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

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

March 25, 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 present the Gaussian mixture model and the second part we describe the well know Expectation-Maximization algorithm (EM). We will also present the limitations of this algorithm in high-dimensional setting.

1.1 The Gaussian mixture model

The Gaussian mixture model is an important framework for clustering problems. The components are Gaussian distributions with parameters (μ_i, Σ_i):

$$\varphi_{(\mu_i,\Sigma_i)}(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)) \tag{1}$$

The density is given by:

$$p_\theta(x) = \sum_{k=1}^{K} \pi_k \varphi_{(\mu_k,\Sigma_k)}(x) \tag{2}$$

with $\theta = \{(\mu_k, \Sigma_k, \pi_k) : 1 \leq k \leq K;\ \Sigma_k \geq 0;\ \pi_k > 0\ \forall k\ and\ \sum_{k=1}^{K} \pi_k = 1\}$

This model can be interpreted from a latent variable perspective. Let Z be a multionomial random variable of parameters $\mu(1, \boldsymbol{\pi})$, we have:

$$p_\theta(x) = \sum_{k=1}^{K} P_\theta(Z^k = 1) p_\theta(x|Z^k = 1) \tag{3}$$

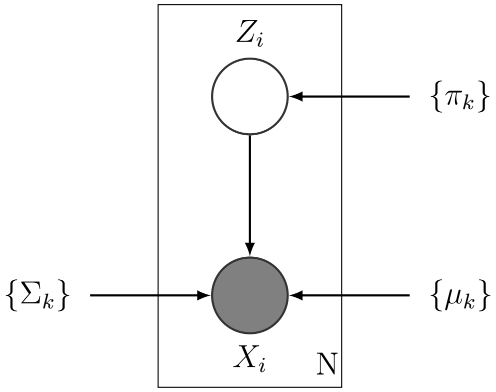
Therefore $\pi_k = p_\theta(Z^k = 1)$ and reflects the probability that x is drawn from the i^{th} mixture component.

In the clustering problem, the goal is to assign X to a cluster and therefore estimate the probability of the latent variable Z conditioned on X which we denote $\tau^k(x, \theta) = p_\theta(Z^k = 1|x)$. From Bayes's rule we have:

$$\tau^k(x,\theta) = \frac{p_\theta(x|Z^k=1)P_\theta(Z^k=1)}{p_\theta(x)} = \frac{\pi_k\varphi_{(\mu_k,\Sigma_k)}(x)}{\sum_{k'=1}^{K}\pi_{k'}\varphi_{(\mu_{k'},\Sigma_{k'})}(x)} \tag{4}$$

The of estimation $\tau^k(x,\theta)$ rely on computing the parameters θ. An approach to this problem is to estimate the parameter that maximize the likelihood.

Let $X_1,\dots,X_N$ with $X_i \in \mathbb{R}^p$ a set of iid observations drawn from p_θ. The related graphical model is:



The log-likelihood is:

$$l(\theta|x_1,\dots,x_N) = \sum_{i=1}^{N}\log p_\theta(x_i) = \sum_{i=1}^{N}\log\Big(\sum_{k=1}^{K}\pi_k\varphi_{(\mu_k,\Sigma_k)}(x_i)\Big) \tag{5}$$

In this equation, we have the logarithm of a sum and the maximization of the log-likelihood is a non-linear problem. (This is not a exponential family distribution where the log acts on a simple probability distribution and therefore yields simple expressions).
A well known approach for computing the maximum of log-likelihood is the Expectation-Maximization Algorithm [2].

