EC503 Final Project Report
Variational Inference: Theory and Applications

Sivaramakrishnan Sankarapandian G Sivaperumal Shruti Kannan Arman Karimian

Abstract

The focus of this project is the Variational Inference (VI) method and some of its applications. In particular, mean-field variational inference is used to find an estimate of the posterior distribution for each of the latent variables. This goal is obtained by decreasing Kullback-Leibler divergence as a measure of distance between the posterior distribution of the latent variables and a candidate distribution from the mean field of distributions. We begin by explaining the idea behind the Expectation-Maximization algorithm, and later we draw the analogy between the EM and the VI methods. Later on, we implement the VI on the univariate and multivariate Gaussian mixture models. In the end, we try to segment images based on their color map.

1. Introduction

1.1. Motivation

In Bayesian inference, the posterior distribution of a latent random variable $\mathbf{z}$, given evidence $\mathbf{x}$ is given by:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})}{p(\mathbf{x})} \tag{1}$$

The denominator in 1 is the marginal distribution of evidence $\mathbf{x}$, and calculating it becomes intractable in high dimensions. It is a common approach to find the posterior distribution using methods in which finding $p(\mathbf{x})$ is not required.

One of such techniques is to use Markov Chain Monte Carlo (MCMC) method for generating samples from the latent variables based on the current evidence $\mathbf{x}$. This method however, takes a lot of time to converge and needs considerations for reducing rejections in the sampling process. Variational Inference is an alternative technique that uses KL-divergence as a measure of distance between two distributions, and tries to minimize this objective function for different families of parametric (and also non-parametric) distributions in order to find the closest one to the posterior.

Variational Inference uses a special family of distributions to find the best estimate of the posterior distribution. In order to measure how close the estimated distributions is to the actual distribution, it uses Kullback Leibler divergence as a measure of distance.

1.2. KL divergence

KL divergence is a measure (but not a metric) of the non-symmetric difference between two probability distributions p and q. For a discrete model, it is defined by:

$$\mathrm{KL}(p||q) = \sum_i p(i) \ln \frac{p(i)}{q(i)} \tag{2}$$

For the continuous models, it is defined to be:

$$\mathrm{KL}(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx \tag{3}$$

From the Gibbs' equation, we know that:

$$\mathrm{KL}(p||q) \geq 0 \tag{4}$$

2. Expectation Maximization Algorithm

We begin this chapter by introducing our notation for the known and unknown data. Our feature matrix is consisted of N data in $\mathbb{R}^D$, and we have N latent variables in $\mathbb{R}^K$. We can think of the latent variables as the soft memberships or responsibilities in the GMM(Gaussian Mixture Model). Model parameters are shown by $\boldsymbol{\theta}$. In order to rule out unfavored model fitting, such as finding zero variances for a Gaussian distribution in the univariate GMM, we assign a prior for these parameters. We also assume that the latent variables are hidden from us, and our goal is to find both the model parameters and latent variables in order to cluster our feature matrix.
To sum up:

$\mathbb{X} \in \mathbb{R}^{N\times D}$	Observed data
$\mathbb{Z} \in \mathbb{R}^{N\times K}$	Latent variables
$\boldsymbol{\theta} \sim \boldsymbol{\pi}(\boldsymbol{\theta})$	Prior assumption
$\{\mathbb{X}, \mathbb{Z}\}$	Complete dataset
$\mathbb{X}$	Incomplete dataset
Our knowledge (based on model)	$p(\mathbb{Z}|\mathbb{X}, \boldsymbol{\theta})$

3. General EM Algorithm

EM algorithm is an iterative algorithm that finds ML or MAP estimates of parameters in statistical models where the model depends on unobserved latent variables. EM is consisted of two recurring steps:

- **Expectation:** creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters.
- **Maximization:** computes parameters maximizing the expected log-likelihood found on the E step.

Likelihood of the complete data is given by the expression: $p(\mathbb{X}, \mathbb{Z}|\boldsymbol{\theta})$, but since the latent variables are hidden it is not possible to maximize it. Instead we try to maximize the likelihood function $\ln p(\mathbb{X}|\boldsymbol{\theta})$. In order to do so, EM uses the following trick:

$$\ln p(\mathbb{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbb{Z}} p(\mathbb{X}, \mathbb{Z}|\boldsymbol{\theta}) \right\} \tag{5}$$

where instead of the likelihood function, it uses the expectation over the joint distribution of the features and the latent variables conditioned over the model parameters $\boldsymbol{\theta}$. Now we make use of the following lemma to relate this trick to an approximate distribution.

