ML classification as a means of peak deconvolution

Shadrach Kwakye-Nimo

July 4, 2023

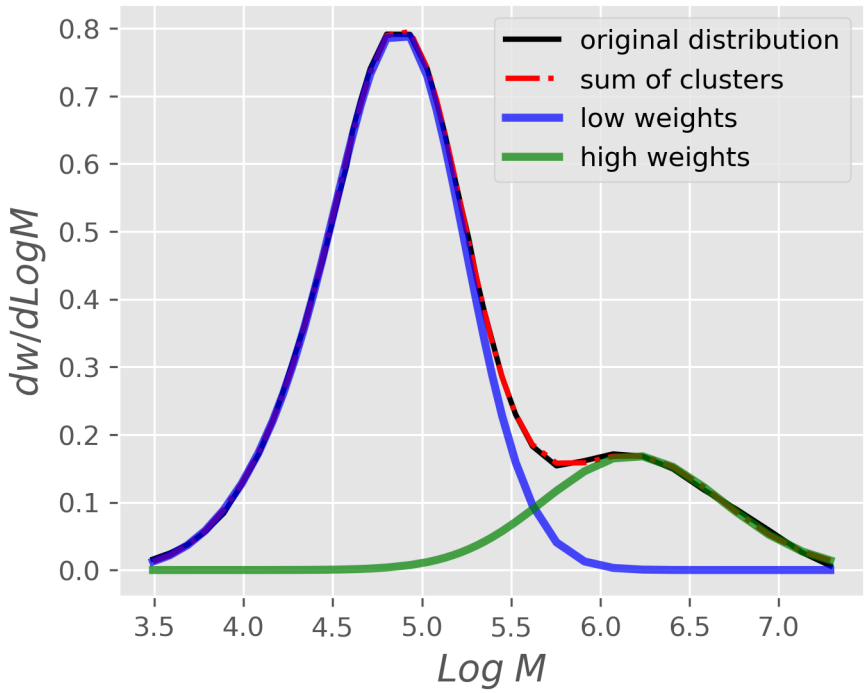

Figure 1: Peak deconvolution using ML

Abstract

This project makes use of machine learning to perform peak deconvolution. A classification algorithm is used to segment polymer chains distribution into subpopulations corresponding to their production site.

Motivation

The motivation for this project is to overcome the challenge a user faces during peak deconvolution, that is, the need to provide the expected peak locations for deconvolution to be done.

Key insight

By using machine learning to identify clusters in a distribution, peak deconvolution can easily be performed. This method most often leads to more reliable results.

Introduction

In polymer science the mechanical properties of a plastic material greatly depends on the characteristics of the two major subpopulations. These populations can be tweaked to produce materials of desirable properties, hence it is sometime required to perform peak deconvolution to fully account for the role played by each population. This has traditionally been achieved by making use commercial softwares. In this project we will using the expectation-maximization algorithm to identify the clusters and subsequently perform deconvolution.

Results and discussion

Choice of model

There are several models which could be used for classification such as K-Nearest Neighbors. Our choice of the Gaussian mixture model in this project was because, not only can it be used for classification purposes but also it can be used as a kernel estimator (see file on data preparation).

Optimizing the number of clusters

The number of clusters required to fully represent the distribution and their characteristics are unknown from the start. To obtain that, we looped over n of possible clusters

- perform classification
- use the cluster parameters to generate a gaussian distribution (ie the chains produce at a site follows a gaussian distribution)

- calculate the difference between the area under the curve of the original distribution and the sum of the distribution from the clusters [plotted in figure 2]

