Weighted pivot balances applied to metabolomics

Alžběta Gardlo[a], Karel Hron[b]

[a] **alzbetagardlo@gmail.com**; Laboratory of Metabolomics, IMTM, Faculty of Medicine and Dentistry, Olomouc, Czech Republic; Faculty of Science, Palacký University Olomouc, Czech Republic; University Hospital Olomouc, Czech Republic;
[b] Faculty of Science, Palacký University Olomouc, Czech Republic

New system of pivot coordinates

Every D-part composition $\mathbf{x} = (x_1, \ldots, x_D)^T$ can be expressed in $D-1$ orthonormal coordinates with respect to the Aitchison geometry. No canonical basis in the Aitchison geometry exist $\Rightarrow$ interpretable coordinates are of primary interest. **The aim** is to capture the relative information to a specific part of interest, x_1, with one of coordinates. **The question** is, which parts need to be considered. Also some parts might have data **problems** that destroy the relative information to x_1. All relative information about part x_1 can be represented by a coordinate z_1, and we can construct orthonormal coordinates (balances) $\mathbf{z} = (z_1, \ldots, z_{D-1})^T$ to z_1 by (ilr coordinates)

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{k=i+1}^{D} x_k}}, \; i = 1, \ldots, D-1. \tag{1}$$

Part x_1 is contained only in z_1, which can also be expressed as scaled sum of $\ln(x_1/x_2) + \cdots + \ln(x_1/x_D)$. Coordinates can be represented also by **log-contrasts**

$$a_1 \ln x_1 + \ldots + a_D \ln x_D = \mathbf{a}^T \ln(\mathbf{x}), \quad \text{where } a_j \in \mathbb{R}, \sum_{j=1}^{D} a_j = 0. \tag{2}$$

Weighted pivot coordinates

Weighted counterpart to z_1 is done by weighted sum of pairwise log-ratios to x_1:

$$\alpha_2 \ln \frac{x_1}{x_2} + \ldots + \alpha_D \ln \frac{x_1}{x_D}, \quad \alpha_k > 0, \; k = 1, \ldots, D; \; \alpha_2 + \cdots + \alpha_D = 1. \tag{3}$$

With unit norm constraint of associated log-contrast we obtain a coordinate

$$w_1 = \frac{1}{\sqrt{1 + \sum_{k=2}^{D} \alpha_k^2}} \ln \frac{x_1}{\prod_{k=2}^{D} x_k^{\alpha_k}}. \tag{4}$$

The remaining $D-2$ coordinates can be obtained by solving a sequence of properly chosen homogeneous linear systems.

Choice of weights

Basic measure of variability for $\mathbf{x} = (x_1, \ldots, x_D)^T$ is **variation matrix** $\mathbf{T} = \left\{ \text{var}\left(\ln \frac{x_i}{x_j}\right) \right\}_{i,j=1}^{D}$, where we assume that the part of interest is x_1, thus look at first column (row) of $\mathbf{T}$: $\mathbf{t}_1 = (t_{11}, \ldots, t_{1D}) = \left(\text{var}\left(\ln \frac{x_1}{x_1}\right), \text{var}\left(\ln \frac{x_1}{x_2}\right) \ldots, \text{var}\left(\ln \frac{x_1}{x_D}\right)\right)$.

Weights - version 1:

$$\tilde{\alpha}_j^{(p)} = \frac{1}{(t_{1j})^p}, \qquad \text{for } j = 2, \ldots, D \text{ and } p > 0, \tag{5}$$

that for $p = 2$ assign reverse values of variances of pairwise logratios with x_1 (further normed to get $\alpha_2, \ldots, \alpha_D$).

Weights - version 2: power is defined with respect to specific groups in data (g_1 =patients/g_2 =controls):

$$m_{ij} = \text{mean}(\log(x_{g_i,1}/x_{g_i,j}) \qquad \text{for } i = 1, 2, \quad j = 2, \ldots, D,$$

$$pow_j = \log\left(\frac{1}{\text{mean}(\text{abs}(m_{1j} - m_{2j}))}\right).$$

Normalized power:

$$pown = \frac{pow - \min(pow)}{\max(pow) - \min(pow)} p \qquad \text{for } p > 0.$$

Weights:

$$\tilde{\alpha}_j^{(p)} = \frac{1}{(t_{1j})^{pown_j}} \qquad \text{for } j = 2, \ldots, D \text{ and } p > 0. \tag{6}$$

Choice of weights II

Weights - version 3/4: weights are defined as the absolute difference between groups: $pown_j = \frac{\text{abs}(m_{1j} - m_{2j})}{\sqrt{t_{1j}}}$.