Lemma 3.1. *Likelihood function given in* (5) *can be written as:*

$$ln\, p(\mathbb{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p) \tag{6}$$

where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbb{Z}} q(\mathbb{Z}) ln \left\{ \frac{p(\mathbb{X}, \mathbb{Z}|\boldsymbol{\theta})}{q(\mathbb{Z})} \right\} \tag{7}$$

$$KL(q||p) = -\sum_{\mathbb{Z}} q(\mathbb{Z}) ln \left\{ \frac{p(\mathbb{Z}|\mathbb{X}, \boldsymbol{\theta})}{q(\mathbb{Z})} \right\} \tag{8}$$

Proof.

$$\ln p(\mathbf{x}) = \ln \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})}$$

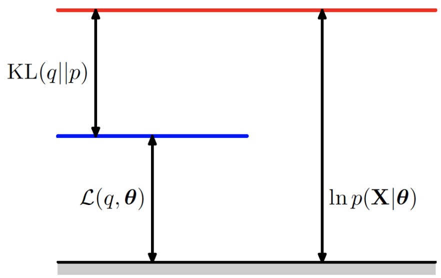
$$\rightarrow \mathrm{E}[\ln p(\mathbf{x})] = \int q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$= \int q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})} \cdot \frac{q(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z}$$

$$\rightarrow \mathrm{E}[\ln p(\mathbf{x})] = \int q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} +$$

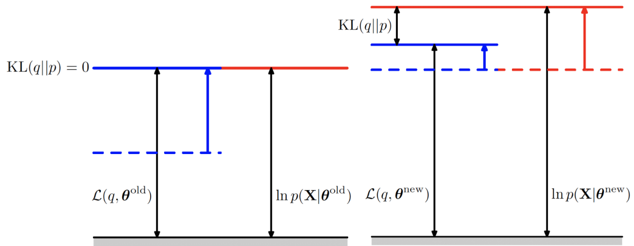
$$\int q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

By conditioning all of the above probabilities over $\boldsymbol{\theta}$, the proof becomes complete. □



$\mathrm{KL}(q||p)$
$\mathcal{L}(q, \boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

Figure 1. Relation between the log of likelihood and the ELBO and KL divergence (figure from [1])



$\mathrm{KL}(q||p) = 0$
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{old}})$
$\mathrm{KL}(q||p)$
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{new}})$

Figure 2. The two steps in the EM algorithm. Left figure shows the **E** step and the right figure shows the **M** step. (figures from [1])

Based on the Gibbs' inequality, we know that the KL divergence is always nonnegative, hence the term $\mathcal{L}(q, \boldsymbol{\theta})$, called *Expectation Lower Bound* or *ELBO* is always less than or equal to the log of likelihood.

$$\mathrm{KL}(q||p) \geq 0 \rightarrow \mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbb{X}|\boldsymbol{\theta})$$

and the equality holds if and only of the estimated distribution over the latent variables $q(\mathbb{Z})$ is equal to the posterior distribution $p(\mathbb{Z}|\mathbb{X}, \boldsymbol{\theta})$.

$$\mathrm{KL}(q||p) = 0 \rightarrow q(\mathbb{Z}) = p(\mathbb{Z}|\mathbb{X}, \boldsymbol{\theta})$$

This relation is best described in Fig. 1. Now, we revisit the EM algorithm and give a more formal description:

- **E step:** $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ is maximized with respect to $q(\mathbb{Z})$ while holding $\boldsymbol{\theta}^{old}$ fixed, by setting $q(\mathbb{Z}) = p(\mathbb{Z}|\mathbb{X}, \boldsymbol{\theta}^{old})$. In this step, $\mathrm{KL}(q||p)$ becomes zero.
- **M step:** the distribution $q(\mathbb{Z})$ is held fix and ELBO $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ to give the new estimate of the model parameters $\boldsymbol{\theta}^{new}$. By maximizing the ELBO, and using the fact that KL divergence is always non-negative, we can see that the log of likelihood increases at each iteration until it converges.

These two steps are depicted in Fig. 1. In the next section, we will use these results to estimate the model parameters and the soft membership values for the Gaussian Mixture Model.

4. EM Algorithm for Gaussian Mixture Model

We include the EM algorithm for GMM in this report so as to be compared with the VI algorithm in section 8. The evidence $p(\mathbf{X})$ of a Gaussian mixture model is given by:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$

Data we have here is given by the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. We use the matrix $\mathbb{X} \in \mathbb{R}^{N \times d}$ where the nth row of this matrix is given by $\mathbf{x}_n^\top$. Latent variables of this problem are z_{nk} which determine to which cluster a point belong, i.e. $\sum_k z_{nk} = 1$. Now we introduce the soft membership formally:

$$\begin{aligned} \gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1) p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1) p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned} \tag{10}$$

where π_k is the prior probability of $z_k = 1$ and $\gamma(z_k)$ is the corresponding posterior probability once we have observed $\mathbf{x}$. Again, we use the latent variable matrix given by: $\mathbb{Z} \in \mathbb{R}^{N \times K}$. In the M step, best estimate of the model parameters, which in this case would be the means and the covariance matrices, is given by the argmin of the following function:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) &= E_Z \Big[\ln \left\{ p(\mathbb{X}, \mathbb{Z}|\boldsymbol{\theta}) \boldsymbol{\pi}(\boldsymbol{\theta}) \right\} \Big| \boldsymbol{\theta}^{old} \Big] \\ &= \sum_{\mathbb{Z}} p(\mathbb{Z}|\mathbb{X}, \boldsymbol{\theta}^{old}) \ln \left\{ p(\mathbb{X}, \mathbb{Z}|\boldsymbol{\theta}) \boldsymbol{\pi}(\boldsymbol{\theta}) \right\} \end{aligned} \tag{11}$$

$$\boldsymbol{\theta}^{new} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{12}$$

