Assessing the Statistical Significance of Clusters Using SigClust

Makena Grigsby and Sharvee Joshi

Contents

Abstract

The goal of this project is to study when a two–cluster structure in high–dimensional data is statistically meaningful versus simply an artifact of running a clustering algorithm. We focus on the SigClust procedure of Liu et al. and the soft–thresholding covariance estimation proposed by Huang et al. (2015), which tests H_0 : one Gaussian cluster against H_1 : more than one cluster. We replicate a subset of the simulation settings in Huang et al. using high–dimensional Gaussian data with a spiked covariance structure. First, we generate data from a single Gaussian distribution, apply k-means with $k = 2$, and use SigClust to test whether the resulting partition corresponds to real clusters. The p–value histogram is concentrated near one, indicating extremely conservative Type I error. Second, we simulate two Gaussian clusters with a large mean shift in a single coordinate; here SigClust almost always rejects, showing high power when the cluster separation is strong. Finally, we consider a dense but weak mean shift in all coordinates and find that, with our moderate dimension and noisy covariance, SigClust has limited power and the p–value distribution resembles the null. Overall, our experiments illustrate how SigClust responds to different signal–to–noise regimes in high–dimensional clustering.