Variable Selection Stepwise Procedure for Compositional Data

S. Donevska [1] P. Filzmoser [2] E. Fišerová [1] K. Hron [1]

[1] Palacký University in Olomouc, Czech Republic
[2] Vienna University of Technology, Austria

LinStat2014

Motivation

- Why do we usually omit variables?

  $\Rightarrow$ We want to simplify the multivariate statistical analysis and also because we want to simplify the interpretation of the results.

- How do we know which variables to exclude?

  $\Rightarrow$ We usually ask the experts...

  POSSIBLE PROBLEMS: Major changes of the multivariate statistical analysis results.

$\Rightarrow$ **SOLUTION: The proposed covariance-based stepwise procedure for variable selection guarantees that the loss of the information when moving from composition to subcomposition will be rather negligible.**

Compositional data

Compositional data (CoDa) = quantitative descriptions of parts of some whole, thus as data carring only **relative information**.

- Simplex with the Aitchison geometry= the sample space of CoDa,

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, \ldots, x_D)',\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa\}.$$

- Aitchison geometry on the simplex is not completely suitable for performing standard statistical methods on the CoDa.
  - $\Rightarrow$ This fact leads to necessity to find proper representations of the CoDa to the real space.

- For this purpose are proposed log-ratio transformations: additive log-ratio (alr) transformation, centered log-ratio (clr) transformation and isometric log-ratio (ilr) transformation.

- Representation of CoDa based on ratio of parts is convenient.

Clr transformation

The clr transformation is an isometric mapping between $\mathcal{S}^D$ and a hyperplane of $\mathbb{R}^D$,

$$\mathbf{y} = clr(\mathbf{x}) = (y_1, y_2, ..., y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, ..., \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)' . \quad (1)$$

- Disadvantages of the clr variables:
  - they are not coordinates with respect to a basis on the simplex,
  - they lead to collinear data, because $y_1 + \cdots + y_D = 0$,
  - they are not subcompositionaly coherent.
- Advantages of the clr variables:
  - they translate perturbation and powering of CoDa into ordinary sum and multiplication by a scalar of vectors of clr coefficients,
  - Euclidean distance between vectors of clr coefficients = Aitchison distance of their corresponding compositions. This also holds for the inner product and the norm.

Measures of variability of CoDa

The basic measure of variability of a random composition $\mathbf{x} = (x_1, \dots, x_D)'$ is the variation matrix defined as

$$\mathbf{T} = \left\{ \mathrm{var}\left(\ln \frac{x_i}{x_j} \right) \right\}_{i,j=1}^{D} .$$

- $\mathbf{T}$ is symmetrical matrix with zeros on the main diagonal.
- The elements of $\mathbf{T}$ describe the variability of the log-ratio between x_i and x_j.

The (normed) sum of the elements of the variation matrix is called total variance,

$$\mathrm{totvar}(\mathbf{x}) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{var}\left(\ln \frac{x_i}{x_j} \right),$$

expressing the total variability of the compositional data set.

Covariance structure

- Total variance of compositional data set $\mathbf{x}$ can be expressed as $\text{totvar}(\mathbf{x}) = \sum_{i=1}^{D} \mathrm{var}(y_i)$, where

$$\mathrm{var}(y_i) = \frac{D-1}{D^2} \sum_{j=1}^{D} \mathrm{var}\left(\ln \frac{x_j}{x_i}\right) - \frac{1}{2D^2} \sum_{\substack{j=1 \\ j \neq i}}^{D} \sum_{\substack{l=1 \\ l \neq i}}^{D} \mathrm{var}\left(\ln \frac{x_j}{x_l}\right).$$

$\Rightarrow$ Strong relation between $\mathrm{var}(y_i)$ and the sum of the i-th row (column) of the corresponding variation matrix $\mathbf{T}$.

Theorem

Consider the clr variables y_i and y_j, $i \neq j$, $i, j = 1, \ldots, D$. Then $\mathrm{var}(y_i) \geq \mathrm{var}(y_j)$, if and only if

$$\sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_i}{x_p}\right) \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_j}{x_p}\right).$$

Proposed stepwise procedure

Let us consider a composition $\mathbf{x} = (x_1, \ldots, x_D)'$, such that

$$\mathrm{var}(y_1) \geq \cdots \geq \mathrm{var}(y_D) \qquad (2)$$

$$\Leftrightarrow$$

$$\sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_1}{x_p}\right) \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_2}{x_p}\right) \geq \cdots \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_D}{x_p}\right). \qquad (3)$$

Algorithm:

1. Omit the part x_D whose variance of the corresponding clr variable is the smallest.
   Consider a subcomposition $\mathbf{x}_1 = (x_1, \ldots, x_{D-1})'$ and perform a clr transformation on $\mathbf{x}_1$.
   Calculate variances of the clr transformed variables of $\mathbf{x}_1$.
2. Repeat step 1.
3. STOP maximally after $D - 2$ steps.