Hence, the complete algorithm would be:

Algorithm 1 EM Algorithm for Gaussian Mixture Model

```
1: Choose an initial $\boldsymbol{\theta}^{old}$
2: **E step:** Evaluate $p(\mathbb{Z}|\mathbb{X}, \boldsymbol{\theta}^{old})$
3: **M step:** Evaluate $\boldsymbol{\theta}^{new} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$
4: $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$
5: **if** log likelihood or $\boldsymbol{\theta}$ didn't converge **then return** to 2
6: **end if**
7: **return** $\mathbb{Z}, \boldsymbol{\theta}^{new}$
```

5. Variational Inference

The basic idea of VI is to take a family of distributions $\mathcal{D}$ and find which amongst those distributions best approximates the posterior distribution that we need[3]. VI uses Kullback-Leibler (KL) divergence as a measure of the difference between the approximate and the actual distributions. Using the same trick that was introduced in the section 2, we can write:

$$\ln p(\mathbb{X}) = \mathcal{L}(q) + \text{KL}(q||p) \tag{13}$$

Where $\mathcal{L}$ is called *Evidence Lower Bound* or *ELBO*, and given by:

$$\mathcal{L}(q) = \sum_{\mathbb{Z}} q(\mathbb{Z}) \ln \left\{ \frac{p(\mathbb{X}, \mathbb{Z})}{q(\mathbb{Z})} \right\} \tag{14}$$

and the KL divergence is defined between $q(\mathbb{Z})$ and $p(\mathbb{Z}|\mathbb{X})$:

$$\text{KL}(q||p) = -\sum_{\mathbb{Z}} q(\mathbb{Z}) \ln \left\{ \frac{p(\mathbb{Z}|\mathbb{X})}{q(\mathbb{Z})} \right\} \tag{15}$$

VI works on the mean field of distributions, sometimes called factorized distributions, and therefore q will be of the following form:

$$q(\mathbb{Z}) = \prod_{i=1}^{M} q_i(\mathbb{Z}_i) \tag{16}$$

Since KL-divergence is always non-negative and $\ln p(\mathbb{X})$ is a constant, maximizing the Evidence Lower Bound (ELBO) term will automatically minimize the KL divergence[2]. The ideal approximate distribution will make the KL divergence zero.

The reason behind using factorized distribution is that they make optimization much easier and there would be a general closed form solution for the optimal distribution $q_i^\star(\mathbb{Z}_i)$ in each iteration.

Lemma 5.1. *For this family of distributions, because of their independence the best solution for each factor q_j is given by:*

$$q_j^\star(\mathbb{Z}_j) = \frac{\exp(E_{i \neq j}[ln\, p(\mathbb{X}, \mathbb{Z})])}{\int \exp(E_{i \neq j}[ln\, p(\mathbb{X}, \mathbb{Z})] d\mathbb{Z}_j)} \tag{17}$$

where $E_{i \neq j}$ indicates that the expectation is taken over all the distributions expect j^{th} distribution.

Proof.

$$
\begin{aligned}
&\mathcal{L}(q) = \int q(\mathbb{Z})\ln\Big\{\frac{p(\mathbb{X},\mathbb{Z})}{q(\mathbb{Z})}\Big\}d\mathbb{Z} \\
&= \int \Big\{\prod_i q_i(\mathbb{Z}_i)\Big(\ln p(\mathbb{X},\mathbb{Z}) - \ln q(\mathbb{Z})\Big)\Big\}d\mathbb{Z} \\
&= \int \Big\{\prod_i q_i(\mathbb{Z}_i)\Big(\ln p(\mathbb{X},\mathbb{Z}) - \sum_i \ln q_i(\mathbb{Z}_i)\Big)\Big\}d\mathbb{Z} \\
&= \int q_j(\mathbb{Z}_j)\Big\{\int \prod_{i\neq j} q_i(\mathbb{Z}_i)\ln p(\mathbb{X},\mathbb{Z})d\mathbb{Z}_i\Big\}d\mathbb{Z}_j \\
&- \int q_j(\mathbb{Z}_j)\ln q_j(\mathbb{Z}_j)d\mathbb{Z}_j + \text{const.} \\
&= \int q_j(\mathbb{Z}_j)\ln \tilde{p}(\mathbb{X},\mathbb{Z}_j)d\mathbb{Z}_j \\
&- \int q_j(\mathbb{Z}_j)\ln q_j(\mathbb{Z}_j)d\mathbb{Z}_j + \text{const.} \\
&\text{where } \tilde{p}(\mathbb{X},\mathbb{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbb{X},\mathbb{Z})] + \text{const.} \\
&= \int \prod_{i\neq j} q_i(\mathbb{Z}_i)\ln p(\mathbb{X},\mathbb{Z})d\mathbb{Z}_i \\
&\rightarrow \mathcal{L}(q) = \text{KL}(q_j(\mathbb{Z}_j)||\tilde{p}(\mathbb{X},\mathbb{Z}_j)) + \text{const.}
\end{aligned}
$$

