Homogeneity testing in high dimensions with applications to graph sparsification and data clustering

Description: Testing whether a sample of observations is statistically homogeneous, is fundamental for assessing the complexity of the underlying distribution. Moreover, it can help in determining the model complexity (model selection) in machine learning applications. Nevertheless, it is unfortunate that, as most statistical hypothesis testing methods, homogeneity testing also suffers from being only effective in very few dimensions.

In this work we first aim to study hypothesis meta-tests, such as the one presented in [1], that devise ways to define and test a hypothesis over the result of multiple univariate unimodality tests [5]. The main idea is based on analyzing multiple histograms of pairwise similarities and then decide whether a high dimensional cloud of points forms one (null hypothesis, H_0) or more (alternative hypothesis, H_a).

We are specifically interested to measure the statistical power of such approaches, their scalability in the size of data, their adequacy to get easily recomputed (to get an updated result) when only small changes have taken place to the sample. Moreover, it is also interesting to try incorporating approaches related to histogram segmentation [3], k-modality testing, or other projection-based preprocessing. In terms of applications, we are planning to use such tests for kernel sparsification and/or data clustering.

Topic keywords: epidemics, social interactions and behavior, diffusion control

Indicative references:

[1] Kalogeratos, A. and Likas, A. (2012). "Dip-means: an incremental clustering method for estimating the number of clusters". NIPS.

[2] Tsapanos, N., Anastasios, T., Nikolaidis, N., and Pitas, I. (2015). "A distributed framework for trimmed kernel k-means clustering", Pattern recognition.

[3] Delon, J., Desolneux, A., Lisani, J.-L., and Petro, A.-B. (2007). "A non parametric approach for histogram segmentation". PAMI.

[4] Daskalakis, C., Diakonikolas, I., Servedio, R.A., Valiant, V., and Valiant, P. (2011). "Testing k-Modal Distributions: Optimal Algorithms via Reductions". arXiv preprint.

[5] Hartigan, J.A. and Hartigan, P. M. (1985). "The dip test of unimodality", Annals of Statistics.

[6] Siffer, A., Fouque, P.-A., Termier, A., and Largouët C. (2018). "Folding test. Are your data gathered?", KDD2018.

[7] Chronis P., Athanasiou, S., and Skiadopoulos, S. (2019). "Automatic clustering by detecting significant density dips in multiple dimensions", ICDM.