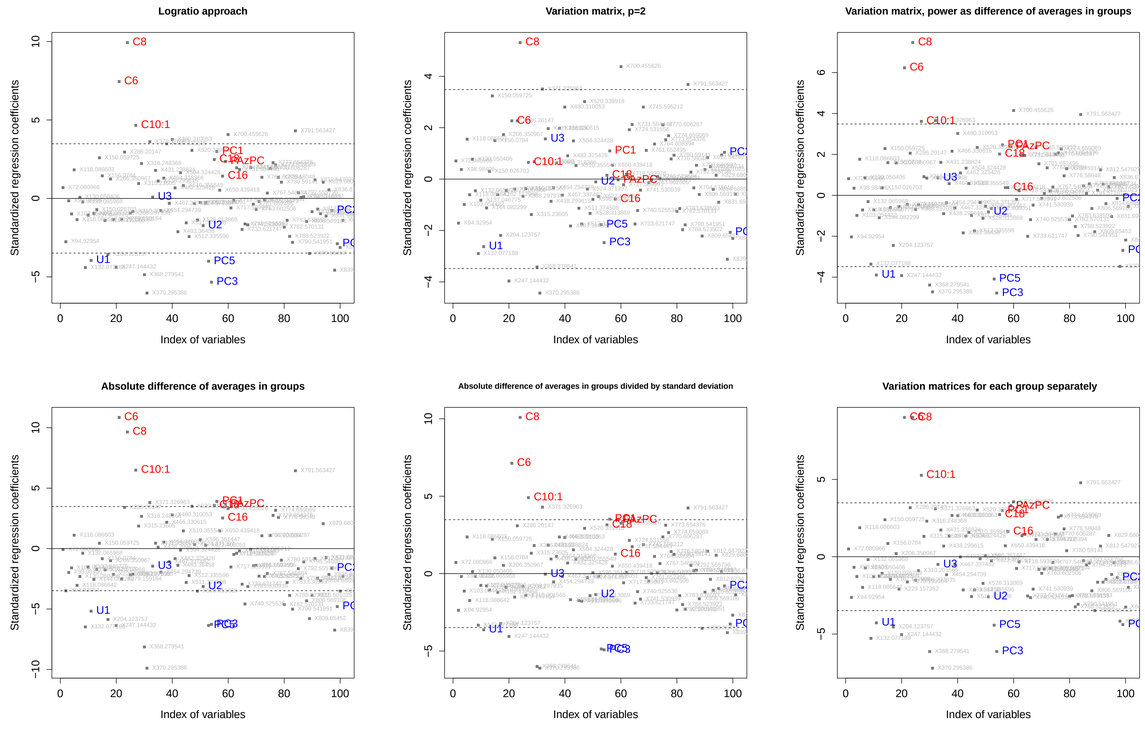
Weights - version 5: weights are based on information from variation matrix of the whole data set $\mathbf{T}$ and variation matrices of specific groups $\mathbf{T}^A$ and $\mathbf{T}^B$: $pown_j = \frac{n \times \sqrt{t_{1j}}}{n_1 \times \sqrt{t_{1j}^A} + n_2 \times \sqrt{t_{1j}^B}}$.

Weights are defined with

$$\tilde{\alpha}_j^{(p)} = pown_j^p \qquad \text{for } j = 2, \ldots, D \text{ and } p > 0. \tag{7}$$

Real data from metabolomics

Data comparing dry blood spots of healthy controls and patients suffered from Medium chain acyl-CoA dehydrogenase deficiency with 25 samples in each group. 101 metabolites was measured → highdimensional data → partial least squares - discriminant analysis (PLS-DA). The significance of the standardized regression coefficients is analyzed using bootstrap with 50 repetitions [2].



Conclusion

The best option for metabolomic data is **version 3**, where weights are defined as absolute difference of averages in specific groups. It identifies the highest amount of markers and only few false positives are present for the group of patients (red ones).

References

1. Hron, K. et al. *Weighted pivot coordinates for compositional data and their application to geochemical mapping.* MATH GEOSCI, to appear, DOI 10.1007/s11004-017-9684-z.

2. Kalivodová, A., et al. *PLS-DA for compositional data with application to metabolomics.* J CHEMOMETR 29, 21-28, 2014.

Acknowledgements

This study was funded by the grant 15-34613L of the Czech Science Foundation (GA ČR) and the project LO1304 of the Ministry of Education, Youth and Sports of the Czech Republic.