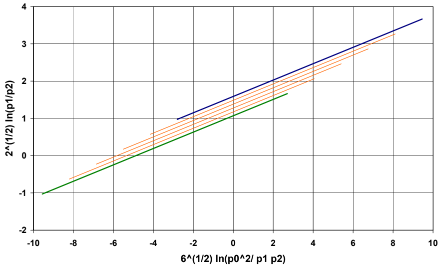




Example: linear model of vulnerability of a dike



Ternary diagram



Coordinates

Design 3.5: green ; Design 16.0: blue

Example: analysis of residuals

ANOVA:

$$c_1 = \frac{1}{6} \ln \frac{p_0^2}{p_1 p_2} \quad , \quad \mathrm{p-value} = 2.69E - 05$$

$$c_2 = \frac{1}{2} \ln \frac{p_1}{p_2} \quad , \quad \mathrm{p-value} = 3.15E - 01$$