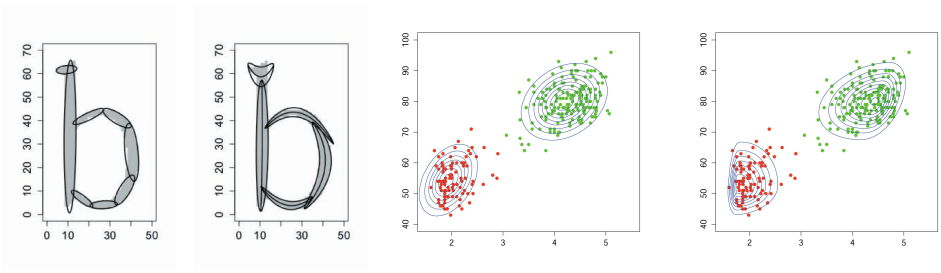
Analiza skupień to jedna z najważniejszych technik analizy danych, która znajduje szerokie zastosowanie w wielu dziedzinach nauki. Jedna z najpopularniejszych metod grupowania danych: Gaussian Mixture Models (GMM) zakłada, że dane koncentrują się na obszarach ograniczonych elipsami – poziomicami rozkładów normalnych, patrz Rysunek 1(a), 1(c). Podobny efekt dostajemy, przy użyciu metody Cross–Entropy Clustering (CEC).

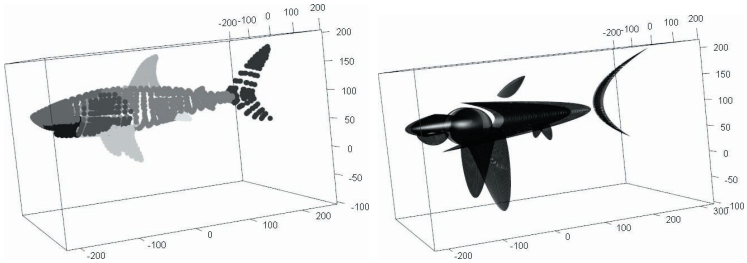
W rzeczywistości analizowane dane często posiadają bardzo skomplikowaną strukturę. Często koncentrują się one wzdłuż krzywych (w wyżej wymiarowych przestrzeniach wzdłuż rozmaitości). Jedną z metod pozwalającą odzwierciedlić takie struktury danych jest dokonanie modyfikacji rozkładów gaussowskich, tak by poziomice gęstości uogólnionych rozkładów normalnych były elipsami w krzywoliniowych układach współrzędnych. Pozwala to lepiej opisać dane za pomocą mniejszej ilości gęstości (patrz Rysunek 1(b)).



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

Rysunek 1: Efekt klasteryzacji za pomocą algorytmów: (a), (c) GMM, (b) afCEC i (d) sgCEC.

Ponadto, istnieje wiele problemów w modelowaniu danych za pomocą mieszaniny rozkładów normalnych. Na przykład założenie normalności może być niespełnione lub zaburzone, gdy zbiór danych zawiera asymetryczne komponenty. Co więcej mieszanina rozkładów normalnych wykazuje tendencję do nadmiernego dopasowania się do danych, ponieważ używa dodatkowych grup aby opisać ewentualną skośność/asymetrię komponentów.



Rysunek 2: Wynik działania algorytmu afCEC na przykładzie danych 3D.

Rosnące zapotrzebowanie na bardziej elastyczne narzędzia do analizy zbiorów danych wymusza użycie bardziej skomplikowanych modeli, które biorą pod uwagę skośność danych, cechują się ciężkimi ogonami lub wielomodalnością. W ostatnim czasie popularnością cieszą się mieszaniny rozkładów skośnych. Prosty przykład ukazujący różnicę między rozkładami klasycznymi i skośnymi jest zaprezentowany na Rysunkach 1(c), 1(d).

Głównym celem grantu jest zbudowanie algorytmów grupowania (klasteryzacji) danych za pomocą rozkładów nie gaussowskich. Umożliwi on opisanie danych, których wewnętrzna struktura przypomina krzywe (powierzchnie lub rozmaitości wyższych rzędów) lub cechuje się skośnością lub grubymi ogonami. Wyniki algorytmu na obiekcie 3D w kształcie rekina widzimy na Rysunku. 2.