Weighting of parts in compositional data analysis

K. Hron[1], J.J. Egozcue[2], V. Pawlowsky-Glahn[3],
J. Palarea-Albaladejo[4], P. Filzmoser[5]

[1]Palacký University, Olomouc, Czech Republic
[2]Polytechnic University of Catalonia, Barcelona, Spain
[3]University of Girona, Girona, Spain
[4]Biomathematics and Statistics Scotland, Edinburgh, United Kingdom
[5]Vienna University of Technology, Vienna, Austria

Robust, 22 January 2018

I'm going to present an ongoing research...

Compositional data

= *D-part positive vectors, describing quantitatively parts of a whole, carrying relative information* (Aitchison, 1986; Pawlowsky-Glahn a kol., 2015)

Compositional data

= *D-part positive vectors, describing quantitatively parts of a whole, carrying relative information* (Aitchison, 1986; Pawlowsky-Glahn a kol., 2015)

- **usual units of measurement**: percentages, mg/kg (*constant sum of parts*), mg/l (*constant sum does not occur*)

Compositional data

= *D-part positive vectors, describing quantitatively parts of a whole, carrying relative information* (Aitchison, 1986; Pawlowsky-Glahn a kol., 2015)

- **usual units of measurement**: percentages, mg/kg (*constant sum of parts*), mg/l (*constant sum does not occur*)
- **examples**: geochemical data - proportions of minerals in a rock; concentrations of fenolic acids in wine (mg/l); household expenditures (foodstuff, housing, clothing) etc.

Compositional data

= *D-part positive vectors, describing quantitatively parts of a whole, carrying relative information* (Aitchison, 1986; Pawlowsky-Glahn a kol., 2015)

- **usual units of measurement**: percentages, mg/kg (*constant sum of parts*), mg/l (*constant sum does not occur*)
- **examples**: geochemical data - proportions of minerals in a rock; concentrations of fenolic acids in wine (mg/l); household expenditures (foodstuff, housing, clothing) etc.
- constant sum of part (1, 100) = *proper representation of equivalence classes of proportional vectors. . .* $\mathcal{S}^D$

Geometrical aspects of compositional data analysis

- principles of compositional data analysis: *scale invariance* (SI), *subcompositional coherence*, *permutation invariance* ⇒ the **Aitchison geometry** (AG; EVS of dimension $D-1$)

- keystones of the Aitchison geometry: operations of perturbation and powering, the Aitchison inner product ($\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $c \in \mathbb{R}$)

$$\mathbf{x} \oplus \mathbf{y} = (x_1 y_1, \ldots, x_D y_D)', \quad c \odot \mathbf{x} = (x_1^c, \ldots, x_D^c)', \tag{1}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}; \tag{2}$$

- most of statistical methods rely on Euclidean geometry in real space (Eaton, 1983) ⇒ express data in orthonormal coordinates w.r.t. the Aitchison geometry (Pawlowsky-Glahn, Egozcue, 2001) as *logconstrasts*: $\sum_{i=1}^{D} a_i \ln x_i$, $\sum_{i=1}^{D} a_i = 0$

Step ahead: weighting of parts

- rarely in practice all variables have the same importance in multivariate analysis (lower precision of measurement devices for minor elements, noise variables in omics data, reliability of variables in questionnaire studies)

⇒ reliable weighting of variables needed

- **first step**: extension of the Aitchison geometry to Bayes spaces (van den Boogaart et al., 2014)

Step ahead: weighting of parts

- rarely in practice all variables have the same importance in multivariate analysis (lower precision of measurement devices for minor elements, noise variables in omics data, reliability of variables in questionnaire studies)

⇒ reliable weighting of variables needed

- **first step**: extension of the Aitchison geometry to Bayes spaces (van den Boogaart et al., 2014)

- **second step**: implications of the Bayes space methodology for weighting of compositional parts (Egozcue and Pawlowsky-Glahn, 2016), but still without considering important theoretical and practical aspects

Edinburgh 2017

We were for a research stay in Edinburgh...

