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

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

March 22, 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). The distribution is given by:

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

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

We will study this mixture from a latent variable perspective. Let Z be a multionomial random variable with component Z^i, we have:

$$p(x|\theta) = \sum_{i=1}^{K} p(Z^i = 1)p(x|Z^i = 1, \theta) \quad (2)$$

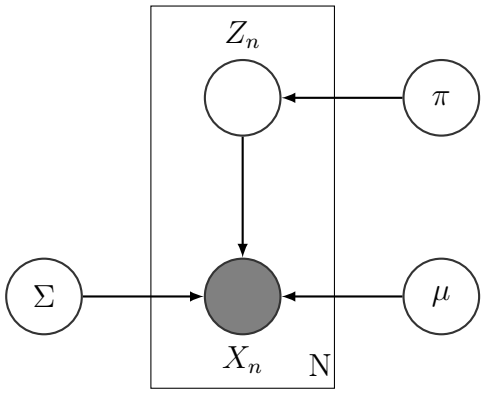
Therefore $\pi_i = p(Z^i = 1)$ and reflects the probability that x is drawn from the i^{th} mixture component.

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:

$$\tau^k = \frac{p(x|Z^k = 1, \theta)p(Z^k = 1)}{p(x|\theta)} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)} \quad (3)$$

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 $X_1, \ldots, X_N$ with $X_i \in \mathbb{R}^d$. The related graphical model is:



The log-likelihood is:

$$l(\theta|D) = \sum_{n=1}^{N} \log p(x_n|\theta) = \sum_{n=1}^{N} \log \sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i) \tag{4}$$

Where D is the set of data points. Here we have the log of a sum and the maximization of the log-likelihood is a non-linear problem (contrary to exponential family distributions where the log acts on a simple probability distribution and therefore yields simple expressions).
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 with respect to μ_k and setting to 0 we have:

$$\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)} \Sigma_k (x_n - \mu_k) = \sum_{n=1}^{N} \tau_n^k \Sigma_k (x_n - \mu_k) = 0 \tag{5}$$

Assuming that Σ_k is non singular, we have:

$$\mu_k = \frac{\sum_{n=1}^{N} \tau_n^k x_n}{\sum_{n=1}^{N} \tau_n^k} \tag{6}$$

Doing a similar calculus for Σ_k, we have:

$$\Sigma_k = \frac{\sum_{n=1}^{N} \tau_n^k (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \tau_n^k} \tag{7}$$

Finally, maximizing the log-likelihood with respect to π_k with the condition $\sum_{k=1}^{K} \pi_k = 1$ (using Lagrange multiplier), we have:

$$\pi_k = \frac{\sum_{n=1}^{N} \tau_n^k}{N} \tag{8}$$

However, as seen in (3), τ_n^k depends on the parameter estimates which depends on τ_n^k. 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.

1.2.1 The EM algorithm for Gaussian Mixtures:

0. Initialize parameters $\mu_k^0,\ \Sigma_k^0,\ \pi_k^0$

1. Calculate (Expectation Step):

$$\tau_n^{k,(t+1)} = \frac{\pi_k^{(t)} \mathcal{N}(x_n|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{i=1}^{K} \pi_i^{(t)} \mathcal{N}(x_n|\mu_i^{(t)}, \Sigma_i^{(t)})} \tag{9}$$

2. Calculate (Maximization Step):

$$\begin{cases} \mu_k^{(t+1)} & = \frac{\sum_{n=1}^N \tau_n^{k,(t+1)} x_n}{\sum_{n=1}^N \tau_n^{k,(t+1)}} \\ \Sigma_k^{(t+1)} & = \frac{\sum_{n=1}^N \tau_n^{k,(t+1)} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_{n=1}^N \tau_n^{(t+1)}} \\ \pi_k^{(t+1)} & = \frac{\sum_{n=1}^N \tau_n^{(t+1)}}{N} \end{cases} \tag{10}$$

3. Evaluate the log-likelihood and check for convergence

1.2.2 Limitations of the EM Algorithm:

EM is conceptually easy and each iteration improves $l(\theta)$. However, the complexity of EM algorithm is $O(dn + Kn^2)$ and it requires many iterations. Unfortunately, in our case, the algorithm slows down in the high dimensional setting. We hope to tackle this problem by making sparsity assumptions on the structure of the precision matrix Σ^{-1}.

2 Draft-A structural analysis on Σ approach

We consider a multivariate Gaussian distribution with mean $\boldsymbol{\mu}^*$ and covariance $\boldsymbol{\Sigma}^*$ and $Y_1, \ldots, Y_N \in \mathbb{R}^p$ 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 $\Sigma^{-1} = \Omega$, Y_n the n-th random variable and Y_n^i the i-th component of this vector.
If $\Sigma_{ij}^{-1} = 0 \Rightarrow Y^i \perp\!\!\!\perp Y^j$ conditionally 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

Let consider a multivariate normal distribution with parameters μ^*, Σ^* with density;

$$\mathcal{N}(x|\mu^*, \Sigma^*) = \frac{1}{(2\pi)^{d/2}|\Sigma^*|^{1/2}} \exp^{-\frac{1}{2}(x-\mu^*)^T \Sigma^{-1*}(x-\mu^*)} \tag{11}$$

