Kernel Methods for Radial Transformed Compositional Data with Many Zeros (ICML 2022)

Junyoung Park, Changwon Yoon, Cheolwoo Park, Jeongyoun Ahn

KAIST

Compositional Data with Many Zeros

- Recently, compositional data with a large proportion of zeros are prevalent in practice.
- E.g., microbiome data are compositional data, having **a significant portion** (about 50 – 80%) **of data are zeros**.
- In the literature, researchers perturb compositional data with zeros slightly so that they all have positive components (**zero replacement**) because the well-developed Aitchison's log-ratio methods are not available for data with zeros.

Main Contributions of Proposed Work

1. Point out a **geometric improperness** of “log-ratio transform after zero replacement” to compositional data with many zeros.
2. Interpret the domain of compositional data alternatively via **radial transformation**, and show that various kernel methods (kernel PCA, SVM, kernel mean embedding,...) can be successfully applied to these data.
3. The better performance of the proposed method than log-ratio methods is provided by experimental results on simulated and real data examples.

Flaws of Zero Replacements + Aitchison Geometry

Some reported flaws:

- There are countless ways to replace zeros but there is no standard.
- Data analysis results are often too sensitive to the choice of the zero replacement strategy.

Underlying premise of zero replacements:

It causes negligible alteration in the data

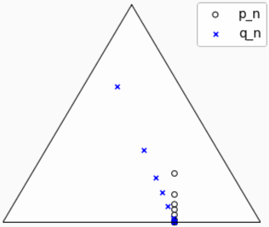
However, the Aitchison geometry tends to **amplify a tiny movement near the boundary** of the simplex, incompatible with the premise above.

Thus, the **combination**:

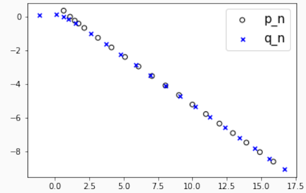
Zero replacements + Log-ratio transformation

does not work the way people want it to.

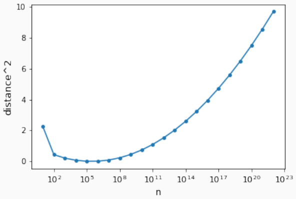
Example: Two Convergent Sequences Do Not Converge



- Two sequence converging to the same point on the boundary.
- We want that p_n and q_n should be almost the same for all large n.



$\mathrm{ilr}(p_n)$ and $\mathrm{ilr}(q_n)$ diverge.



The Aitchison distance $\|\mathrm{ilr}(p_n) - \mathrm{ilr}(q_n)\|_{\mathbb{R}^2}^2$ diverges.

Radial Transformation and Equivalence of Function Spaces

Radial transformation

$$\psi : \Delta^d \to \mathbb{S}^d_{\geq 0}, \qquad x \longmapsto \frac{x}{\|x\|_2}$$

Here, $\mathbb{S}^d_{\geq 0}$ denotes the nonnegative part of the hypersphere $\mathbb{S}^d \subset \mathbb{R}^{d+1}$.

Theorem (Equivalence of kernel mean embeddings)

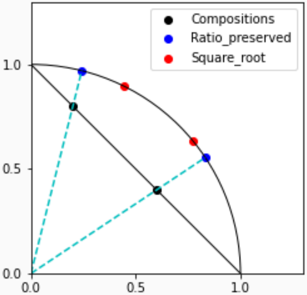
The following diagram

$$\begin{tikzcd}
\mathcal{P}(\Delta^d) \arrow[r] \arrow[d, "\psi_*"] & \mathcal{H}_{K \circ \psi} \\
\mathcal{P}(\mathbb{S}^d_{\geq 0}) \arrow[r] & \mathcal{H}_K \arrow[u, "\psi^*"]
\end{tikzcd}$$

is commutative where the horizontal maps are kernel mean embeddings.

The theorem establishes an equivalence between kernel-based data analysis between these two domains.

Sphere Should Be Better Than Simplex



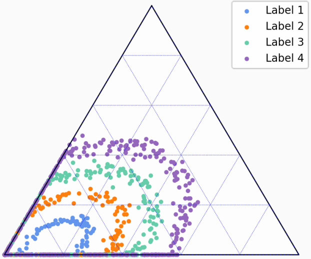
- Given a compositional data $x = (x_1, \ldots, x_{d+1}) \in \Delta^d$, it is clear that $x_i/x_j = cx_i/cx_j$ for all $c > 0$.
- It is natural to interpret compositional data as radial vectors!

1. There are a rich class of well-understood and easily computable kernels on hyperspheres with desirable decay of eigenvalues.
2. The non-smooth boundary of the simplex makes it hard to apply theoretical results of kernel methods on manifold data as those theory often assume smoothness of manifold.

Simulated Data

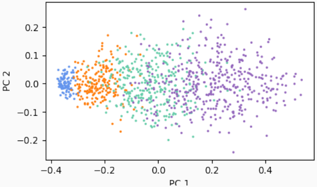
We generated high dimensional compositional data in Δ^d

(Zero proportion: about 40%).

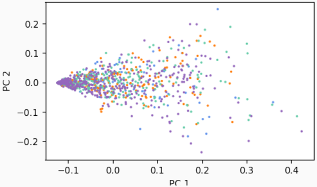


Visualization in case $d = 2$.

kPCA projection plots (rbf kernel):



(a) radial transform ($\gamma = 60$)



(b) clr transform ($\gamma = 0.005$),

replace zeros by $0.5x_{\text{min}}$

Conclusions

1. Data analysis based on

   $$\text{“zero-replacement} + \text{log-ratio transform”}$$

   might lose their justification as it distorts the original data significantly.
2. Kernel methods after the radial transform are successfully applied to compositional data with many zeros, showing better performance than the log-ratio transformations.

Please refer to our paper for more details, and more experimental results on the other datasets.

Thank you for listening