ROBUST PRINCIPAL COMPONENT ANALYSIS FOR COMPOSITIONAL TABLES

J. RENDLOVÁ[1], K. FAČEVICOVÁ[1], K. HRON[1], P. FILZMOSER[2]

[1]Palacký University Olomouc, Czech Republic, julie.rendlova@gmail.com

[2]Vienna University of Technology, Austria

Many practical examples contain relative information about the distribution according to two factors which leads to a $(I \times J)$-dimensional extension of compositional data called compositional tables. This contribution proposes a robust approach to principal component analysis of the compositional tables in order to explore information about a relationship between the given factors. The theoretical background is illustrated on a real data set containing unemployment information with gender distribution and age structure from OECD Statistics. Data from several different countries are processed as a set of 2×4 compositional tables using R, namely the robCompositions package.

COMPOSITIONAL TABLES

- direct extension of vector compositional data [1,7]

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}, x_{ij} > 0, i = 1, ..., I, j = 1, ..., J$$

- accordingly, also the Aitchison geometry can be extended [4,5]

$$\mathcal{C}(\mathbf{x}) = \begin{pmatrix} \frac{\kappa x_{11}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{1J}}{\sum_{i,j} x_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\kappa x_{I1}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{IJ}}{\sum_{i,j} x_{ij}} \end{pmatrix},$$

$$\mathcal{S}^{IJ} = \left\{ \boldsymbol{x} = (x_{11}, ..., x_{IJ}) | x_{ij} > 0, i = 1, ..., I, j = 1, ..., J, \sum_{i,j} x_{ij} = \kappa \right\},$$

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C} \begin{pmatrix} x_{11}y_{11} & \cdots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \cdots & x_{IJ}y_{IJ} \end{pmatrix}, \quad \alpha \odot \mathbf{x} = \mathcal{C} \begin{pmatrix} x_{11}^{\alpha} & \cdots & x_{1J}^{\alpha} \\ \vdots & \ddots & \vdots \\ x_{I1}^{\alpha} & \cdots & x_{IJ}^{\alpha} \end{pmatrix},$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}}, \quad \|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A}, \quad \text{and} \quad d_A(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_A$$

- vectorized compositional table $\boldsymbol{x}$ used in $\mathcal{S}^{IJ} \Rightarrow$ dimension of the simplex is $IJ - 1$

Decomposition

- *independence* and *interaction tables* obtained through an orthogonal decomposition $\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int}$ done by row and column projections of the compositional table onto marginal subspaces [1]

$$x_{ij}^{ind} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} x_{kj} x_{il} \right)^{\frac{1}{IJ}} \quad \text{and} \quad x_{ij}^{int} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} \frac{x_{ij}}{x_{kj} x_{il}} \right)^{\frac{1}{IJ}}$$

- independence table: all information describing relationships within the given factors in case of their independence (by geometric marginals)
- interaction table: information about the "true" relationship between the row and column factors
- analysis of $\mathbf{x}_{ind}$ and $\mathbf{x}_{int}$ allows for a deeper insight into the original data
- dimensions of $\mathbf{x}_{ind}$ and $\mathbf{x}_{int}$ lower to $I + J - 2$ for $\mathcal{S}_{ind}^{IJ}$, and to $(I-1)(J-1)$ for $\mathcal{S}_{int}^{IJ}$

Pivot coordinates and centered logratio coefficients

- ilr coordinate representation of compositional data needed to perform robust methods
- $(IJ-1)$-dimensional alternative to pivot coordinates [2] for two factor analysis
- first $I + J - 2$ coordinates of the entire compositional table coincide with pivot coordinates of the independence table, the remaining $(I-1)(J-1)$ with the interaction table
- types of pivot coordinates correspond to row, column and "odds ratio" partitioning of the compositional table

$$z_i^r = \sqrt{\frac{(I-i)J}{1+I-i}} \ln \frac{g(\boldsymbol{x}_{i\bullet})}{[g(\boldsymbol{x}_{i+1\bullet}) \cdots g(\boldsymbol{x}_{I\bullet})]^{1/(I-i)}},$$

$$z_j^c = \sqrt{\frac{I(J-j)}{1+J-j}} \ln \frac{g(\boldsymbol{x}_{\bullet j})}{[g(\boldsymbol{x}_{\bullet j+1}) \cdots g(\boldsymbol{x}_{\bullet J})]^{1/(J-j)}},$$

$$z_{rs}^{OR} = \sqrt{\frac{1}{(I-r)(J-s)(I-r+1)(J-s+1)}} \ln \prod_{i=r+1}^{I} \prod_{j=s+1}^{J} \frac{x_{ij}x_{rs}}{x_{is}x_{rj}},$$

