A Forward Search Algorithm for Compositional Data

Filippo Palombi

filippo.palombi@istat.it

Istat

in collaboration with

Simona Toti (ISTAT), Romina Filippini (ISTAT), Valeria Tomeo (ISTAT)

Conference of European Statisticians 2011 - UNECE

Work Session on Statistical Data Editing

Ljubljana - May 11^{th}, 2011

what are compositional data ?

- multivariates with components interpreted as parts of a whole
  - Example 1: Financial Portfolio (Unit: $)

Portfolio	Stocks	Bonds	Options	Cache	Total
P_1	2'000	8'000	100	30'000	40'100
P_2	3'000	4'000	500	20'000	27'500
⋮	⋮	⋮	⋮	⋮	⋮

  - Example 2: Agricultural surface (Unit: are)

Farm	Soft Wheat	Barley	Corn	Other Cereals	Total
F_1	889	231	281	72	1473
F_2	1199	480	1191	85	2955
⋮	⋮	⋮	⋮	⋮	⋮

- we look at the relative importance of different components . . .

what are compositional data ?

- multivariates with components interpreted as parts of a whole
  - Example 1: Financial Portfolio (Unit: $)

Portfolio	Stocks	Bonds	Options	Cache	Total
P_1	0.0499	0.1995	0.0025	0.7481	1.0
P_2	0.1091	0.1455	0.0182	0.7273	1.0
⋮	⋮	⋮	⋮	⋮	⋮

  - Example 2: Agricultural surface (Unit: are)

Farm	Soft Wheat	Barley	Corn	Other Cereals	Total
F_1	0.6035	0.1568	0.1908	0.0489	1.0
F_2	0.4058	0.1624	0.4030	0.0288	1.0
⋮	⋮	⋮	⋮	⋮	⋮

- we look at the relative importance of different components . . .
  . . . by normalizing each variable to its total

what are compositional data ?

- multivariates with components interpreted as parts of a whole
  - Example 1: Financial Portfolio (Unit: $)

Portfolio	Stocks	Bonds	Options	Cache	Total
P_1	0.0499	0.1995	0.0025	0.7481	1.0
P_2	0.1091	0.1455	0.0182	0.7273	1.0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

  - Example 2: Agricultural surface (Unit: are)

Farm	Soft Wheat	Barley	Corn	Other Cereals	Total
F_1	0.6035	0.1568	0.1908	0.0489	1.0
F_2	0.4058	0.1624	0.4030	0.0288	1.0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- we look at the relative importance of different components . . .

  . . . by normalizing each variable to its total
- how do we search for **compositional outliers** ?

Relevant papers for Compositional Analysis

- seminal work

  "The Statistical Analysis of Compositional Data"
  J. Aitchison
  Monograph on Statistics and Applied Probability, Chapman & Hall Ltd. (1986)

- distributional hypothesis

  "Logistic-Normal Distributions: Some Properties and Uses"
  J. Aitchison, S. M. Shen
  Biometrika, Vol. 67, No. 2 (1980), pp. 261–272

- isometric logratio transformation

  "Isometric Logratio Transformations for Compositional Data Analysis"
  J. J. Egozcue, V. Pawlowsky–Glahn, G. Mateu-Figueras and C. Barcelò–Vidal
  Mathematical Geology, Vol. 35, No. 3 (2003), pp. 279–300

Relevant papers for the Forward Search Algorithm (FS)

- original proposal

  "Fast very robust methods for the detection of multiple outliers"
  A. C. Atkinson
  Journal of the American Statistical Association, 89 (1994), pp. 1329–1339

- multivariate version

  "Finding an unknown number of of multivariate outliers"
  M. Riani, A. C. Atkinson, A. Cerioli
  J. R. Statist. Soc. B (2009) **71**, *Part 2, pp. 447–466*

- mathematical foundations

  "Discussion of The FS: Theory and Data Analysis by Atkinson, Riani, and Cerioli"
  S. Johansen and B. Nielsen
  Center for Research in Econometric Analysis of Time Series, Research Paper 2010-6

Standard Forward Search Algorithm

(Riani *et al.* (2009))