Edinburgh 2017

Edinburgh - Promised City of compositional people...

Edinburgh 2017

. . . so inspiration comes immediately

Towards a general reference measure

- consider D categories $c_1, \ldots, c_D$ constituting a partition of the whole measurable space Ω

- a reference (finite) measure P on Ω assigns the "volume" p_i to the category c_i

⇒ the total volume of Ω is then $P(\Omega) = \sum_i p_i$

⇒ P provides information about how volumes are distributed on categories (and thus scale invariance of the composition $\mathbf{p} = (p_1, \ldots, p_D)$ applies), but also indicates the total volume $P(\Omega)$

- the standard reference measure on Ω assigns a unit to each category, that is $P_0(\Omega) = D$ and $p_{0i} = 1$ for $i = 1, 2, \ldots, D$, a uniform measure on the discrete space Ω

Towards a general reference measure

- given a composition $\mathbf{x} = (x_1, \ldots, x_D)$, there is a measure $\mu_{\mathbf{x}}$ which assigns the value x_i to each c_i; consider $A \subset \Omega$

$$\mu_{\mathbf{x}}(A) = \int_A \frac{d\mu_{\mathbf{x}}}{dP_0}\, dP_0 = \sum_{c_i \in A} x_i = \sum_{c_i \in A} \frac{x_i}{1}\, 1 = \sum_{c_i \in A} \frac{x_i}{p_i}\, p_i = \int_A \frac{d\mu_{\mathbf{x}}}{dP}\, dP$$

⇒ Radon-Nikodym derivatives are obtained

$$\mathbf{x} = \frac{d\mu_{\mathbf{x}}}{dP_0} = (x_1, x_2, \ldots, x_D),\ \mathbf{x}^{(\mathbf{p})} = \frac{d\mu_{\mathbf{x}}}{dP} = \left(\frac{x_1}{p_1}, \frac{x_2}{p_2}, \ldots, \frac{x_D}{p_D}\right)$$

⇒ $\mathbf{x}^{(\mathbf{p})} = (x_1^{(\mathbf{p})}, \ldots, x_D^{(\mathbf{p})}) = \mathbf{x} \ominus \mathbf{p}$ can be considered as a composition (SI) with respect to reference measure P

× weights $\mathbf{p}$ and their positive multiple $\alpha\mathbf{p} = (\alpha p_1, \ldots, \alpha p_D)$ do not represent the same information (shrinkage, expansion)

$$\sum_{c_i \in A} x_i^{(\mathbf{p})} p_i = \sum_{c_i \in A} x_i^{(\alpha\mathbf{p})} \alpha p_i = \sum_{c_i \in A} \frac{x_i}{\alpha p_i} \alpha p_i .$$

Geometry of a weighted space

- the vector space operations for compositions, perturbation and powering, change their expressions using densities when changing the reference measure

- perturbation and powering of measures (van den Boogaart et al., 2014)

$$(\mu_{\mathbf{x}} \oplus \mu_{\mathbf{y}})(A) = \int_A \frac{d\mu_{\mathbf{x}}}{dP} \cdot \frac{d\mu_{\mathbf{y}}}{dP}\ dP\ = \sum_{c_i \in A} x_i^{(\mathbf{p})} y_i^{(\mathbf{p})}\ p_i,$$

$$(c \odot \mu_{\mathbf{x}})(A) = \int_A \left(\frac{d\mu_{\mathbf{x}}}{dP}\right)^c\ dP = \sum_{c_i \in A} (x_i^{(\mathbf{p})})^c\ p_i$$

