Identifying informative distance measures in high-dimensional feature spaces

Real-world data in biochemistry, material science and beyond typically contain a large number of features that are often heterogeneous in nature, relevance, and also units of measure. When assessing the similarity between data points, one can build various distance measures using subsets of these features. Finding a small set of features that still retains sufficient information about the dataset is important for the successful application of many statistical learning approaches.

We introduce a statistical test that can assess the relative information retained when using two different distance measures, and determine if they are equivalent, independent, or if one is more informative than the other. This test can be used to identify the most informative distance measure and, therefore, the most informative set of features, out of a pool of candidates. The approach can be used to identify the most appropriate set of collective variables in molecular systems and to infer causality in high-dimensional dynamic processes and time series.