1.2 The EM Algorithm

We will infer the values of $\{z_n\}$ conditioned to the data $\{x_n\}$. A natural approach to infer the parameters θ is to estimate the mean of each class by computing the partial derivative of the log-likelihood with respect to μ_k and setting to 0. We have:

$$\sum_{i=1}^{N} \frac{\pi_k \varphi_{(\mu_k,\Sigma_k)}(x_i)}{\sum_{k'=1}^{K} \pi_{k'} \varphi_{(\mu_{k'},\Sigma_{k'})}(x_i)} \Sigma_k^{-1}(x_i - \mu_k) = \sum_{i=1}^{N} \tau_i^k \Sigma_k^{-1}(x_i - \mu_k) = 0 \tag{6}$$

Assuming that Σ_k is non singular, we have:

$$\mu_k = \frac{\sum_{n=1}^{N} \tau_n^k x_n}{\sum_{n=1}^{N} \tau_n^k} \tag{7}$$

Doing a similar computation for Σ_k, we have:

$$\Sigma_k = \frac{\sum_{n=1}^{N} \tau_n^k (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \tau_n^k} \tag{8}$$

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:

$$\pi_k = \frac{\sum_{n=1}^{N} \tau_n^k}{N} \tag{9}$$

However, as seen in (4), τ_n^k depends on the parameter estimatation $\widehat{\theta}$ which depends on τ_n^k as seen in (7),(8) and (9). 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 μ_k^0, Σ_k^0, π_k^0

1. Calculate (Expectation Step):

$$\tau_i^k(t+1) = \frac{\pi_k(t)\varphi_{(\mu_k(t),\Sigma_k(t))}(x_i)}{\sum_{k'=1}^K \pi_{k'}(t)\varphi_{(\mu_{k'},\Sigma_{k'})}(x_i)} \tag{10}$$

2. Calculate (Maximization Step):

   We note: $\rho_i^k(t+1) = \frac{\tau_i^k(t+1)}{\sum_{j=1}^N \tau_j^k(t+1)}$, hence:

$$\mu_k(t+1) = \sum_{i=1}^N \rho_i^k(t+1)x_i \tag{11}$$

$$\Sigma_k(t+1) = \sum_{i=1}^N \rho_i^k(t+1)(x_i - \mu_k(t+1))(x_i - \mu_k(t+1))^T \tag{12}$$

$$\pi_k(t+1) = \frac{\sum_{i=1}^N \tau_i^k(t+1)}{N} \tag{13}$$

3. Evaluate the log-likelihood and check for convergence as explained in the following section

1.2.2 An analysis of the EM Algorithm convergence

As defined previously, we denote all observed variables $\boldsymbol{X}$ and discrete variables $\boldsymbol{Z}$. Our goal is to maximize the likelihood function[1]:

$$l(\theta|\boldsymbol{X}) = log(p_\theta(\boldsymbol{X})) = log\Big(\sum_{\boldsymbol{Z}} p_\theta(\boldsymbol{X},\boldsymbol{Z})\Big) \tag{14}$$

We introduce a distribution $q(\boldsymbol{Z})$ over the latent variables. We have:

$$l(\theta|\boldsymbol{X}) = \mathcal{L}(q,\theta) + KL(q||p_\theta) \tag{15}$$

with:

$$\mathcal{L}(q,\theta) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log\Big(\frac{p_\theta(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\Big) \tag{16}$$

$$KL(q||p_\theta) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log\Big(\frac{p_\theta(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})}\Big) \tag{17}$$

To verify this identity, we will start by decomposing $\mathcal{L}(q, \theta)$:

$$\begin{aligned}
\mathcal{L}(q,\theta) &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log\Big(\frac{p_\theta(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\Big) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log\Big(\frac{p_\theta(\boldsymbol{X})p_\theta(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})}\Big) \quad (18)\\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log\Big(\frac{p_\theta(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})}\Big) + \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log p_\theta(\boldsymbol{X}) \quad (19)\\
&= -KL(q||p_\theta) + l(\theta|\boldsymbol{X}) \quad (20)
\end{aligned}$$

$KL(q||p_\theta)$ is the Kullback-Leibler divergence between $q(\boldsymbol{Z})$ and $p_\theta(\boldsymbol{Z}|\boldsymbol{X})$. It satisfies $KL(q||p_\theta) \geq 0$. Therefore, $\mathcal{L}(q,\theta) \leq l(\theta|\boldsymbol{X})$. With the decomposition (15) we will prove that EM maximize the log-likelihood.