⇒ perturbation and powering in terms of densities when using a general reference measure (w.r.t. P; SI)

$$\mathbf{x}^{(\mathbf{p})} \oplus^{(\mathbf{p})} \mathbf{y}^{(\mathbf{p})} = \mathbf{x}^{(\mathbf{p})} \oplus \mathbf{y}^{(\mathbf{p})},\ \ c \odot^{(\mathbf{p})} \mathbf{x}^{(\mathbf{p})} = c \odot \mathbf{x}^{(\mathbf{p})}$$

Geometry of a weighted space

- × the scale of **p** impacts the weighted Aitchison inner product, given by

$$\langle \mathbf{x}^{(\mathbf{p})}, \mathbf{y}^{(\mathbf{p})} \rangle_P = \frac{1}{2\sum_{k=1}^{D} p_k} \sum_{i=1}^{D} \sum_{j=1}^{D} p_i p_j \ln \frac{x_i^{(\mathbf{p})}}{x_j^{(\mathbf{p})}} \ln \frac{y_i^{(\mathbf{p})}}{y_j^{(\mathbf{p})}}; \qquad (3)$$

  for $P = P_0$ the usual Aitchison inner product $\langle \mathbf{x}, \mathbf{y} \rangle_A$ from (1) would be obtained

- it is now easy to work with compositions in the "weighted space" given the respective geometrical background

Geometry of a weighted space

- the scale of **p** impacts the weighted Aitchison inner product, given by

$$\langle \mathbf{x}^{(\mathbf{p})}, \mathbf{y}^{(\mathbf{p})} \rangle_P = \frac{1}{2\sum_{k=1}^{D} p_k} \sum_{i=1}^{D} \sum_{j=1}^{D} p_i p_j \ln \frac{x_i^{(\mathbf{p})}}{x_j^{(\mathbf{p})}} \ln \frac{y_i^{(\mathbf{p})}}{y_j^{(\mathbf{p})}}; \qquad (3)$$

  for $P = P_0$ the usual Aitchison inner product $\langle \mathbf{x}, \mathbf{y} \rangle_A$ from (1) would be obtained

- it is now easy to work with compositions in the "weighted space" given the respective geometrical background

- a natural step further is to provide coordinate representations, where standard multivariate statistics can be performed

Logratio coordinates for compositional data

- the main focus in logratio methodology is devoted to express compositions from their original sample space in the standard real space
- **the staying-in-the-simplex approach** (Pawlowsky-Glahn and Egozcue, 2001): to find proper (*isometric*) coordinate representations of compositional data
- centred logratio (clr) coefficients

$$clr(\mathbf{x}) = \left(\frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)$$

- isometric logratio (ilr) coordinates

$$ilr(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_A, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_A)$$

Logratio coordinates for compositional data

- clr coefficients are coefficients with respect to a generating system, thus a singular covariance matrix is obtained due to zero sum constraint of $clr(\mathbf{x})$
- an important class of interpretable ilr coordinates: balances (Egozcue and Pawlowsky-Glahn, 2005), obtained from sequential binary partition (SBP) of the original composition

$$\tilde{x}_j = \sqrt{\frac{r_j s_j}{r_j + s_j}} \ln \frac{\prod_{+} x_i^{1/r_j}}{\prod_{-} x_k^{1/s_j}}, \quad j = 1, \ldots, D-1$$

- parts in numerator and denominator of the logratio correspond to partition of parts in the j-th step, r_j and s_j being numbers of such parts

Logratio coordinates and reference measure

- when weighting of parts $\mathbf{p} = (p_1, \ldots, p_D)$ and the respective reference measure P are considered, coordinate representations need to be adapted accordingly

