Coordinate Ascent Variational Inference

Diego Garrido

Variational Inference (VI) is an approach used to approximate complicated distribution $p(z|x)$. It learn an approximate distribution $q(z)$ by optimization. In this paper, we use Coordinate Ascent Variational Inference (CAVI) to approximate the posterior distribution of a mixture of two 1D Gaussian distribution.

Jupyter Notebook

1 Gaussian Mixture Model

Consider a 1D Gaussian mixture model with prior variance of component means σ^2. The full hierarchical model is

$$\mu_k \sim \mathcal{N}(0, \sigma^2), \qquad k = 1, \ldots, K \tag{1}$$
$$c_i \sim Cat(1/K, \ldots, 1/K), \qquad i = 1, \ldots, n \tag{2}$$
$$x_i|c_i, \mu \sim \mathcal{N}(c_i^T \mu, 1) \qquad i = 1, \ldots, n \tag{3}$$

, and the mean-field variational family instead

$$q(\mu, c) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i) \tag{4}$$

2 ELBO

The Evidence Lower Bound (ELBO) to maximize is

$$ELBO(m, s^2, \varphi) = \sum_{k=1}^{K} \mathbb{E}\big[\log p(\mu_k); m_k, s_k^2\big] \tag{5}$$
$$+ \sum_{i=1}^{n} \Bigg(\mathbb{E}\big[\log p(c_i); \varphi_i\big] + \mathbb{E}\big[\log p(x_i|c_i, \mu); \varphi_i, m, s^2\big] \Bigg) \tag{6}$$
$$- \sum_{i=1}^{n} \mathbb{E}\big[\log q(c_i; \varphi_i)\big] - \sum_{k=1}^{K} \mathbb{E}\big[\log q(\mu_k; m_k, s_k^2)\big] \tag{7}$$

Using that $q(\mu_k; m_k, s_k^2) = \mathcal{N}(\mu_k; m_k, s_k^2)$ and $q(c_i; \varphi_i) = Cat(c_i; \varphi_i)$ the ELBO can be written as

$$ELBO(m, s^2, \varphi) = -\frac{1}{2\sigma^2} \sum_{k=1}^{K} m_k^2 + s_k^2 - \frac{1}{2} \log 2\pi\sigma^2 \tag{8}$$
$$- n \log K + \sum_{i=1}^{N} \sum_{k=1}^{K} \varphi_{ik} (m_k x_i - \frac{m_k^2 + s_k^2}{2}) \tag{9}$$
$$- \frac{1}{2} \sum_{i=1}^{n} x_i^2 - \frac{n}{2} \log 2\pi \tag{10}$$
$$- \sum_{i=1}^{n} \sum_{k=1}^{K} \varphi_{ik} \log \varphi_{ik} - \frac{1}{2} \sum_{k=1}^{K} \log s_k^2 \tag{11}$$

, and can be simplified to

$$ELBO(m, s^2, \varphi) = -\frac{1}{2\sigma^2} \sum_{k=1}^{K} m_k^2 + s_k^2 \tag{12}$$
$$+ \sum_{i=1}^{N} \sum_{k=1}^{K} \varphi_{ik} (m_k x_i - \frac{m_k^2 + s_k^2}{2}) \tag{13}$$
$$- \sum_{i=1}^{n} \sum_{k=1}^{K} \varphi_{ik} \log \varphi_{ik} - \frac{1}{2} \sum_{k=1}^{K} \log s_k^2 \tag{14}$$
$$+ const \tag{15}$$

, so we can omit the constant.

3 CAVI

Below the particular CAVI algorithm:

Algorithm 1: CAVI

```
input     : Data $x_{1:n}$, number of components $K$,
            prior variance of component means $\sigma^2$
output    : Variational densities
            $q(\mu_k; m_k, s_k^2)$(Gaussian) and $q(c_i; \varphi_i)$
            (K-categorical)
initialize: Variational parameters $m = m_{1:K}$,
            $s^2 = s_{1:K}^2$, and $\varphi = \varphi_{1:n}$
while the ELBO has not converged do
    for $i \in \{1, \dots, n\}$ do
        Set $\alpha_{ik} = \exp\left(m_k x_i - \frac{m_k^2 + s_k^2}{2}\right)$
        Set $\phi_{ik} = \frac{\alpha_{ik}}{\sum_{k=1}^{K} \alpha_{ik}}$
    end
    for $k \in \{1, \dots, K\}$ do
        Set $m_k = \frac{\sum_{i=1}^{n} \varphi_{ik} x_i}{1/\sigma^2 + \sum_{i=1}^{n} \varphi_{ik} x_i}$
        Set $s_k^2 = \frac{1}{1/\sigma^2 + \sum_{i=1}^{n} \varphi_{ik} x_i}$
    end
end
Compute ELBO$(m, s^2, \varphi)$
```

The convergence criterion used is $ELBO_i - ELBO_{i-1} < tol$ and $ELBO_i - ELBO_{i-1} \geq 0$, where *tol* means the numerical tolerance, in this case it was set to $1e - 16$.

