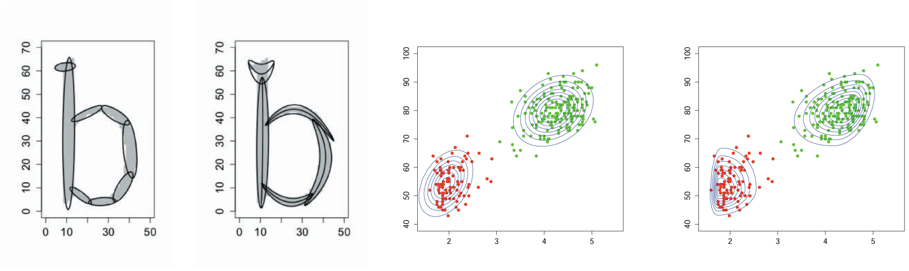
Clustering plays a basic role in many parts of data engineering, pattern recognition, and image analysis. One of the most important clustering algorithms is based on Gaussian Mixture Models. It is hard to overestimate the role of GMM in computer science; it include object detection, object tracking, learning and modelling, feature selection, classification, and statistical background subtraction.

GMM accommodates data with distributions that lie on affine subspaces of lower dimensions obtained by principal components (PCA). However, by the manifold hypothesis, real world data presented in high dimensional spaces are likely to concentrate in the vicinity of non-linear sub-manifolds of lower dimensionality. The classical approach approximates this manifold by a mixture of Gaussian distributions. Since one non-Gaussian component can be approximated by several Gaussian ones, these clusters are, in practice, represented by introducing more Gaussian components which can be seen as a form of piecewise linear approximation, see Fig. 1(a).



(a) GMM (b) afCEC (c) GMM (d) sgCEC

Figure 1: Fitting a b-type set by using (a), (c) GMM, (b) afCEC, (d) sgCEC.

Moreover, there still exist several problems in statistical modeling of normal mixture models. For instance, normality assumptions for component densities could be violated when a set of data contains asymmetric outcomes for each component. Moreover, the classical normal mixture model tends to over-fit the data since they need to include additional components to capture possibly excess skewness.

The growing need for more flexible tools to analyze datasets that exhibit non-normal features, including asymmetry, multimodality, and heavy tails, has led to intense development in non-normal model-based methods. The difference between classical method and our approach is presented in Fig. 1(c) and Fig. 1(d).

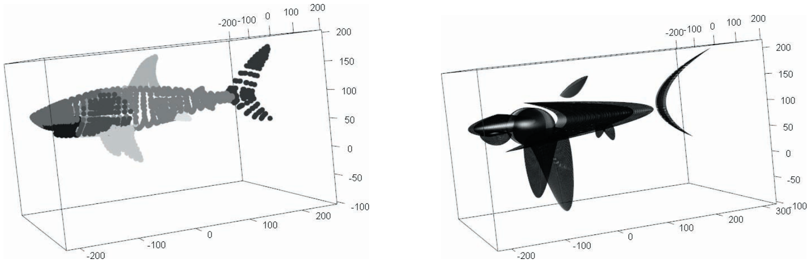


Figure 2: Result of afCEC algorithm in the case of a 3D shark-type set.

The main purpose of the grant is to construct a general afCEC (Active Function Cross-Entropy Clustering) and sgCEC (Split Gaussian Cross-Entropy Clustering) theory, which allows the clustering of data on sub-manifolds of $\mathbb{R}^d$ and datasets that exhibit non-normal features, including asymmetry, multimodality, and heavy tails. The motivation for this idea was a result of the observation that it is often profitable to describe non-linear (non-Gaussian) data by smaller numbers of components with more complicated shapes to obtain a better fit of data, see Fig. 1. The effect of afCEC in $\mathbb{R}^3$ on a shark-type set is shown in Fig. 2.