- weighted clr coefficients with respect to reference P ($\mathbf{p}$-SI)

$$clr_P^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})}) = clr_P^{(\mathbf{p})}\left(\frac{d\mu_{\mathbf{x}}}{dP}\right) = \ln\frac{d\mu_{\mathbf{x}}}{dP} - \frac{1}{P(\Omega)}\int_{\Omega}\ln\frac{d\mu_{\mathbf{x}}}{dP}\,dP,$$

resulting in (Egozcue and Pawlowsky-Glahn, 2016)

$$clr_P^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})}) = \left(\ln\frac{x_1^{(\mathbf{p})}}{g_{\mathbf{p}}(\mathbf{x}^{(\mathbf{p})})}, \ldots, \ln\frac{x_D^{(\mathbf{p})}}{g_{\mathbf{p}}(\mathbf{x}^{(\mathbf{p})})}\right), \; g_{\mathbf{p}}(\mathbf{x}^{(\mathbf{p})}) = \exp\left(\frac{\sum_{i=1}^{D} p_i \ln x_i^{(\mathbf{p})}}{\sum_{i=1}^{D} p_i}\right)$$

- weighted clr coefficients are automatically obtained with respect to P, thus a practical task is how to express them under uniform reference P_0

Properties of weighted clr coefficients

- one-to-one mapping, i.e. inverse mapping can be used
- enable to avoid dealing with perturbation and powering

$$clr_P^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})} \oplus^{(\mathbf{p})} \mathbf{y}^{(\mathbf{p})}) = clr_P^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})}) + clr_P^{(\mathbf{p})}(\mathbf{y}^{(\mathbf{p})}),$$

$$clr_P^{(\mathbf{p})}(\alpha \odot^{(\mathbf{p})} \mathbf{x}^{(\mathbf{p})}) = \alpha \cdot clr_P^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})})$$

- from the weighted inner product

$$\langle \mathbf{x}^{(\mathbf{p})}, \mathbf{y}^{(\mathbf{p})} \rangle_P = \sum_{i=1}^{D} p_i \ln \frac{x_i^{(\mathbf{p})}}{g_{\mathbf{p}}(\mathbf{x}^{(\mathbf{p})})} \ln \frac{y_i^{(\mathbf{p})}}{g_{\mathbf{p}}(\mathbf{y}^{(\mathbf{p})})}$$

it is easy to see that

$$clr_{P_0}^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})}) = \left(\sqrt{p_1} \ln \frac{x_1^{(\mathbf{p})}}{g_{\mathbf{p}}(\mathbf{x}^{(\mathbf{p})})}, \ldots, \sqrt{p_D} \ln \frac{x_D^{(\mathbf{p})}}{g_{\mathbf{p}}(\mathbf{x}^{(\mathbf{p})})} \right)$$

are weighted clr coefficients w.r.t. uniform reference (P_0)

Towards weighted balances

- $clr_P^{(\mathbf{p})}$ is not a coordinate representation, and integration along a $clr_P^{(\mathbf{p})}$ should be carried out with respect to the reference measure

× ilr coordinates are normalized coordinates, i.e. the change of reference has been used to define them

- in general (van den Boogaart et al., 2014), the clr-representation of a density is a function with the same support as the density, whereas the Fourier-coefficients (ilr-coordinates) are discrete sequences of coordinates

Weighted balances

- weighted balances are automatically considered with respect to uniform reference:

$$\tilde{x}_j^{(\mathbf{p})} = \sqrt{\frac{r_j s_j}{r_j + s_j}} \ln \frac{\prod_{+} (x_j^{(\mathbf{p})})^{p_j/r_j}}{\prod_{-} (x_j^{(\mathbf{p})})^{p_j/s_j}}, \quad j = 1, \dots, D-1.$$

…and some further properties can be expected:

Weighted balances

- weighted balances are automatically considered with respect to uniform reference:

$$\tilde{x}_j^{(\mathbf{p})} = \sqrt{\frac{r_j s_j}{r_j + s_j}}\, \ln \frac{\prod_{+} (x_j^{(\mathbf{p})})^{p_j/r_j}}{\prod_{-} (x_j^{(\mathbf{p})})^{p_j/s_j}}, \quad j = 1, \ldots, D-1.$$

. . . and some further properties can be expected:

- the respective basis vectors are orthogonal and of unit length as expected for an orthonormal coordinate representation;

