High Dimensional Density Estimation by HMM-VB for Clustering and Classification

WASP Mini-Course, May 2024

Jia Li, Penn State University, WASP Guest Professor at Umeå University

Course coordinator: Jun Yu, jun.yu@umu.se

1 Introduction

Statistical mixture modeling is a fundamental paradigm for unsupervised clustering. This course will cover the basic principles of mixture modeling, with a focus on Gaussian Mixture Models (GMM) and the Expectation-Maximization (EM) algorithm. Additionally, we will delve into the challenges associated with clustering in high-dimensional spaces. Specifically, we will address the issue of non-elliptical clusters by employing mode-based clustering. This approach utilizes an optimization algorithm known as Modal EM, which is designed to identify local maxima in the density function of a GMM. The challenge of effectively modeling high-dimensional data using GMM will be addressed through the Hidden Markov Model on Variable Blocks (HMM-VB). Moreover, the concept of clustering by mode association will be expanded from GMM to HMM-VB by extending Modal EM to Modal Baum-Welch algorithm.

While the primary focus of this course is on clustering, the models and algorithms introduced have wider applications. Both GMM and HMM-VB serve as versatile tools for density estimation, which can facilitate a range of downstream analyses, including classification tasks. The EM algorithm is extensively applied in estimating probabilistic graphical models that incorporate latent states. Furthermore, the Modal EM algorithm, alongside the related Ridgeline EM algorithm, provides valuable insights into the geometric characteristics of high-dimensional data. These algorithms are particularly useful for visualizing the separation between clusters, regardless of the original dimension, by one-dimensional curves.

We plan to cover the following topics.

1. General background on Gaussian mixture model (GMM) and EM algorithm.

2. Modal EM algorithm with applications to clustering by the principle of mode association and Ridgeline EM algorithm for visualizing the separation between clusters.

3. Hidden Markov model on variable blocks (HMM-VB): model formulation, Baum-Welch algorithm for estimation, and Modal Baum-Welch algorithm for finding modes of the density.

4. An introduction to the R CRAN package: HDclust. If time allows, a brief introduction to variable selection for clustering.

2 Prerequisites

Students are expected to have a basic knowledge of linear algebra, and courses in probability theory and mathematical statistics at second cycle level.

3 Schedule

The preliminary time is 20-22 May.

1. Day 1
   (a) Class 1: Topic 1
   (b) Class 2: Topic 2
2. Day 2
   (a) Class 1: Topic 3
   (b) Class 2: Topic 4

4 Assignment

1. Select at least one related reference to read in detail. Recommended papers are listed below, but not limited to those.
   (a) J. Li, S. Ray, B. G. Lindsay, "A nonparametric statistical approach to clustering via mode identification," Journal of Machine Learning Research, 8(8):1687-1723, 2007.
   (b) L. Lin and J. Li, "Clustering with hidden Markov model on variable blocks," Journal of Machine Learning Research, 18(110):1-49, 2017.
   (c) B. Seo, L. Lin, J. Li, "Block-wise variable selection for clustering via latent states of mixture models," Journal of Computational and Graphical Statistics, 31(1):138-150, 2021.
2. Experiment with the R package HDclust.
   (a) Learn the basics about how to use the package.
   (b) Use HMM-VB as a density estimation method to perform classification using a generative modeling approach. Specifically, the Bayes formula is applied to compute the class posterior. Compare the result, both in terms of classification result and the posteriors, with popular black-box classifiers such as random forest (RF), gradient boosting tree, and multi-layer perceptron (MLP) neural networks.
   (c) Use HMM-VB for clustering, either using the default mode-association clustering provided by the package, or develop your own schemes based on the Viterbi state sequences identified for each data point.
   (d) Write a report on the analysis including some background on the datasets used, methods experimented with, comparison of results, etc.
3. Expect to submit in one week.