4 Data

In this section is explained the process used to generate the data. Firstly, was setted the number of components to discover in $K = 2$ and the prior variance of component means in $\sigma^2 = 100$. Secondly, two means are sampled using $\mu_k \sim \mathcal{N}(0, \sigma)$. The means gotten are $\mu_1 = 2.210$ and $\mu_2 = -3.405$. Thirdly, the mixtures components are sampled using $\pi \sim Dir(1_K)$ and the clusters assignments from $c_i \sim Cat(\pi)$ with $i \in 1 \dots, n$, where the sample size n is not fixed. The π gotten is $[0.656, 0.344]$. Finally, the data is sampled using $x_i | c_i, \mu \sim \mathcal{N}(c_i^T \mu, 1)$. It is ilustrated in Figure 1 using a sample size $n = 10000$.

5 ELBO convergence

This section shows ELBO's convergece in function of different samples sizes and initial parameters. The ELBO's convergence for different sample sizes is illustrated in Figure 2. The initial variational parameters used are the same, they are $m_k = 0$ (the prior mean), $s_k = 1$ and $\phi_{ik} = 1/K$. In this figure it is observed that the greater sample size takes less iterations in converge. However, a larger sample size implies more expensive iterations. The behavior of the ELBO's convergence for different initial variational parameter is shown in Figure 3. The sample size used are the same and is $n = 100$. In this figure it is observed that different initialization of the variational parameters may lead CAVI to find different local optima of the ELBO.

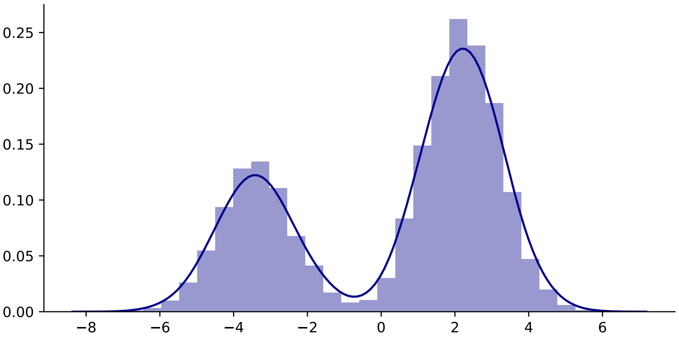

Figure 1: Random sample of a mixture of two Gaussians with sample size $n = 10000$.

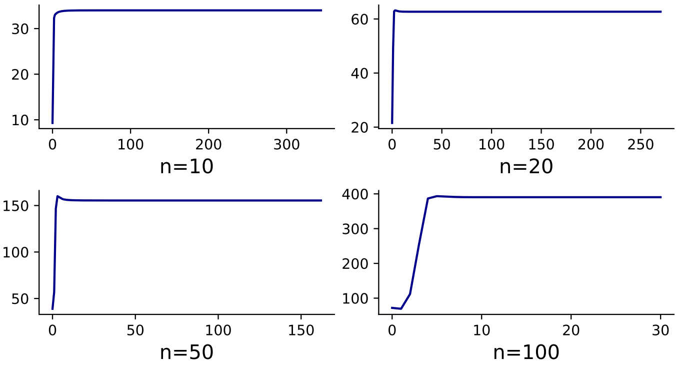

Figure 2: Iteration that the ELBO takes to converge for different sample sizes.

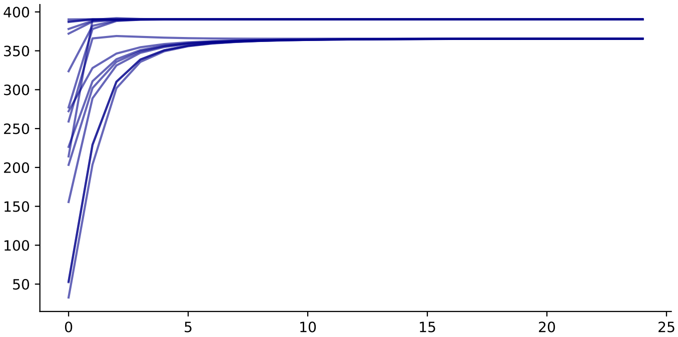

Figure 3: The ELBO using different initial variational parameters on CAVI.

6 CAVI approximation

The mean-field approximation is shown in Figure 4. In this figure can be senn that the mean-field approximation underestimate the posterionr variance. This is consequence of its objective function, since it penalizes more to put mass in regions where $p(z|x)$ has no mass. Despite underestimating the variance, the approximation is quite good, where for a sample size $n = 100$ the expected value of the means are $\mathbb{E}(\mu_1) = 2.064$ and $\mathbb{E}(\mu_2) = -3.689$, very close to the means used to generate the data.

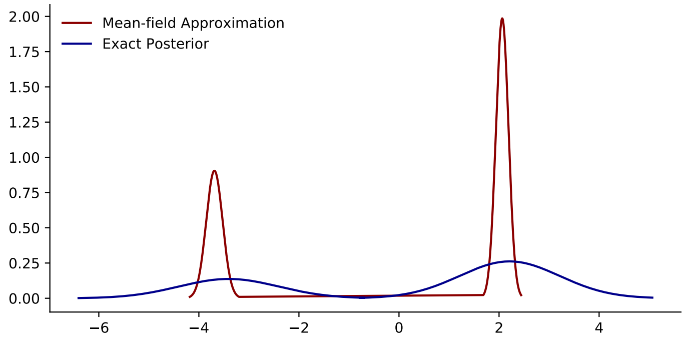

Figure 4: Mean-field approximation of a Gaussian mixture model.