Weighted balances

- weighted balances are automatically considered with respect to uniform reference:

$$\tilde{x}_j^{(\mathbf{p})} = \sqrt{\frac{r_j s_j}{r_j + s_j}}\ \ln \frac{\prod_{+} (x_j^{(\mathbf{p})})^{p_j/r_j}}{\prod_{-} (x_j^{(\mathbf{p})})^{p_j/s_j}}, \quad j = 1, \dots, D-1.$$

…and some further properties can be expected:

- the respective basis vectors are orthogonal and of unit length as expected for an orthonormal coordinate representation;

$$ilr^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})} \oplus^{(\mathbf{p})} \mathbf{y}^{(\mathbf{p})}) = ilr^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})}) + ilr^{(\mathbf{p})}(\mathbf{y}^{(\mathbf{p})}),$$

$$ilr^{(\mathbf{p})}(\alpha \odot^{(\mathbf{p})} \mathbf{x}^{(\mathbf{p})}) = \alpha \cdot ilr^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})}),$$

$$\langle \mathbf{x}^{(\mathbf{p})}, \mathbf{y}^{(\mathbf{p})} \rangle_P = \langle ilr^{(\mathbf{p})}(\mathbf{x}^{(\mathbf{p})}), ilr^{(\mathbf{p})}(\mathbf{y}^{(\mathbf{p})}) \rangle$$

Choice of weights

- from practical reasons it is crucial how the weights $\mathbf{p} = (p_1, \ldots, p_D)$ are chosen

- any such reasonable choice should reflect "importance" of the compositional part according to measurement precision, number of outliers etc.

- one possibility is to weight with reverse log-variances, i.e. penalize parts with higher (absolute) variability

- **the choice of the uniform weights $\mathbf{p}_0$ and setting $p_{i_1,0} \to 0, \ldots, p_{i_k,0} \to 0$ is going towards a subcomposition of $\mathbf{x} = (x_1, \ldots, x_D)$ after excluding parts $x_{i_1}, \ldots, x_{i_k}$** (Egozcue and Pawlowsky-Glahn, 2016)

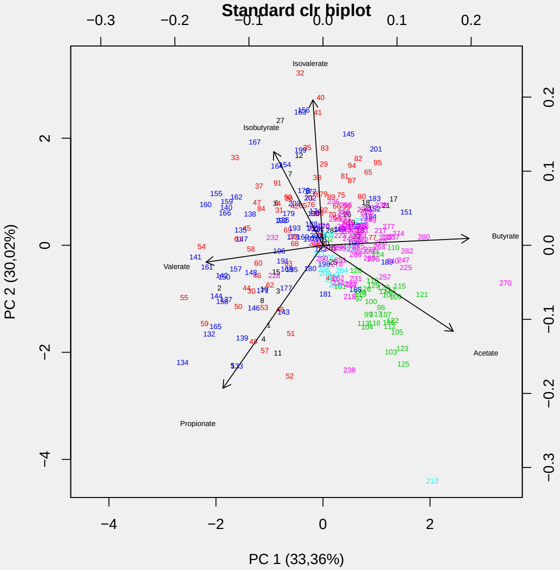
VFA data

- six volatile fatty acids (Acetate, Propionate, Butyrate, Isobutyrate, Isovalerate, Valerate) were measured in 284 compositional samples
- six natural groups of samples occur, relative structure of observations of main interest
- for the original compositions as well as for their weighted counterparts PCA in clr coefficients was applied
- the resulting loadings and scores were displayed in a compositional biplot (Aitchison and Greenacre, 2002)