Therefore, the minimum happens when $q_j(\mathbb{Z}_j) = \tilde{p}(\mathbb{X},\mathbb{Z}_j)$. In that case we have:

$$
\begin{aligned}
\ln q_j^\star(\mathbb{Z}_j) &= \mathbb{E}_{i\neq j}[\ln p(\mathbb{X},\mathbb{Z})] + \text{const.} \\
\rightarrow q_j^\star(\mathbb{Z}_j) &\propto \exp\big\{\mathbb{E}_{i\neq j}[\ln p(\mathbb{X},\mathbb{Z})]\big\}
\end{aligned}
\tag{18}
$$

By normalizing this, the proof becomes complete. □

6. Example explaining VI

Here we implement VI to find the distribution of mean μ and precision λ (inverse of variance) given the data D of a univariate gaussian distribution [4].Normal distribution and gamma distribution are assumed as priors to the mean and the precision respectively. Mathematically,

$$
\begin{aligned}
\mathbb{X} &\sim \mathcal{N}(\mu, \lambda^{-1}) \\
p(\mu,\lambda) &= N(\mu|\mu_0, (\kappa_0\lambda)^{-1})Ga(\lambda|a_0, b_0)
\end{aligned}
$$

In accordance with the mean field assumption, the family of distribution chosen to approximate $p(\mu,\lambda|D)$ is:

$$q(\mu,\lambda) = q(\mu)q(\lambda)$$

As seen in equation 18, the best distribution for $q(\mu)$ is given by:

$$\ln q_\mu(\mu) = E_{q_\lambda}[\ln P(\mu,\lambda,D)]$$

Solving the expectation, we get:

$$q_\mu(\mu) = (N)(\mu|\mu_N, k_N^{-1}) \tag{19}$$

Similarly,for $q(\lambda)$ it can be found that:

$$q_\lambda(\lambda) = \text{Ga}(\lambda|a_N, b_N) \tag{20}$$

The parameters of these distributions are given by:

$$
\begin{aligned}
a_N &= a_0 + \frac{N+1}{2} \\
b_N &= b_0 + k_0(\frac{1}{k_N} + \mu_N^2 + \mu_0^2 - 2\mu_N\mu_0) \\
&\quad + \frac{1}{2}\sum_{i=1}^{N}(x_i^2 + \frac{1}{k_N} + \mu_N^2 - 2\mu_N x_i) \\
\mu_N &= \frac{k_0\mu_0 + N\bar{x}}{k_0 + N} \\
k_N &= (k_0 + N)\frac{a_N}{b_N}
\end{aligned}
$$

From the equations it is clear that a_N and μ_N just depend on a_0, μ_0, k_0. However, the optimal values of b_N, k_N are found by iterative technique called Coordinate Ascent Variational Inference **(CAVI)** which is described as follows:

Algorithm 2 Coordinate Ascent Algorithm

```
1: Find $\mu_0, k_0, a_0, b_0$
2: Using the intial values find $\mu_N, k_N, a_N, b_N$
3: Evaluate **ELBO**
4: **if** ELBO didn't converge **then return** to 2
5: **end if**
```

The ELBO plot for the example we ran is given in figure3

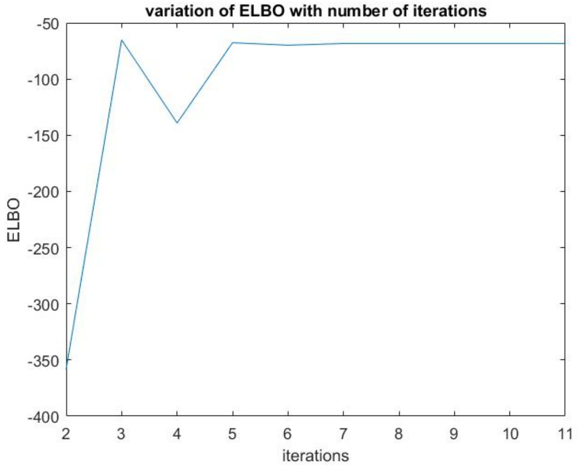


Figure 3. Variation of ELBO with iterations

7. Implementation of VI on Univariate GMM

We have implemented VI using MATLAB for unit-variance univariate mixture of Gaussians. We considered a mixture of K Gaussian distributions with their means considered to have a prior Gaussian distribution of the form $\mathcal{N}(0, \sigma^2)$ and σ^2 is a hyper-parameter. Each data point x_i is assigned to a cluster c_i which is assumed to follow a categorical distribution over all the clusters. Each component of the mixture is also a univariate Gaussian distribution of the form $\mathcal{N}(c_i^T \boldsymbol{\mu}, 1)$.