Null hypothesis

$$\mathcal{D} = \{y^{(k)} \in \mathbb{R}^v\}_{k=1,\ldots,n}$$

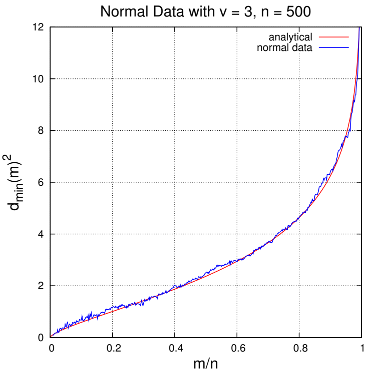
$$H_0 : \quad \{y^{(1)} \sim \mathcal{N}(\mu,\Sigma)\} \cap \{y^{(2)} \sim \mathcal{N}(\mu,\Sigma)\} \cap \cdots \cap \{y^{(n)} \sim \mathcal{N}(\mu,\Sigma)\}$$

Construction of the signal $d_{\min}(m)$

- initialization
  0. choose a subset $S(m_0) \subset \mathcal{D}$ of m_0 elements of $\mathcal{D}$
- m^{th} – step $(m_0 \leq m \leq n-1)$:
  1. compute mean $\mu(m)$ and covariance matrix $\Sigma(m)$ of $S(m)$
  2. compute the Mahalanobis distance d of all $y \in S(m)$ from $\mu(m)$
  3. define $d_{\min}(m) = d_{[m+1]}$ {the $(m+1)^{\text{th}}$-ordered distance}
  4. define $S(m+1)$ the set of the first $m+1$ y's closest to $\mu(m)$

- forward plot: $d^2_{\min}(m)$ vs. m
- under H_0, $d^2_{\min}(m)$ fluctuates around

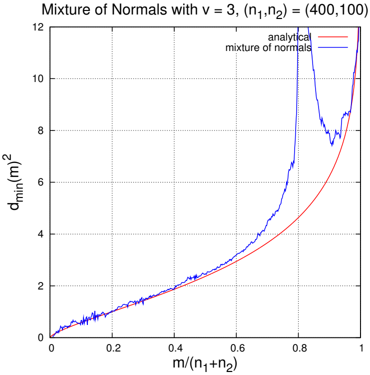
  $$d^2_{\min}(m) = (\chi^2_v)^{-1}\left(\frac{m}{n}\right) + \mathrm{O}\left(\frac{1}{n^k}\right)$$
  $$k \geq 1$$
- in presence of outliers, distortions in the forward plot are observed
- outliers $\equiv$ statistically relevant distortions
- $\Rightarrow$ need for quantitative assessment $\Leftarrow$



- forward plot: $d^2_{\min}(m)$ vs. m
- under H_0, $d^2_{\min}(m)$ fluctuates around

  $d^2_{\min}(m) = (\chi^2_v)^{-1}\left(\frac{m}{n}\right) + \mathrm{O}\left(\frac{1}{n^k}\right)$

  $k \geq 1$
- in presence of outliers, distortions in the forward plot are observed
- outliers $\equiv$ statistically relevant distortions
- $\Rightarrow$ need for quantitative assessment $\Leftarrow$



Construction of the envelopes

- $d^2_{\min}(m) = d^2_{[m+1]}$ is the $(m+1)^{\text{th}}$ order statistics
- "An easy method for obtaining percentage points of order statistics"
  W. C. Guenther, *Technometrics, Vol. 19, No. 3 (1977), pp. 319–321*

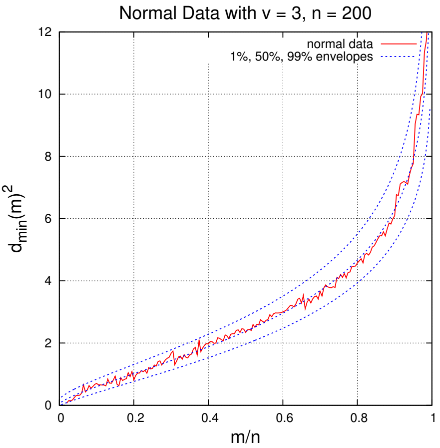
Theorem (Guenther)

Given $\{y^{(1)}, \ldots, y^{(n)}\}$ with $\{y^{(k)} \sim G\}_{k=1,\ldots,n}$, the α–percentile of $y_{[m+1]}$ is given by the equation

$$y_{[m+1];\alpha} = G^{-1}\left(\frac{m+1}{m+1+(n-m)f_{2(n-m),2(m+1);1-\alpha}}\right)$$

where $f_{a,b;\alpha}$ denotes the α–percentile of the Fisher distribution with parameters (a, b).