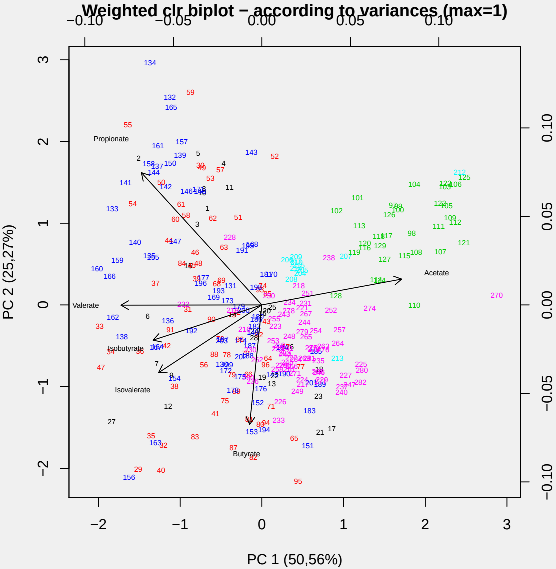
VFA data

Acetate mmol/mol	Propionate mmol/mol	Butyrate mmol/mol	Isobutyrate mmol/mol	Isovalerate mmol/mol	Valerate mmol/mol
557,7331	310,1811	95,7238	12,2168	7,2203	16,9249
477,1865	417,3198	64,0036	11,3342	14,8450	15,3110
542,9800	326,6339	94,0684	9,8870	12,7990	13,6318
562,4834	329,2575	76,3558	9,2198	7,6811	15,0024
551,1794	329,2686	77,9234	8,9214	5,9381	26,7691
537,4469	292,3413	112,9408	14,0573	27,7506	15,4631
567,0722	242,8060	125,4521	14,8281	33,2699	16,5716
557,0352	330,2433	76,1455	8,4567	13,1425	14,9767
562,7468	235,7689	143,4106	11,6374	30,1151	16,3213
555,8863	332,0050	75,5206	10,1947	13,2999	13,0936
⋮	⋮	⋮	⋮	⋮	⋮