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0,\, \sigma^2) & k &= 1, \cdots, K \\ c_i &\sim \text{Categorical}(1/K, \cdots, 1/K) & i &= 1, \cdots, n \\ x_i | c_i, \boldsymbol{\mu} &\sim \mathcal{N}(c_i^T \boldsymbol{\mu},\, 1) & i &= 1, \cdots, n \end{aligned}$$

The joint distribution of the latent variables $\mathbf{z} = \{\boldsymbol{\mu}, \mathbf{c}\}$ and the data $\mathbf{x}$ is given by the following equation,

$$p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \boldsymbol{\mu})$$

The variational family of distributions after mean-field assumption for the mixture model is defined as follows,

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i) \tag{21}$$

m_k and s_k^2 are the mean and variance of k^{th} Gaussian distribution in the family of variational distributions. φ_i represents the K-vector probability of assigning each data point i to a particular Gaussian distribution. Evidence Lower Bound (ELBO) for the mixture of univariate Gaussian is given by the equation,

$$\begin{aligned} &\text{ELBO}(\mathbf{m}, \mathbf{s^2}, \boldsymbol{\varphi}) = k \\ &\quad \sum_{k=1}^{K} \Big(\text{E}[\,\ln p(\mu_k); m_k, s_k^2] - E[\ln q(\mu_k; m_k, s_k^2)] \Big) \\ &\quad + \sum_{i=1}^{n} \Big(\text{E}[\,\ln p(c_i); \varphi_i] + \text{E}[\,\ln p(x_i | c_i, \mu); \varphi_i, \mathbf{m}, \mathbf{s^2}] \\ &\quad - \text{E}[\,\ln q(c_i; \varphi_i)] \Big) \end{aligned} \tag{22}$$

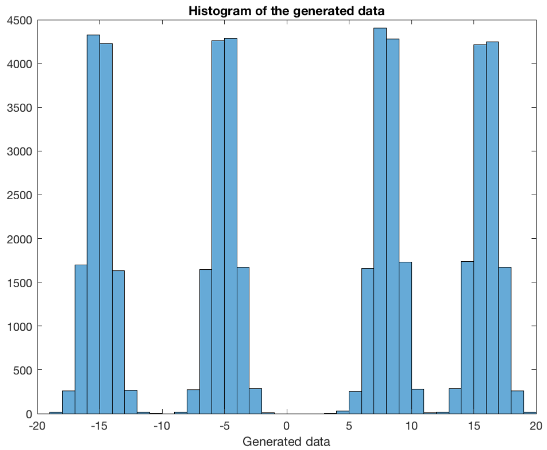


Figure 4. Univariate GMM data generated by MATLAB with four clusters

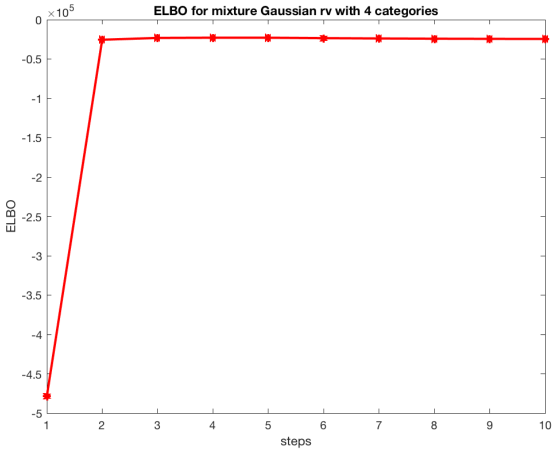


Figure 5. Convergence of ELBO

where

$$\begin{aligned} \text{E}[\ln p(\mu_k)] &= -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(s_k^2 + m_k^2) \\ \text{E}[\ln q(\mu_k; m_k, s_k^2)] &= -\frac{1}{2} \ln 2\pi s_k^2 - \frac{1}{2} \\ \text{E}[\log q(c_i; \varphi_i)] &= -\sum_{k=1}^{K} \varphi_{ik} \ln \varphi_{ik} \\ \text{E}[\ln p(c_i; \varphi_i)] &= \ln \frac{1}{K} \\ \text{E}[\ln p(x_i | c_i; \boldsymbol{\mu}); \varphi_i, \mathbf{m}, \mathbf{s^2}] &= -\frac{1}{2}[(\varphi_i^T \mathbf{m})^2 + \varphi_i^T \mathbf{s}^2 + (\varphi_i^T \mathbf{m})^2 \\ &\quad - 2\varphi_i^T \mathbf{m^2}] \end{aligned}$$