Proposed stepwise procedure

- Will the order of the clr variances be maintained after omitting x_D?

$\Rightarrow$ The order of the clr variances when moving from a D-part to a $(D-1)$-part composition is maintained only under the assumption

$$\mathrm{var}\left(\ln\frac{x_1}{x_D}\right) \geq \mathrm{var}\left(\ln\frac{x_2}{x_D}\right) \geq \cdots \geq \mathrm{var}\left(\ln\frac{x_{D-1}}{x_D}\right).$$

- When the selection of parts should be stopped?

$\Rightarrow$ After using a stop criterion that will compare the total variance of the $\mathbf{x}_i$, obtained in the i-th step of the algorithm, $i = 1, \ldots, D-2$, with the total variance of $\mathbf{x}_{i-1}$.

Proposed stepwise procedure - STOP criterion

$\mathrm{H}_0: \ \mathrm{totvar}(\mathbf{x}_i) = \mathrm{totvar}(\mathbf{x}_{i-1})$ v.s. $\mathrm{H}_A: \ \mathrm{totvar}(\mathbf{x}_i) < \mathrm{totvar}(\mathbf{x}_{i-1})$

- For this purpose we use the following test statistic:

$$U_i^+ = \frac{\widehat{\mathrm{totvar}}(\mathbf{x}_i) - \mathrm{totvar}(\mathbf{x}_{i-1})}{\sqrt{\frac{2}{n-1}\mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}}_i^2\right)}},$$

  where $\widehat{\boldsymbol{\Sigma}}_i$ stands for the sample covariance matrix of the composition $\mathbf{x}_i$ in (arbitrarily chosen) ilr coordinates.

- H_0 is rejected if $u_i^+ \in \mathrm{W} = (-\infty, u_\alpha\rangle$, where u_i^+ is the realization of U_i^+ and u_α denotes the α-quantile (preferably $\alpha = 0.05$) of the standard normal distribution.

- In each step we compute U_i^+ and the procedure is stopped when $u_i^+ \in \mathrm{W}$ for the first time.

Example - Kola Data

Kola data set is a result of a large geochemical mapping project, carried out from 1992 to 1998 by the Geological Surveys of Finland and Norway, and the Central Kola Expedition, Russia.

- An area covering 188 000 km^2 at the peninsula Kola in Northern Europe was sampled.
- In total, around 600 samples of soil were taken in 4 different layers (moss, humus, B-horizon, C-horizon).
- The samples were analyzed by a number of different techniques for more than 50 chemical elements.
- The primary idea of the project was to reveal the environmental conditions in the area.
- The data are available in the package `StatDA` of the software environment R (R Development Core Team, 2012).

Example - Kola Data - First experiment

- 15 variables are selected randomly from 31 elements of the moss layer.
- The stepwise procedure is applied until is reached a 2-part subcompostiion.
- In each step is computed the total variance.
- Whole procedure is repeated for 1000 times.

Example - Kola Data - First experiment

- 15 variables are selected randomly from 31 elements of the moss layer.
- The stepwise procedure is applied until is reached a 2-part subcompostiion.
- In each step is computed the total variance.
- Whole procedure is repeated for 1000 times.

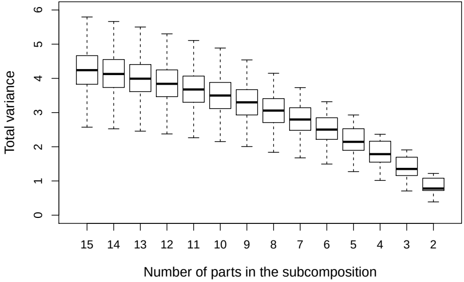

Figure: Total variances of subcompositions obtained from the stepwise algorithm.

Example - Kola Data - Second experiment

- Again 15 variables are selected randomly from 31 elements of the moss layer.
- The stepwise procedure is applied until the test statistic suggests to stop the process.
- Whole procedure is repeated for 1000 times.

Example - Kola Data - Second experiment

- Again 15 variables are selected randomly from 31 elements of the moss layer.
- The stepwise procedure is applied until the test statistic suggests to stop the process.
- Whole procedure is repeated for 1000 times.

