Double sparsity in high-dimensional Gaussian mixture estimation and clustering
Subject Overview

Laboratory Supervisor: A.S. Dalalyan
PHd Student: M. Sebbar

March 3, 2015

Contents

1 Introduction

The broad goal of this thesis is to tackle a clustering problem in the scope of mixtures model framework. More precisely, we will study the clustering of points drawn from high-dimensional Gaussian mixtures distributions.
Thus, in the first part of this section we study the gaussian mixture model and the second part we describe the well know algorithm Expectation-Maximization (EM) and the limitations in high-dimensional setting.

1.1 The Gaussian mixture model

The Gaussian mixture model is an important framework where the components are Gaussian distributions with parameters (μ_i, Σ_i). We obtain the following distribution:

$$p(x|\theta) = \sum_{i=1}^{K} \pi_i \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} = \sum_{i=1}^{K} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$$

with $\theta = \{\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K\}$ and $\forall i, \pi_i > 0$ and $\sum_{i=1}^{K} = 1$

In the clustering problem, we would like to calculate the probability of the latent variable Z conditioned on X in order to assign X to a cluster.
We denote $\tau_k = P(z_k = 1|x, \theta)$, from Bayes's rule we have:

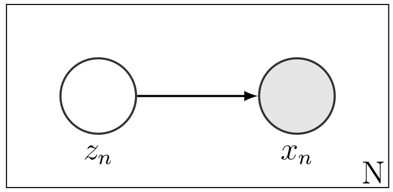
$$\tau_k = \frac{P(x|z_k = 1, \theta)P(z_k = 1)}{P(x)} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} \mathcal{N}(x|\mu_i, \Sigma_i)}$$

where $\pi_i = P(z_i = 1)$ the prior probability and τ_i the posterior.

We would like to estimate θ from a set of iid observations $X_1, \ldots, X_N$. The related graphical model is:

The log-likelihood is:

$$l(\theta|D) = \sum_{n=1} N \log p(x_n|\theta)$$



Here we have the log of a sum (contrary to exponential family distribution where the log acts on a simple probability distribution) and the maximization of the log-likelihood is a non-linear problem.
An approach for the estimation of the maximum of log-likelihood is the Expectation-Maximization Algorithm.

1.2 The EM Algorithm

We will infer the values of $\{z_n\}$ conditioned to the data $\{x_n\}$. A natural approach to estimate the parameters θ is to estimate the mean of each class by deriving the log-likelihood:

$$\widehat{\mu}_i = \frac{\sum_{n=1}^{N} \tau_n^i x_n}{\sum_{n=1}^{N} \tau_n^i}$$

However, as seen in ?, τ_n^i depends on the parameter estimates which depends on τ_n^i. An idea would be to initialize the parameters and iterate. We calculate the posterior probability and then estimate the parameter θ. This is the idea of the EM algorithm.

The EM algorithm for Gaussian Mixtures would be:

0. Init parameters

1. Calculate (Expectation Step): $\tau_n^i(t+1)$

2. Calculate (Maximization Step):

    - $\mu_i(t+1) =$
    - $\Sigma_i(t+1) =$

- $\pi_i(t+1) =$

#Explain why complicated, pro and cons with p large

2 A structural analysis on Σ approach

We consider a multivariate Gaussian distribution with mean $\boldsymbol{\mu}^*$ and covariance $\boldsymbol{\Sigma}^*$ and $Y_1, \dots, Y_N \in \mathbb{R}^p$ iid drawn from this distribution. We would like to estimate $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$. We know that $\widehat{\boldsymbol{\mu}}_n = \bar{Y}_n$, then WLOG we consider $\mu^* = 0$, the problem is to estimate $\boldsymbol{\Sigma}^*$. We will study the precision matrix and consider that Σ^{-1} is sparse. We note $\Sigma^{-1} = \Omega$.
If $\Sigma^{-1}_{ij} = 0 \Rightarrow Y_i \perp\!\!\!\perp Y_j$ conditionnaly to $Y_{l \neq \{i,j\}}$. Thus, it makes sense to impose a L_1 penalty on Σ^{-1} to increase its sparsity.

2.1 Graphical Lasso

$$\mathcal{N}(x|\mu^*, \Sigma^*) = \frac{1}{(2\pi)^{d/2}|\Sigma^*|^{1/2}} \exp^{-\frac{1}{2}(x-\mu^*)^T \Sigma^{-1*}(x-\mu^*)}$$

The log-likelihood, with $\mu = 0$ is given by:

$$\mathcal{L}(\Sigma) = \log\left(\prod_{n=1}^{N} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp^{-\frac{1}{2}(x_n)^T \Sigma^{-1}(x_n)}\right)$$

write eqs

$$L(\Sigma) = C + \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}tr(S_n \Sigma^{-1})$$

Thus, considering the sparsity of Ω, we impose a penalization to the maximum likelihood estimator of Σ^{-1}

$$\widehat{\Omega} \in argmin\left\{\log(|\Omega|) - tr(S_n\Omega) - \lambda||\Omega||_1\right\}$$

A reason to use the L_1 penalization instead of the ridge is that for an L_p penalization, the prpblem is convex for $p \geq 1$ and we have parsimonious property for $p \leq 1$.

2.2 Column-Wise Lasso

3 Comments

4 References