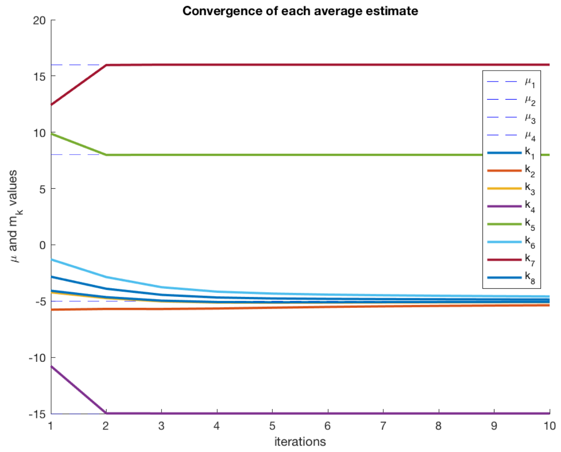


Figure 6. Means of the GMM estimated by VI for K=8

8. Implementation of VI on Multivariate GMM

The only difference in the multivariate Gaussian Mixture Model from the univariate case is that the data is represented by a mixture of Gaussian each with its own mean and covariance matrix. The procedure is similar to what we have done before - assuming a prior on the latent variables, deriving the mean field approximation and finally deriving the objective function. Then, using CAVI algorithm each parameter can be maximized with respect to other parameters resulting in the maximization of Evidence Lower Bound (ELBO). The specification of the data is as before with $\mathbb{X} = \{x_1, ..., x_N\}$ and latent variables corresponding to each data point is given by $\mathbb{Z} = \{z_1, ..., z_N\}$. The underlying GMM of the data is given by,

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \Lambda^{-1})^{z_{nk}} \tag{23}$$

The distribution of $\mathbb{Z}$ given the membership coefficients($\boldsymbol{\pi}$) is given by (a multinomial distribution).

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{24}$$

Now, for assuming priors on the latent variables, we assume a Dirichlet prior on the membership coefficients given by,

$$p(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}|\alpha_0) = C(\boldsymbol{\alpha_0}) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \tag{25}$$

and a Gaussian-Wishart Prior on the mean and precision given by,

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \tag{26}$$

$$= \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu_k}|m_0, (\beta_0 \Lambda_k)^{-1}) W(\boldsymbol{\Lambda_k}|\mathbf{W}_0, \nu_0) \tag{27}$$

By mean-field variational inference, we can assume all the latent variables to be independent of each other.

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\Lambda})q(\mathbf{Z}) \tag{28}$$

With respect to $\mathbb{Z}$, the objective function that maximizes ELBO is given by,

$$\ln q^*(\mathbf{Z}) = \boldsymbol{E}_{\pi,\mu,\Lambda}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \tag{29}$$

$$\ln q^*(\mathbf{Z}) = \boldsymbol{E}_{\pi}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \boldsymbol{E}_{\mu,\Lambda}[\ln p(\mathbf{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \tag{30}$$

Similarly for membership coefficients, mean and precision, the objective functions are,

$$\ln q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu_k}, \boldsymbol{\Lambda_k}) \tag{31}$$

$$q^*(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}|\alpha) \tag{32}$$

$$q^*(\boldsymbol{\mu_k}, \boldsymbol{\Lambda_k}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu_k}|m_k, (\beta_k \Lambda_k)^{-1}) W(\boldsymbol{\Lambda_k}|\mathbf{W}_k, \nu_k) \tag{33}$$

Implementation of CAVI algorithm involves cycling through maximizing all of these objective function one at a time till there is no further changes to the ELBO (the time at which convergence is achieved).

9. Image segmentation

Image segmentation is the process of partitioning image into regions which are visually similar in appearance. We converted the image from a three-dimensional representation (i.e) height X width X channels into a two-dimensional representation (i.e) pixel X rgb_value where each pixel can be represented in a three-dimensional space. Although this is not a great way to do image segmentation, we tried clustering the pixels based on their rgb values. As expected, the results were not great because the spatial locality is lost during the change of dimensions of the image.

References

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

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.

[3] J. D. M. David M. Blei, Alp Kucukelbir. Variational inference: A review for statisticians. *CoRR*, abs/1601.00670, 2016.

[4] K. P. Murphy. *Machine Learning A Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts, 2012.

10. Contribution

Task	**Contribution**
EM algorithm	Arman Karimian
Univariate gaussian analysis	G Sivaperumal Shruti Kannan
Univariate gaussian mixture	Arman Karimian
Multivariate gaussian analysis	Sivaramakrishnan
Literature survey	all
Presentation	all

11. Appendix

11.1. plots for simple gaussian distribution described in section 6

Plots for 4th,5th and the last iterations in figure:

