Optimizing Clustering of Compositional Data: A Comparative Study of Divergence Measures

Muhammad Shoaib, Eva Riccomagno muhammad.shoaib@dima.unige.it University of Genova, Italy

Università di Genova DIMA Dipartimento di Matematica SMiG LAB CLOE

1. CLOE

CLOE is an interdisciplinary and inter-sectoral Doctoral Programme co-funded by the EU and developed by the University of Genoa in collaboration with academic and non-academic host organisations.

1. Abstract

The study compares various divergence measures for clustering compositional data, focusing on Kaniadakis' divergence, a special case of the C-KL divergence with escort probabilities. Kaniadakis' divergence performs well with K-means clustering, achieving the lowest WSS (0.5), highest CHI (189.97), and highest AS (0.37) for simulation data. For real alimentation data, it also shows strong performance with CHI = 21 and AS = 0.39, highlighting its effectiveness for compositional clustering.

2. Introduction

A composition is formally defined as a vector x on the $(D-1)$ dimensional simplex space.

$$\mathcal{S}^D = \{\mathbf{x} = [x_1, .., x_D] :$$
$$x_i > 0, i = 1, ,.., D; \sum_{i=1}^{D} x_i = \kappa\},$$

In [1] describes a completely algebraic form to summarise Kaniadakis' logarithm. According to [2], the reciprocal derivative function A is related to generalised logarithms. We cannot incorporate with zeros because of logratios.Kaniadakis' logarithm is defined as.

$$log_\kappa(x) = \frac{1}{2}\left(x - \frac{1}{x}\right) = \int_1^x \frac{du}{A(u)}$$
$$where \quad A(u) = \frac{2u^2}{1+u^2}$$

3. Divergence Criteria for CoDa

key compositional properties:

1. **Scale Invariance:** $d(\lambda x, \mu y) = d(x,y), \ \forall\ \lambda,\ \mu \in R^+ \quad \forall\ x,\ y \in \mathcal{S}^D.$
2. **Subcompositional Dominance:** $d(s_x, s_y) \leq d(x,y)$
3. **Perturbation Invariance:** $d(x \oplus z, y \oplus z) = d(x,y) \quad \forall\ x,\ y,\ z\ \in\ \mathcal{S}^D.$

8. References

[1] Giovanni Pistone and Muhammad Shoaib. Kaniadakis's information geometry of compositional data. *Entropy*, 25(7):1107, 2023.

[2] Jan Naudts. Generalised exponential families and associated entropy functions. *Entropy*, 10(3):131–149, 2008.

5. Framework

The Kaniadakis divergence is defined by changes in the usual definition of the logarithm to the Kaniadakis logarithm and the escort probability $\tilde{p}$. Aitchision distance quantifies the difference between the compositional vectors while considering the constraints that the components sum to a constant. Jeffreys divergnece measures how much more information is needed to describe one dataset in terms of the other. If Jeffreys divergence is zero, it indicates that the two datasets are identical. C-KL quantifies how one compositional distribution (e.g., the composition of a sample) differs from another (e.g., a reference composition).