If we observe the E-Step of EM, we hold $\theta^{old} = \{(\mu_k(t), \Sigma_k(t), \pi_k(t))\}$ and optimize the lower bound $\mathcal{L}(q, \theta^{old})$ with respect to q. Because $l(\theta^{old}|\boldsymbol{X})$ does not depend on $q(\boldsymbol{Z})$, we need to decrease $KL(q||p_{\theta^{old}})$ which disappear when $q^{new}(\boldsymbol{Z}) = p_{\theta^{old}}(\boldsymbol{Z}|\boldsymbol{X})$ as seen in (10). At this Step:

$$\mathcal{L}(q^{new}, \theta^{old}) = l(\theta^{old}|\boldsymbol{X}) \quad (21)$$

In the M-Step, $q^{new}(\boldsymbol{Z})$ is fixed and we optimize $\mathcal{L}(q^{new}, \theta)$ with respect to θ. $l(\theta|\boldsymbol{X})$ depends on θ and will increase, fixing a new likelihood. And by optimizing θ, $q(\boldsymbol{Z})$ is no longer equal to $p_{\theta^{new}}(\boldsymbol{Z}|\boldsymbol{X})$ and we have a nonzero Kullback-Leibler divergence. Therefore, $l(\theta|\boldsymbol{X})$ will increase greater than $\mathcal{L}(q^{new}, \theta)$

If we decompose $\mathcal{L}(q, \theta)$:

$$\begin{aligned}
\mathcal{L}(q^{new},\theta) &= \sum_{\boldsymbol{Z}} p_{\theta^{old}}(\boldsymbol{Z}|\boldsymbol{X}) \log\Big(\frac{p_\theta(\boldsymbol{X},\boldsymbol{Z})}{p_{\theta^{old}}(\boldsymbol{Z}|\boldsymbol{X})}\Big) \quad (22)\\
&= \sum_{\boldsymbol{Z}} p_{\theta^{old}}(\boldsymbol{Z}|\boldsymbol{X}) \log p_\theta(\boldsymbol{Z}|\boldsymbol{X}) - \sum_{\boldsymbol{Z}} p_{\theta^{old}}(\boldsymbol{Z}|\boldsymbol{X}) \log p_{\theta^{old}}(\boldsymbol{Z}|\boldsymbol{X}) \\
&\quad (23)\\
&= \mathcal{Q}(\theta, \theta^{old}) + H(q) \quad (24)
\end{aligned}$$

$\mathcal{Q}(\theta, \theta^{old})$ is the expectation of the complete log-likelihood and $H(q)$ is the entropy of q. Therefore, the M-Step is equivalent to maximize $\mathcal{Q}(\theta, \theta^{old})$ with respect to θ:

$$\theta^{new} = \underset{\theta}{\operatorname{argmax}}(\mathcal{Q}(\theta, \theta^{old})) \tag{25}$$

A variant of EM, the Generalized EM (GEM) is based on the idea that it suffices to increase Q and not necessarily find the maximum.(ref)

As we saw in this section, each steps increase the log-likelihood and yields to a local maximum. By running EM over different initial parameters, we improve the estimation by choosing the best local maximum found.

1.2.3 Stopping criterion of EM

.