We consider $\mu = 0$. Given N datapoints $X_1, \ldots, X_N$ and $X_i \in \mathbb{R}^d$, the log-likelihood is given by:

$$\begin{aligned}\mathcal{L}(\Sigma) &= \log\left(\prod_{n=1}^{N} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp^{-\frac{1}{2}(x_n)^T\Sigma^{-1}(x_n)}\right) \\ &= -\frac{dN}{2}\log 2\pi - \frac{N}{2}\sum_{n=1}^{N}\log|\Sigma^*| - \frac{1}{2}\sum_{n=1}^{N} x_n^T\Sigma^{*,-1}x_n\end{aligned} \tag{12}$$

Note that $x_n^T\Sigma^{*,-1}x_n = tr(x_n^T\Sigma^{*,-1}x_n)$, and therefore:

$$\sum_{n=1}^{N} x_n^T\Sigma^{*,-1}x_n = tr\big(\sum_{n=1}^{N} x_n^T\Sigma^{*,-1}x_n\big) = tr\Big(\big[\sum_{n=1}^{N} x_n^T x_n\big]\Sigma^{*,-1}\Big) = tr(S_N\Sigma^*) \tag{13}$$

Where S_N is the empirical covariance matrix. We can replace that in the log-likelihood expression:

$$\mathcal{L}(\Sigma) = -\frac{dN}{2}\log 2\pi - \frac{N}{2}\sum_{n=1}^{N}\log|\Sigma^*| - \frac{1}{2}tr(S_N\Sigma^*) \tag{14}$$

Finally:

$$\mathcal{L}(\Sigma) = C + \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}tr(S_n\Sigma^{-1}) \tag{15}$$

Where C is a constant (dependent on N). Thus, considering the sparsity of the precision matrix $\Omega = \Sigma^{-1}$, we impose a penalization to the maximum likelihood estimator of Ω

$$\widehat{\Omega} \in argmin\big\{\log|\Omega| - tr(S_N\Omega) - \lambda||\Omega||_1\big\} \tag{16}$$

A reason to use the L_1 penalization instead of the ridge is that for an L_p penalization, the problem is convex for $p \geq 1$ and we have parsimonious property for $p \leq 1$.
This is a convex optimization problem, however the complexity is $O(p^3)$ (Source, high dim & var select Buhlmann 2006 ? Wassermann)

2.2 Column-Wise Lasso

We consider a gaussian vector $Y \in \mathbb{R}^d$, $Y \sim \mathcal{N}(0, \Sigma)$. We can write $Y = (Y^1, Y^{2:d})$. With this decomposition we can write the covariance matrix as following:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \tag{17}$$

and according to theorem[?]: If Σ_{22} is inversible, then:

$$\begin{array}{rcl} \mathbb{E}[Y^1|Y^{2:d}] & = & \Sigma_{12}\Sigma_{22}^{-1}Y^{2:d} \\ Var[Y^1|Y^{2:d}] & = & \sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \end{array} \tag{18}$$

We have the following identity:

$$\begin{pmatrix} \omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_{p-1} \end{pmatrix} \tag{19}$$

Which gives the following equations:

$$\begin{cases} \omega_{11}\sigma_1^2 + \Omega_{12}\Sigma_{12}^T & = 1 \qquad (*) \\ \omega_{11}\Sigma_{12} + \Omega_{12}\Sigma_{22} & = 0 \qquad (**) \\ \Omega_{12}^T\Sigma_{12} + \Omega_{22}\Sigma_{22} & = I_{p-1} \quad (***) \end{cases} \tag{20}$$

With (**) we have $-\omega_{11}\Sigma_{12}\Sigma_{22}^{-1} = \Omega_{12}$ and injected to (*) we have:

$$\begin{cases} \mathbb{E}[Y^1|Y^{2:d}] & = -\frac{1}{\omega_{11}}\Omega_{12}Y^{2:d} \\ Var[Y^1|Y^{2:d}] & = \frac{1}{\omega_{11}} \end{cases} \tag{21}$$

Finally, $Y^1 - \mathbb{E}[Y^1|Y^{2:d}]$ is a gaussian vector of $\mathbb{R}^{d-1}$, centered, independent of $Y^{2:d}$ and of covariance matrix $\sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$. If we denote $\xi^1 \sim \mathcal{N}(0,1)$ we have $Y^1 - \mathbb{E}[Y^1|Y^{2:d}] = \frac{1}{\sqrt{\omega_{11}}}\xi^1$.

Therefore, for $Y_1, \ldots, Y_n$ iid of law $\mathcal{N}(0, \Sigma^*)$ we have:

$$\begin{array}{rcl} Y_i^1 & = & -\frac{1}{\omega_{11}^*}\Omega_{12}Y_i^{2:d} + \frac{1}{\sqrt{\omega_{11}^*}}\xi_i^1 \\ & = & -\sum_{j=2}^d \frac{w_{ij}^*}{\omega_{11}^*}Y_i^j + \frac{1}{\sqrt{\omega_{11}^*}}\xi_i^1 \end{array} \tag{22}$$

and

$$\beta_1^{*T} Y_i = \frac{1}{\sqrt{\omega_{11}^*}} \xi_i^1 \Rightarrow \beta_1^{*T} \boldsymbol{Y} = \frac{1}{\sqrt{\omega_{11}^*}} \boldsymbol{\xi}^1 \tag{23}$$

with

$$\beta_1^* = \frac{1}{\sqrt{\omega_{11}^*}} \begin{bmatrix} w_{11}^* \\ w_{12} \\ \vdots \\ w_{1d} \end{bmatrix} \in \mathbb{R}^d \quad \text{and} \quad \boldsymbol{Y} = \begin{bmatrix} verifier \end{bmatrix} \tag{24}$$

2.3 The Square-Root Lasso

3 Comments

4 References