- we choose a global confidence level α and, for each m, we draw α–and $(1-\alpha)$–percentiles (percentile envelopes)


Normal Data with v = 3, n = 200
normal data
1%, 50%, 99% envelopes
$d_{min}(m)^2$
m/n
0
2
4
6
8
10
12
0
0.2
0.4
0.6
0.8
1

Compositional Analysis

(Aitchison (1986))

- compositional data live on the simplex

$$\mathcal{S}^{(v)} = \left\{x \in \mathbb{R}^v : \quad 0 < x_k < \kappa\,, \quad \sum_{k=1}^{v} x_k = \kappa\right\}$$

- Euclidean–type distances are not appropriate on $S^{(v)}$. Example:

$$d_{\mathrm{M}}(x,y) = \sqrt{\sum_{i,j=1}^{v}(x_i - y_i)\Sigma_{ij}^{-1}(x_j - y_j)}$$ **not defined**: $\det\Sigma = 0$

- Aitchison has proposed a better definition:

$$d_{\mathrm{A}}(x,y) = \sqrt{\frac{1}{2v}\sum_{i,j=1}^{2v}\left[\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{y_i}{y_j}\right)\right]^2}$$

- d_{A} emerges naturally from a vector space construction

Q: can we develop a FS for Compositional Data ?

The question breaks up into three sub–questions:

Q1. how do we construct the signal ?

Q2. how do we compute the percentile envelopes ?

Q3. how does H_0 change ?

A: YES!

A1. replace the Mahalanobis distance with the Aitchison distance

A2. questions Q2. & Q3. are related; the answer rests in the ILR

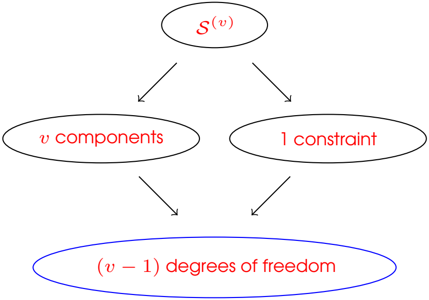
Isometric Logratio Transformation

(Egozcue et al. (2003))

$\mathcal{S}^{(v)}$ has a vector space structure

1. closure: $\mathcal{C}(x) = \left\{ \frac{\kappa x_1}{\sum_{k=1}^{v} x_k}, \ldots, \frac{\kappa x_v}{\sum_{k=1}^{v} x_k} \right\}$
2. vector sum: $x \oplus y = \mathcal{C}(x_1 y_1, \ldots, x_v y_v)\,, \quad \forall x, y \in \mathcal{S}^{(v)}$
3. product by a real: $\alpha \otimes y = \mathcal{C}(x_1^{\alpha}, \ldots, x_v^{\alpha})\,, \quad \forall \alpha \in \mathbb{R},\ x \in \mathcal{S}^{(v)}$
3. scalar product: $\langle x, y \rangle_{\mathrm{A}} = \frac{1}{2v} \sum_{i,j=1}^{v} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right)$
4. vector norm: $||x||_{\mathrm{A}} = \sqrt{\langle x, x \rangle_{\mathrm{A}}}$
5. distance: $d_{\mathrm{A}}(x, y) = ||x - y||_{\mathrm{A}}$

Q: how to find a basis of $\mathcal{S}^{(v)}$?



A: Gram–Schmidt orthonormalization makes the job!

Orthonormal vectors:

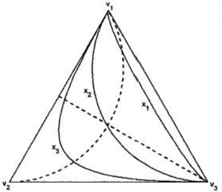
$$e_k = \mathcal{C}\left\{\underbrace{\exp\sqrt{\frac{1}{k(k+1)}},\ldots,\exp\sqrt{\frac{1}{k(k+1)}}}_{k\ \text{times}},\exp\left[-\sqrt{\frac{1}{k(k+1)}}\right],\underbrace{1\ \ldots,1}_{v-1-k\ \text{times}}\right\}$$

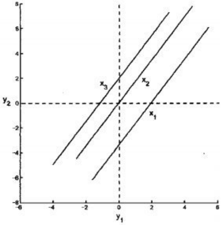
$k = 1,\ldots,v-1$

$$x \in \mathcal{S}^{(v)}$$

$$\langle x, e_1\rangle_{\mathrm{A}} \qquad \langle x, e_2\rangle_{\mathrm{A}} \qquad \ldots \qquad \langle x, e_{v-2}\rangle_{\mathrm{A}} \qquad \langle x, e_{v-1}\rangle_{\mathrm{A}}$$

$$\mathrm{ilr}(x) \stackrel{\mathrm{def}}{=} \{\langle x, e_1\rangle_{\mathrm{A}}, \ldots, \langle x, e_{v-1}\rangle_{\mathrm{A}}\}$$





$$\Rightarrow \quad \boxed{d_{\mathrm{A}}(x,y) = d_{\mathrm{E}}(\mathrm{ilr}(x), \mathrm{ilr}(y))\,, \quad \forall x, y \in \mathcal{S}^{(v)}} \quad \Leftarrow$$