1.2.4 Limitations of the EM Algorithm:

EM is conceptually easy and each iteration increases $l(\theta)$. However, the complexity at each step 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;

$$\mathcal{N}(x|\mu^*, \Sigma^*) = \frac{1}{(2\pi)^{d/2}|\Sigma^*|^{1/2}} \exp^{-\frac{1}{2}(x-\mu^*)^T\Sigma^{-1*}(x-\mu^*)} \tag{26}$$

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:

$$\begin{aligned}\mathcal{L}(\Sigma) &= \log\left(\prod_{n=1}^{N} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp^{-\frac{1}{2}(x_n)^T\Sigma^{-1}(x_n)}\right) \\ &= -\frac{dN}{2}\log 2\pi - \frac{N}{2}\sum_{n=1}^{N}\log|\Sigma^*| - \frac{1}{2}\sum_{n=1}^{N} x_n^T\Sigma^{*,-1}x_n\end{aligned} \tag{27}$$

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

$$\sum_{n=1}^{N} x_n^T\Sigma^{*,-1}x_n = tr\big(\sum_{n=1}^{N} x_n^T\Sigma^{*,-1}x_n\big) = tr\Big(\big[\sum_{n=1}^{N} x_n^T x_n\big]\Sigma^{*,-1}\Big) = tr(S_N\Sigma^*) \tag{28}$$

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

$$\mathcal{L}(\Sigma) = -\frac{dN}{2}\log 2\pi - \frac{N}{2}\sum_{n=1}^{N}\log|\Sigma^*| - \frac{1}{2}tr(S_N\Sigma^*) \tag{29}$$

Finally:

$$\mathcal{L}(\Sigma) = C + \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}tr(S_n\Sigma^{-1}) \tag{30}$$

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

$$\widehat{\Omega} \in argmin\big\{\log|\Omega| - tr(S_N\Omega) - \lambda||\Omega||_1\big\} \tag{31}$$

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:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \tag{32}$$

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

$$\begin{array}{rcl} \mathbb{E}[Y^1|Y^{2:d}] & = & \Sigma_{12}\Sigma_{22}^{-1}Y^{2:d} \\ Var[Y^1|Y^{2:d}] & = & \sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \end{array} \tag{33}$$

We have the following identity:

$$\begin{pmatrix} \omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_{p-1} \end{pmatrix} \tag{34}$$

Which gives the following equations:

$$\begin{cases} \omega_{11}\sigma_1^2 + \Omega_{12}\Sigma_{12}^T & = 1 \qquad (*) \\ \omega_{11}\Sigma_{12} + \Omega_{12}\Sigma_{22} & = 0 \qquad (**) \\ \Omega_{12}^T\Sigma_{12} + \Omega_{22}\Sigma_{22} & = I_{p-1} \quad (***) \end{cases} \tag{35}$$

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

$$\begin{cases} \mathbb{E}[Y^1|Y^{2:d}] & = -\frac{1}{\omega_{11}}\Omega_{12}Y^{2:d} \\ Var[Y^1|Y^{2:d}] & = \frac{1}{\omega_{11}} \end{cases} \tag{36}$$

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

$$\begin{aligned} Y_i^1 &= -\tfrac{1}{\omega^*_{11}}\Omega_{12}Y_i^{2:d} + \tfrac{1}{\sqrt{\omega^*_{11}}}\xi_i^1 \\ &= -\textstyle\sum_{j=2}^{d} \tfrac{w^*_{ij}}{\omega^*_{11}} Y_i^j + \tfrac{1}{\sqrt{\omega^*_{11}}}\xi_i^1 \end{aligned} \tag{37}$$

and

$$\beta_1^{*T} Y_i = \frac{1}{\sqrt{\omega^*_{11}}}\xi_i^1 \Rightarrow \beta_1^{*T}\boldsymbol{Y} = \frac{1}{\sqrt{\omega^*_{11}}}\boldsymbol{\xi}^1 \tag{38}$$

with

$$\beta_1^* = \frac{1}{\sqrt{\omega^*_{11}}}\begin{bmatrix} w^*_{11} \\ w_{12} \\ \vdots \\ w_{1d} \end{bmatrix} \in \mathbb{R}^d \quad \text{and} \quad \boldsymbol{Y} = \begin{bmatrix} verifier \end{bmatrix} \tag{39}$$

2.3 The Square-Root Lasso

3 Comments

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, No. 1:1–38, 1977.