- obtained pivot coordinates might be transformed back to simplex or directly to clr coefficients

$$\mathrm{clr}(\mathbf{x}) = \mathbf{V}\mathbf{z}, \quad \text{where} \quad \mathbf{V} = (\mathrm{clr}(\boldsymbol{e}_1), \mathrm{clr}(\boldsymbol{e}_2), ..., \mathrm{clr}(\boldsymbol{e}_{IJ-1})),$$

$$\mathrm{clr}(\boldsymbol{e}^r) = \begin{cases} \sqrt{\frac{I-i}{(I-i+1)J}} & \text{for elements in pivot row } i, \\ -\sqrt{\frac{1}{(I-i+1)J(I-i)}} & \text{for elements in rows } i+1, ..., I, \\ 0 & \text{otherwise,} \end{cases} \quad \mathrm{clr}(\boldsymbol{e}^c) = \begin{cases} \sqrt{\frac{J-j}{(J-j+1)I}} & \text{for elements in pivot column } j, \\ -\sqrt{\frac{1}{(I-i+1)J(I-i)}} & \text{for elements in columns } j+1, ..., J, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathrm{clr}(\boldsymbol{e}^{OR}) = \begin{cases} \sqrt{\frac{1}{rs(r-1)(s-1)}} & \text{for elements on positions } i = r+1, ..., I, \\ & j = s+1, ..., J, \\ \sqrt{\frac{(r-1)(s-1)}{rs}} & \text{for pivot elements } rs, \\ -\sqrt{\frac{r-1}{rs(s-1)}} & \text{for elements in pivot row } r, j = s+1, ..., J, \\ -\sqrt{\frac{s-1}{rs(r-1)}} & \text{for elements in pivot column } s, i = r+1, ..., I, \\ 0 & \text{otherwise} \end{cases}$$

- resulting clr coefficients can be expressed by geometric means

$$\mathrm{clr}(\mathbf{x}_{ind})_{ij} = \ln \frac{g(\boldsymbol{x}_{i\bullet}) g(\boldsymbol{x}_{\bullet j})}{g(\boldsymbol{x}_{\bullet\bullet})^2}, \quad \text{and} \quad \mathrm{clr}(\mathbf{x}_{int})_{ij} = \ln \frac{x_{ij} g(\boldsymbol{x}_{\bullet\bullet})}{g(\boldsymbol{x}_{i\bullet}) g(\boldsymbol{x}_{\bullet j})},$$

where $g(\boldsymbol{x}_{i\bullet})$, $g(\boldsymbol{x}_{\bullet j})$ and $g(\boldsymbol{x}_{\bullet\bullet})$ stand for geometric mean of i-th row, j-th column and the whole compositional table, respectively

Interpretation of centered logratio coefficients

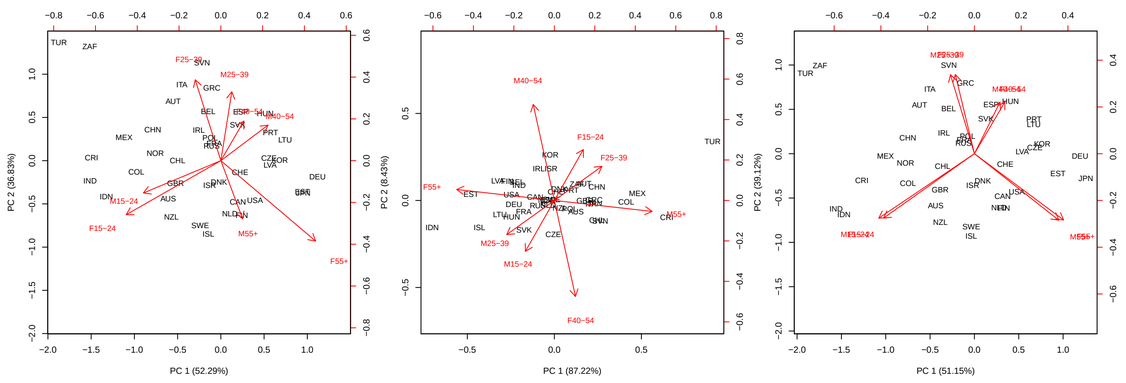
- each $\mathrm{clr}(\mathbf{x}_{ind})_{ij}$ is a dominance of a given combination of factor values in case of independence
- dominance amplified or weakened according to the interaction table (depending on whether it is shifted in a positive or a negative direction)
- interaction table refers also about sources of departures from independence
- information obtained only from $\mathrm{clr}(\mathbf{x}_{int})_{ij}$ does not provide complete picture about the dominance