Need for a distributional hypothesis to be tested by the FS

- how does $d_{\mathrm{A}}(x, y)$ distribute ? it depends on how x, y distribute!
- one can easily prove the following logical chain:

$$y \sim \ln \mathcal{N}_v \quad \Leftrightarrow \quad \mathcal{C}(y) \sim L_{v-1} \quad \Leftrightarrow \quad \mathrm{ilr}(\mathcal{C}(y)) \sim \mathcal{N}_{v-1}$$

- lognormal distribution is more natural for positive quantities
- distributional parameters can be easily related

Therefore, we turn H_0 into:

$$H_0: \quad \{y^{(1)} \sim \ln \mathcal{N}_v(\mu, \Sigma)\} \cap \{y^{(2)} \sim \ln \mathcal{N}_v(\mu, \Sigma)\} \cap \cdots \cap \{y^{(n)} \sim \ln \mathcal{N}_v(\mu, \Sigma)\}$$

and

$$d^2_{\mathrm{A}}(y^{(j)}, y^{(k)}) = d^2_{\mathrm{E}}(\mathrm{ilr}(\mathcal{C}(x^{(j)}), \mathrm{ilr}(\mathcal{C}(y^{(k)}))) \sim$$ **Euclidean square distance under normality hypothesis**

- distribution of quadratic forms has been studied a long time ago

  "Probability Content of Regions Under Spherical Normal Distributions, IV: The Distribution of Homogeneous and Non-Homogeneous Quadratic Functions of Normal Variables"

  H. Ruben, *Annals of Mathematical Statistics, Vol. 33, No. 2 (1962), pp. 542–570*

Theorem (Ruben)

The c.d.f. of a quadratic form t^2 of normal v–dimensional variables can be represented as a series of χ^2 distributions

$$H_{\mu,\Sigma}(t^2) = \sum_{j=0}^{\infty} \omega_j(\mu, \Sigma, p)\, \chi^2_{v+2j}(t^2/p)$$

- μ, Σ depend on t^2 and the distributional parameters of the variates
- coeffs $\{\omega_k\}_{k=v,v+2,v+4,\ldots}$ can be recursively computed (cfr. paper)
- $p > 0$ is a properly chosen scale factor

- the series converges rapidly: first few terms are sufficient

Compositional Forward Search

Null hypothesis

$$\mathcal{D} = \{y^{(k)} \in \mathbb{R}^v_+\}_{k=1,\dots,n}$$

$$H_0: \quad \{y^{(1)} \sim \ln\mathcal{N}(\mu,\Sigma)\} \cap \{y^{(2)} \sim \ln\mathcal{N}(\mu,\Sigma)\} \cap \cdots \cap \{y^{(n)} \sim \ln\mathcal{N}(\mu,\Sigma)\}$$

- initialization
  - **0.a** choose $\kappa = 1$ and close $\mathcal{D}$: $\mathcal{D} \rightarrow \mathcal{C}(\mathcal{D})$
  - **0.b** apply the isometric logratio transform: $\mathcal{C}(\mathcal{D}) \rightarrow \operatorname{ilr}[\mathcal{C}(\mathcal{D})]$
- construction of the signal
  - **1.** run the FS algorithm on $\operatorname{ilr}[\mathcal{C}(\mathcal{D})]$ with d^2_{M} replaced by d^2_{E}
- construction of the percentile envelopes
  - **2.** compute the envelopes of d^2_{E} from Ruben's distribution

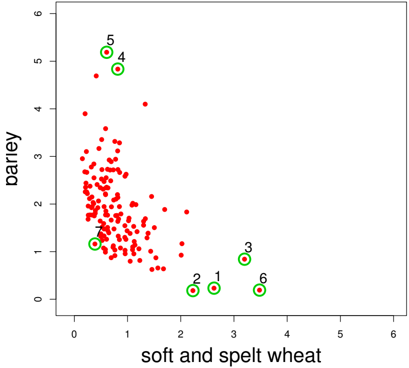
$\Rightarrow$ numerical technicalities: no time for a discussion $\Leftarrow$

An example from the Italian Agricultural Census 2010

$n = 148$

$v = 3$ (surfaces @: barley, soft & spelt wheat, corn)

Data refer to the Province of Alessandria (Piedmont)



Plot in log–log scale. Unit : are