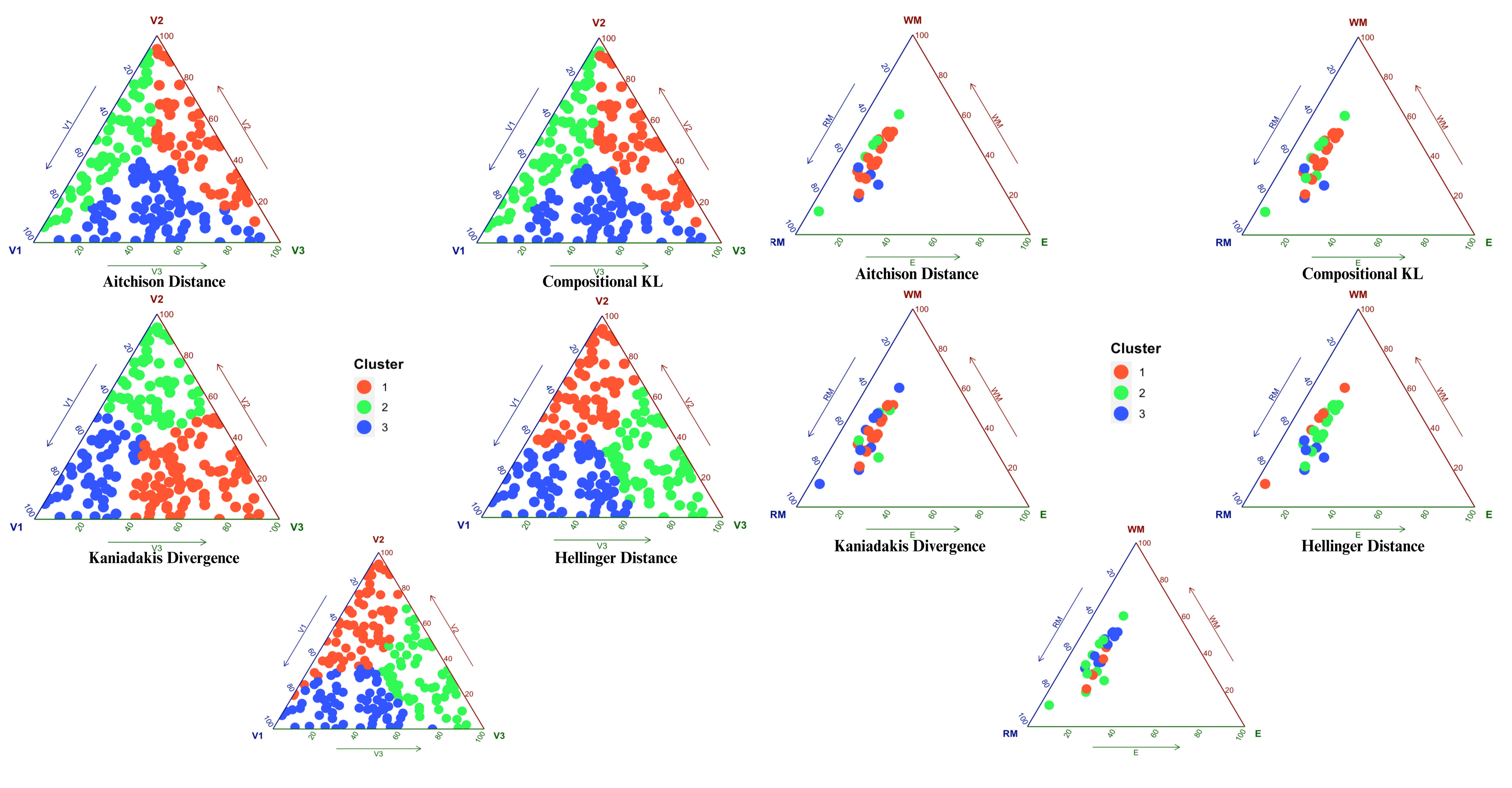
Divergences

- Aitchison Distance, *J Aitchison 1982*: $\sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(ln\frac{x_i}{x_j} - ln\frac{y_i}{y_j}\right)^2}$
- Compositional KL Divergence, *J Palarea et al. 2012*: $\sqrt{\frac{D}{2} log\left(\frac{1}{D^2}\sum_{i=1}^{D}\frac{x_i}{y_i}\sum_{i=1}^{D}\frac{y_i}{x_i}\right)}$
- Kaniadakis' Divergence [1]: $\sum_{x\in\Omega}(log_\kappa p - log_\kappa q)\widetilde{p}(x)$
  - $\widetilde{p}(x)$: $\frac{A(p(x))}{\sum_{x\in\Omega} A(p(x))}$
  - q: is reference in $\mathcal{E}(\Omega)$
- Jeffreys Divergence, *Martín et al. 1996*: $\sqrt{\sum_{i=1}^{D}(log\ x_i - log\ y_i)(x_i - y_i)}$
- Hellinger distance, *V Pawlowsky et al. 2011*: $\sqrt{\sum_{i=1}^{D}\left(\sqrt{x_i} - \sqrt{y_i}\right)^2}$

5. Clustering Results on Simulation and Alimentation Data Set

		Divergences	Within sum of Squares	Calinski-Harabasz Index	Average Silhouette
Simulation	200 samples from the Dirichlet distribution with three compositional components Mathematically $\{\mathbf{x}_i\}_{i=1}^{200} \sim \mathrm{Dir}(\alpha_1, \alpha_2, \alpha_3)$	**Aitchison Distance**	0.543	79.86	0.19
		Kullback Leibler divergence	0.543	80.96	0.20
		Kaniadakis' divergence	**0.5**	**189.97**	**0.37**
		Hellinger divergence	0.59	164	0.35
		j-divergence	0.61	131	0.32
Alimentation Data	Data set contains the percentages of the consumption of several types of food during the 1980s of 25 European countries, grouped into 25 ethnic groups.	**Aitchison Distance**	0.45	21	0.34
		Kullback Leibler divergence	0.6	23	0.41
		Kaniadakis' divergence	**0.47**	**21**	**0.39**
		Hellinger divergence	0.36	17	0.21
		j-divergence	0.34	15	0.24

6. Compositional Clustering on Simplex



Results for simulation data are on the left, while results for Alimentation data are on the right(WM is White Meat, RM is Red Meat and F is Fish

7. Findings and Future Work

- Kaniadakis' divergence performs well in terms of WSS (lower is better), CHI, and AS, with higher values for simulation data. For real data, it also shows higher AS, indicating that Kaniadakis' divergence is effective for compositional clustering.
- Similarities were observed between the Aitchison Distance and Kullback-Leibler methods on simulation data, while the Hellinger and Jeffreys Distance produced different results.When applied to Alimentation data, the Aitchison Distance and Kullback-Leibler methods yield different results.
- These differences will be better investigated both by understanding the mathematics behind the clustering algorithms and the measures of efficiency, and by running further examples.
- Other divergences and measures of dissimilarities will be considered for compositional data, including data with zero components.