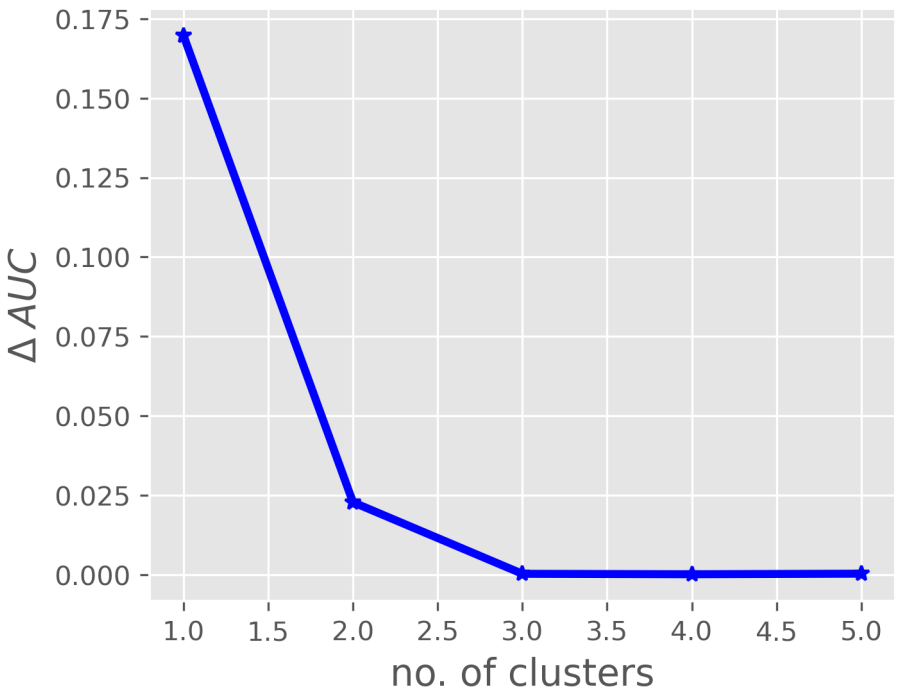

Figure 2: Calculation of error

From Fig.2, the error seems to be minimal when using 3 cluster, beyond which no further usefull information is obtained.

This can be observed from Fig.3, where the curve on the right has cluster 2 and 3 are all embedded in cluster 1

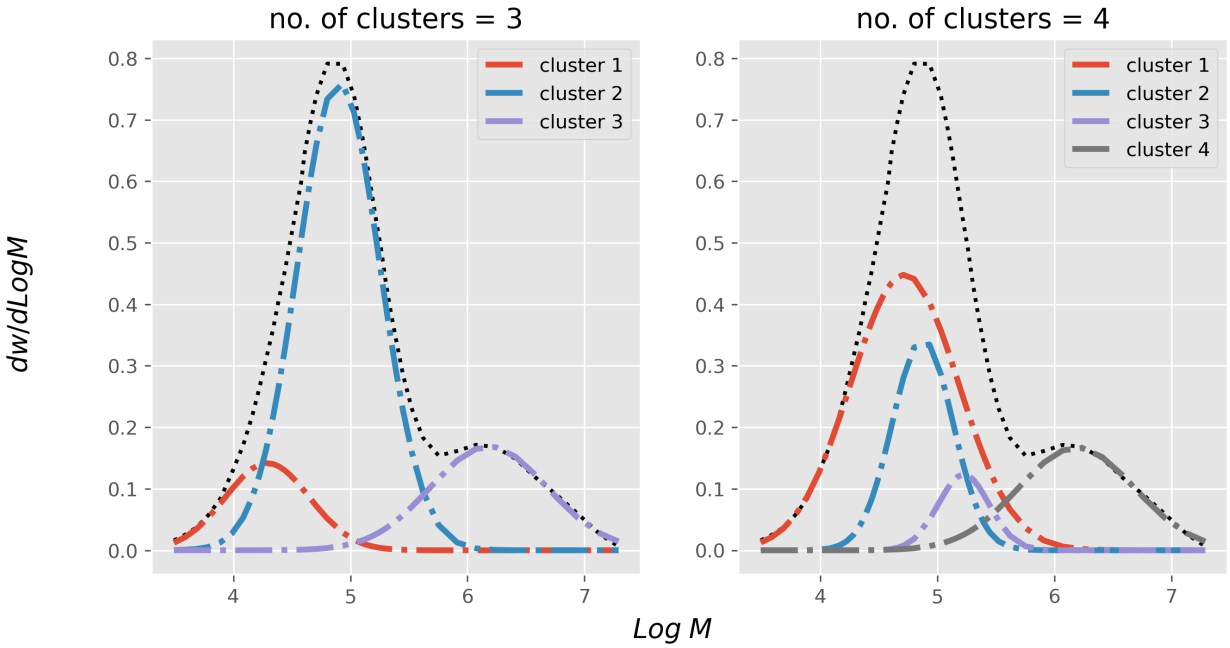

Figure 3: comparing output of 3 and 4 clusters