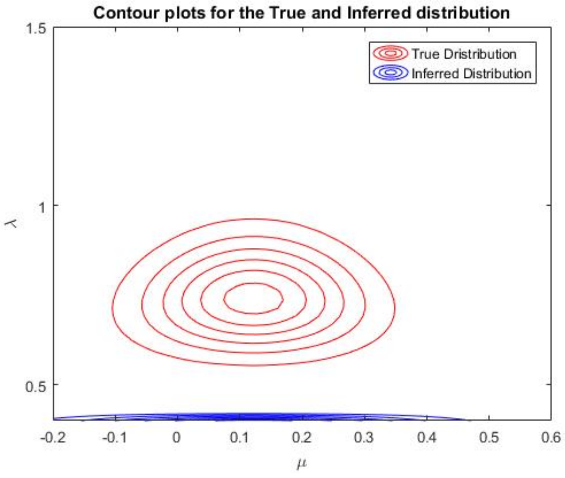


Figure 7. 4th iteration

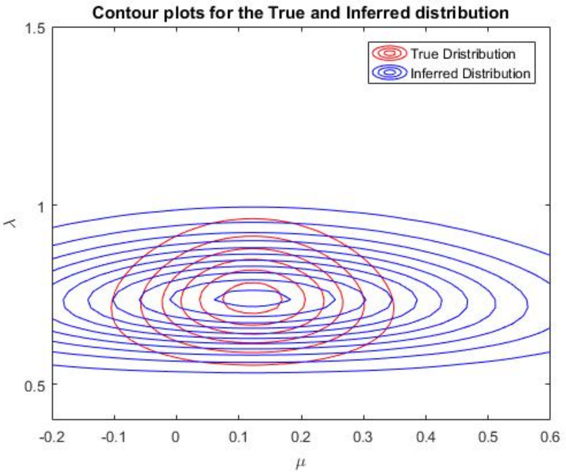


Figure 8. 5th iteration

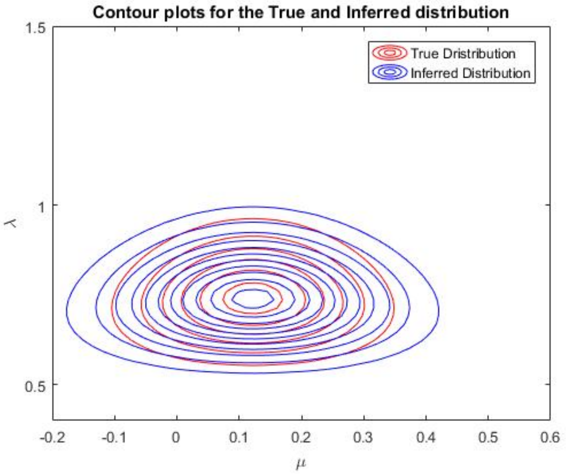


Figure 9. last iteration

11.2. plots for multivariate gaussian mixtures described in section 8

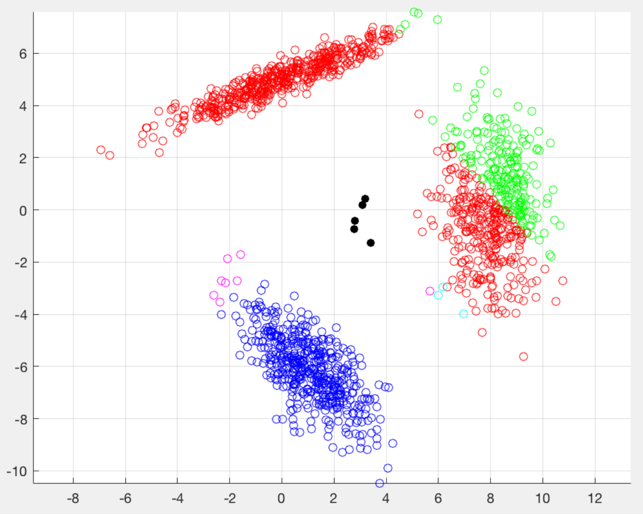


Figure 10. 2nd iteration

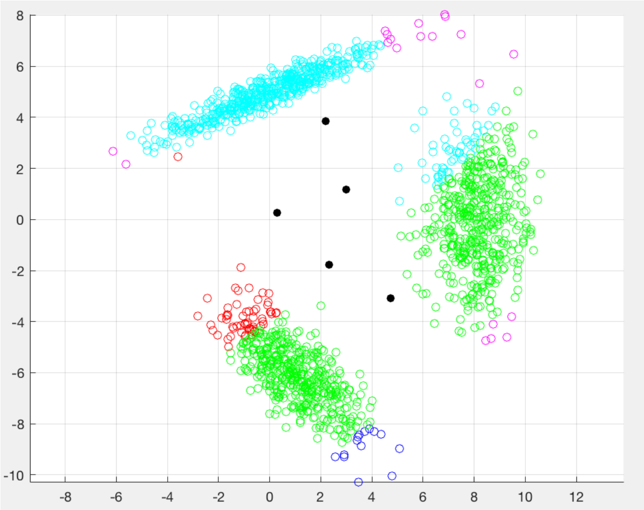


Figure 11. 14th iteration