ROBUST PCA FOR COMPOSITIONAL TABLES

- classical PCA might be strongly affected by outliers
- robust PCA transformation defined as $\mathbf{X}^* = (\mathbf{X} - \mathbf{1}\boldsymbol{t}^T)\mathbf{G}$ [3], where $\boldsymbol{t}$ is the *Minimum Covariance Determinant (MCD)* estimator of location and $\mathbf{1}$ is a vector of ones of length n
- ilr coordinates are needed to obtain full rank MCD estimate of covariance matrix
- pivot coordinates on input are transformed into scores $\boldsymbol{z}^* = \mathbf{G}^T(\boldsymbol{z} - \boldsymbol{t})$
- loadings (columns of $\mathbf{G}$) need to be transformed back to clr coefficients $\mathbf{G}^* = \mathbf{V}\mathbf{G}$ accounting for compositional biplot construction with meaningful interpretation

UNEMPLOYMENT DATA ANALYSIS

- data from OECD Statistics about unemployed people from 42 different countries in 2010 depending on their gender and age category (15-24, 25-39, 40-54 and 55+) [6]
- analysis done using the statistical software R, namely the robCompositions package



Robust biplots of the unemployment compositional (left figure), interaction (middle figure), and independence tables (right figure)

Overall picture from the biplot on the left

- most European countries together with the USA and Canada tend to have higher unemployment among older people, say 40+
- central and south America together with China, India and Indonesia face the higher rate rather in the opposite situation
- outlyingness of Turkey and South Africa is caused by unexpectedly high ratio of young unemployed

How to interpret the biplot in the middle

- focus e.g. on Costa Rica (in the direction of the arrow for a loading "men 55+")
- arrow for a loading "men 55+" just marks the group causing imbalance for Costa Rica; the true ratio of unemployment for "men 55+" can be actually lower
- higher dominance of unemployed men in the oldest group than expected in the hypothetical case of independence can be concluded only after looking at the independence tables

What the last biplot represents

- "ideal" situation in case the relationships between gender and age factors would be filtered away
- nearly gender equity; disproportionately weaker relationships between the age levels

Please note that using classical PCA (which does not handle outliers) the resulting figures would be far less illustrative. Especially the gender equity achieved in the robust biplot of independence tables would not be present in the classical one.

References:

[1] Egozcue, J. J., Díaz-Barrero, J. L., Pawlowsky-Glahn, V. (2008). Compositional Analysis of Bivariate Discrete Probabilities. In Daunis-i-Estadella, J., Martín-Fernández, J. A. (Eds.), *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop*. University of Girona, Spain.

[2] Fačevicová, K., Hron, K., Todorov, V., Templ, M. (2016). Compositional Tables Analysis in Coordinates. *Scandinavian Journal of Statistics 43*(4), pp. 962–977.

[3] Filzmoser, P., Hron, K., Reimann, C. (2009). Principal Component Analysis for Compositional Data with Outliers. *Environmetrics 20*(6), pp. 621–632.

[4] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*, Wiley, Chichester.

[5] Aitchison, J. (1986). *The statistical analysis of compositional data*, Chapman and Hall, London.

[6] OECD Statistics. (2017). *Unemployment by sex and age - 2010*. (Available from http://stats.oecd.org/.) [Accessed March 10, 2017].

[7] Egozcue, J.J., Pawlowsky-Glahn, V., Templ, M., Hron, K. (2015). Independence in contingency tables using simplicial geometry. *Commun. Stat. Theory 44*(18), pp. 3978–3996.