T-ARC: Topology-Aware Randomized Clustering via Distributionally Robust Optimization

Serena Grazia De Benedictis

Classical centroid-based clustering, such as K-means [1], relies on Euclidean distances and implicitly assumes convex, isotropic cluster shapes. This geometric bias leads to failure on non-linear or entangled structures, where topological connectivity rather than metric proximity defines the underlying classes.

We propose *T-ARC (Topology-Aware Randomized Clustering)*, a framework that integrates a graph-cut regularization term directly into the K-means objective to enforce topological awareness in the cluster assignment:

$$\min_{\substack{C\in\mathbb{R}_+^{n\times k},\ \mu\in\mathbb{R}_+^{k\times d}\\ L\in\mathcal{L}_{\mathcal{G}}(X)}} \frac{1}{2}\operatorname{Tr}(C^\top LC) + \frac{\lambda_\mu}{2}\|X - C\mu\|_F^2,$$

where C is the cluster assignment matrix, μ is the centroids matrix, and $L \in \mathcal{L}_{\mathcal{G}}(X)$ is the graph Laplacian associated with any admissible graph structure on the dataset X.

To overcome the combinatorial intractability of optimizing over all $\mathcal{L}_{\mathcal{G}}(X)$, we model the latent graph as a random realization from a Stochastic Block Model (SBM) [2] and optimize its parameters via a Distributionally Robust Optimization (DRO) framework [3], yielding closed-form proximal updates [4] for the SBM parameter b. The graph prior is grounded in topological data analysis [5]: the similarity matrix S driving the SBM is constructed from zero-dimensional persistent homology (H_0) [6], translating the multi-scale connectivity of the point cloud into a fixed topological prior.

The overall optimization proceeds via Block Coordinate Descent (BCD) [7]. While the updates for C and μ are deterministic and convex, the stochastic graph update requires a dedicated convergence analysis: we introduce a Lyapunov functional and prove a quadratic bound on its expected variation, which suffices to guarantee convergence to a stationary point.

Numerical experiments demonstrate that T-ARC recovers latent topological structures where baselines fail, achieving superior supervised accuracy while maintaining competitive unsupervised coherence. The persistence-based similarity matrix consistently outperforms an Euclidean-based variant, confirming the value of topological information as a structural prior.

This is a joint work with Andersen Ang, Nicoletta Del Buono, Flavia Esposito and Laura Selicato.

References

[1] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[2] C. Lee and D. J. Wilkinson, "A review of stochastic block models and extensions for graph clustering," *Applied Network Science*, vol. 4, no. 1, 2019.

[3] D. Kuhn, S. Shafiee, and W. Wiesemann, "Distributionally robust optimization," *Acta Numerica*, vol. 34, pp. 579–804, 2025.

[4] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[5] H. Schenck, *Algebraic Foundations for Applied Topology and Data Analysis*. Springer International Publishing, 2022.

[6] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[7] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.