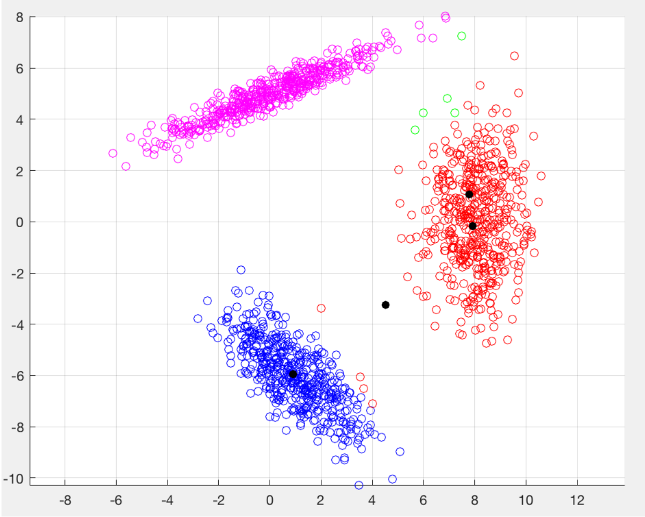


Figure 12. 22nd iteration

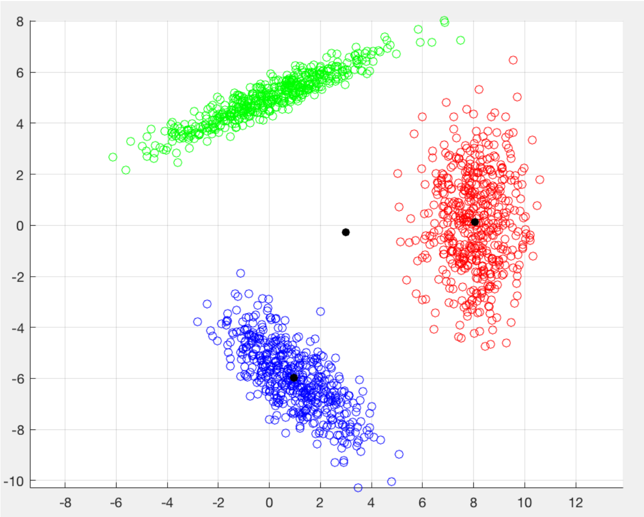


Figure 13. 51th iteration

11.3. image segmentation results

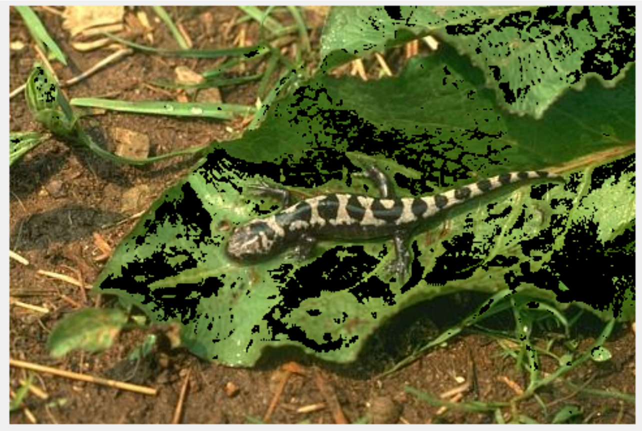


Figure 14. 1st image

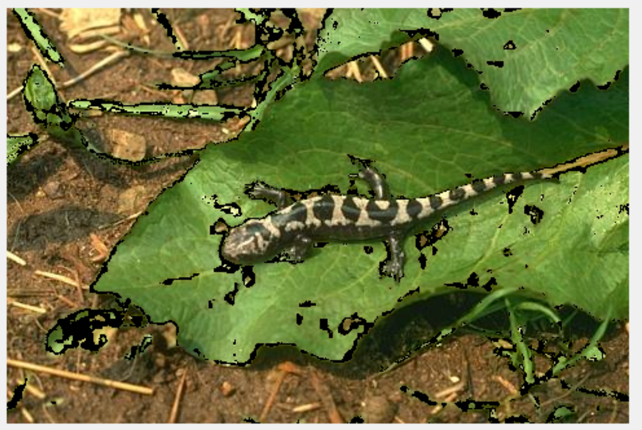


Figure 15. 2nd image

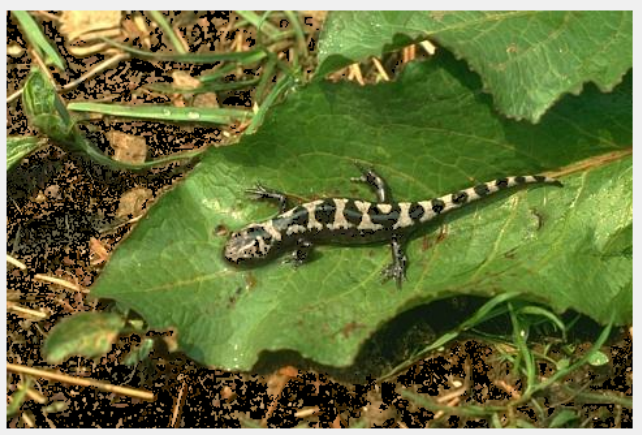


Figure 16. 3rd image

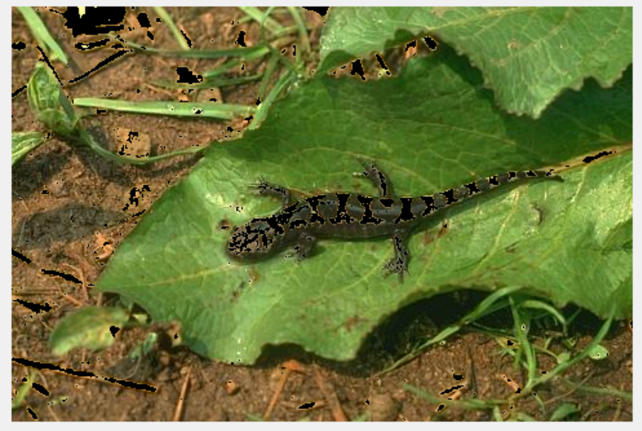


Figure 17. 4th image