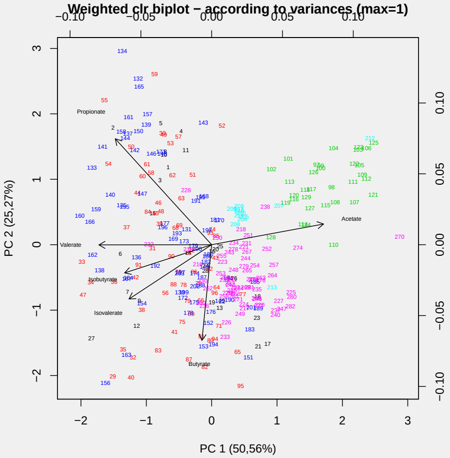
VFA data: the unweighted case



VFA data: penalized variances

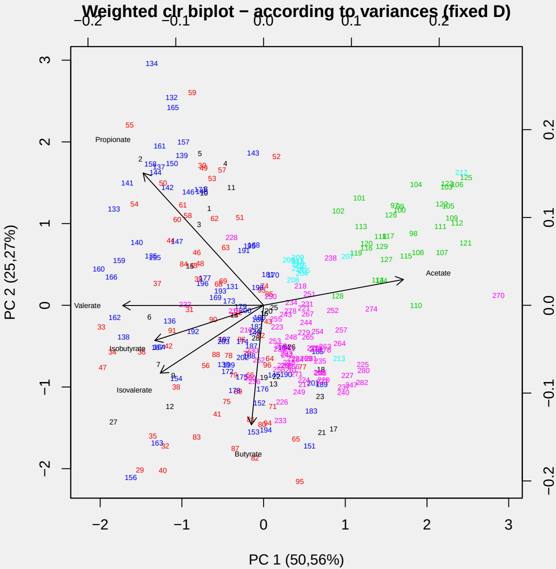


VFA data: penalized variances

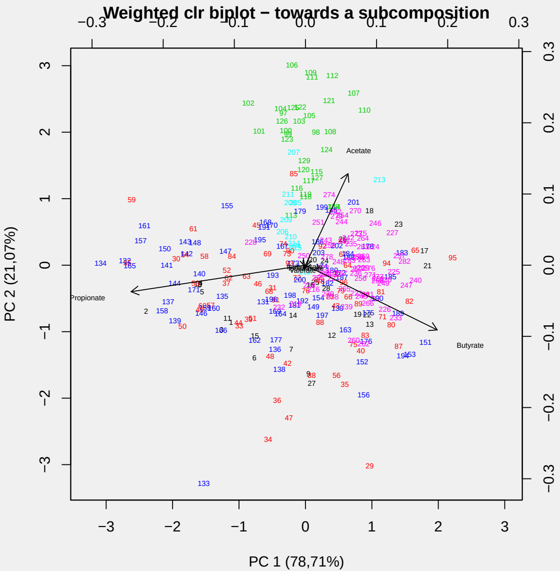


$$\mathbf{p} = (1.00, 0.12, 0.15, 0.07, 0.07, 0.07)$$

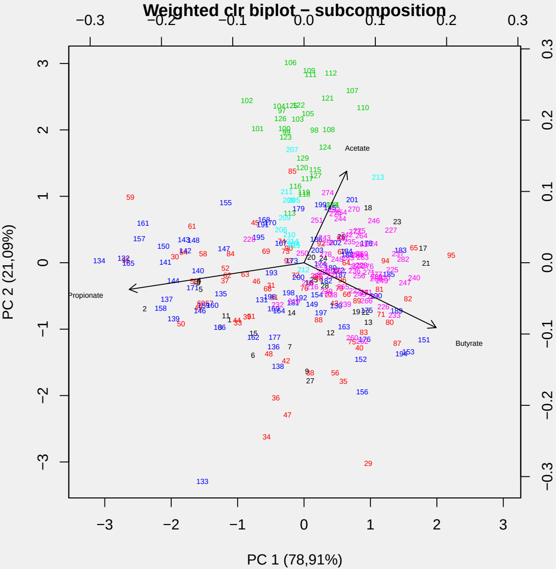
VFA data: penalized variances, fixed sum of **p** (6)



VFA data: towards a subcomposition (0.001)



VFA data: subcomposition



Conclusions

- the Bayes space methodology enables for a systematic approach to weighting of compositional parts

Conclusions

- the Bayes space methodology enables for a systematic approach to weighting of compositional parts
- first results were presented, further research needed

Conclusions

- the Bayes space methodology enables for a systematic approach to weighting of compositional parts
- first results were presented, further research needed
- for practical purposes is crucial to choose the weights properly

Conclusions

- the Bayes space methodology enables for a systematic approach to weighting of compositional parts
- first results were presented, further research needed
- for practical purposes is crucial to choose the weights properly
- possible extension to functional case (densities) and to multifactorial case (compositional tables and cubes)

References

Aitchison, J. : *The statistical analysis of compositional data*. Chapman and Hall, London, 1986.

Aitchison, J., Greenacre, M.: *Biplots of compositional data*. Journal of the Royal Statistical Society, Series C (Applied Statistics) 51, 375–392, 2002.

Eaton, M.L.: *Multivariate statistics: A vector space approach*. Wiley, New York, 1983.

Egozcue, J.J., Pawlowsky-Glahn, V.: *Groups of parts and their balances in compositional data analysis*. Mathematical Geology 37, 795-828, 2005.

Egozcue, J.J., Pawlowsky-Glahn, V.: *Changing the reference measure in the simplex and its weighting effects*. Austrian Journal of Statistics 45, 25-44, 2016.

Pawlowsky-Glahn, V., Egozcue, J.J.: *Geometric approach to statistical analysis on the simplex*. Stochastic Environmental Research and Risk Assessment (SERRA) 15, 384–398, 2001.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: *Modeling and analysis of compositional data*. Wiley, Chichester, 2015.

van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V.: *Bayes Hilbert spaces*. Australian & New Zealand Journal of Statistics 56, 171-194, 2014.