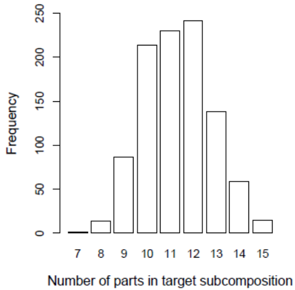

Figure: Barplot of the number of parts of the subcomposition resulting from the stepwise procedure using the stop-criterion.

Example - Kola Data - Second experiment

Consists the resulting target compositions of parts with large clr variances of the initial compositions, or not?

- The parts of all 1000 initial subcompositions are sorted according to decreasing values of their clr variances.
- We count how often the top k clr variables were included in the target compositions, where $k = 1, \ldots, 15$.

Example - Kola Data - Second experiment

Consists the resulting target compositions of parts with large clr variances of the initial compositions, or not?

- The parts of all 1000 initial subcompositions are sorted according to decreasing values of their clr variances.
- We count how often the top k clr variables were included in the target compositions, where $k = 1, \ldots, 15$.

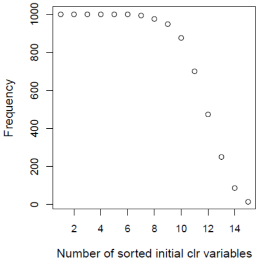

Figure: Clr variables of the initial composition, sorted according to decreasing variance, versus number of times the corresponding compositional parts were included in the resulting subcomposition.

Example - Kola Data - Third experiment

- We use the same simulation setting as before, but select as initial composition 5, 10, 20, and 25 parts of the Kola moss data, respectively.
- Repeat each case 1000 times.

Example - Kola Data - Third experiment

- We use the same simulation setting as before, but select as initial composition 5, 10, 20, and 25 parts of the Kola moss data, respectively.
- Repeat each case 1000 times.

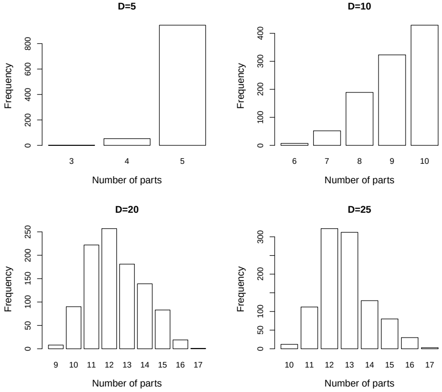

Figure: Barplots of the number of parts of the subcomposition resulting from the stepwise procedure using the stop-criterion with D-part original compositions.

Example - Kola Data - Fourth experiment

- The stepwise procedure is applied to the whole moss layer data set (31 compositional parts).

Example - Kola Data - Fourth experiment

- The stepwise procedure is applied to the whole moss layer data set (31 compositional parts).

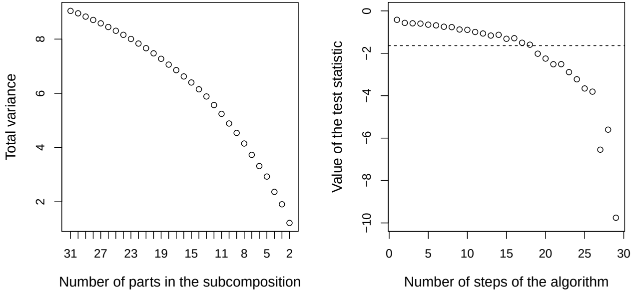

Figure: Total variances of subcompositions obtained from the stepwise algorithm for the whole moss layer data set (left), corresponding values of the test statistic U_i^+ together with the cut-off value (right).

Conclusion

- **The proposed stepwise procedure for variable selection guarantees the presence of compositional parts in the resulting subcomposition, conveying important information about multivariate data structure.**

- **The reduction of the compositional parts leads to consequent facilitation of the analysis and simultaneously to simplification of the interpretation of the results of the multivariate statistical analysis.**

References

- Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London.
- Egozcue JJ (2009) Reply to ”On the Harker Variation Diagrams; . . . ” by J. A. Cortés. Math Geosci 41:829–834.
- Filzmoser P, Hron K, Reimann C (2012) Interpretation of multivariate outliers for compositional data. Computers & Geosciences 39: 77–85.
- Hron K, Filzmoser P, Donevska S, Fišerová E (2013) Covariance-based variable selection for compositional data. Mathematical Geosciences 45: 487–498.
- Hron K, Kubáček L (2011) Statistical properties of the total variation estimator for compositional data. Metrika 74: 221–230.