

Essay title  
Machine Learning, Advanced Course/DD2434/mladv24

Aurhor  
KTH Royal Institute of Technology  
School of Engineering Sciences in Chemistry, Biotechnology and Health

November 20, 2024

1.1

The answers to the questions are as follows:

- Yes.
- No.
- Yes.
- Yes
- No.
- Yes.

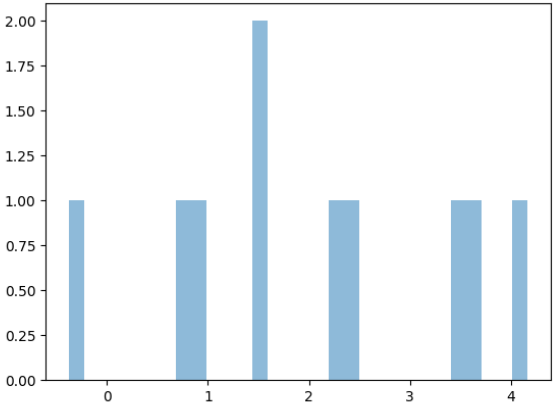


Figure 1: N=10

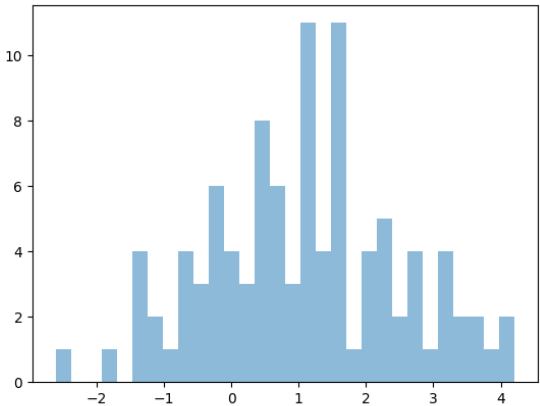


Figure 2: N=100

1.2

1.2.7

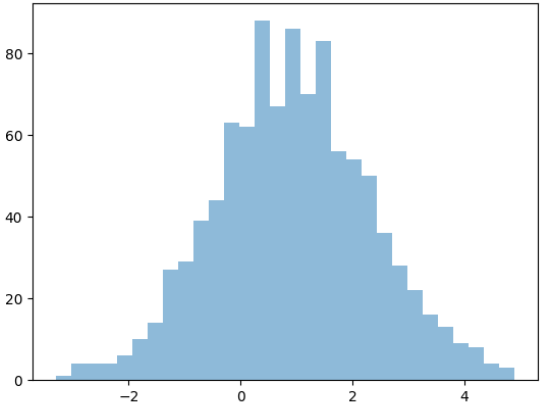


Figure 3: N=1000

Figure 4: Histograms of generated datasets

We implement a function that generates a dataset of  $N$  points in the plane, where each point is drawn from a normal distribution with mean  $\mu$  and precision  $\tau$ . For reproducibility we set the seed of the random number generator to 0. We generate datasets for  $N = 10, 100, 1000$  and plot the generated values as histograms, which results in figure Figure 4.

We see that the more datapoints we have the more the histogram resembles a normal distribution which is as expected.

## 1.2.8

We compute the ML estimates as follows:

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^N X_i$$

$$\tau_{ML} = \frac{N}{\sum_{i=1}^N (X_i - \mu_{ML})^2}.$$

We get our results for the different datasets, with random seed 0 as shown in table Table 1.

N	$\mu_{ML}$	$\tau_{ML}$
10	2.043	0.535
100	1.085	0.492
1000	0.936	0.513

Table 1: ML estimates for different datasets

We see, that the higher the number of datapoints, the closer the ML estimates are to the true values of  $\mu = 1$  and  $\tau = 0.5$ .

## 1.2.9

We derive an expression for the exact posterior. We have the likelihood function  $p(X|\mu, \tau)$  and the prior distributions  $p(\mu|\tau)$  and  $p(\tau)$ . We can write the joint posterior as:

$$p(\mu, \tau|X) \propto p(X|\mu, \tau)p(\mu|\tau)p(\tau) \quad (1)$$

The likelihood is given by:

$$p(X|\mu, \tau) = \prod_{i=1}^N p(X_i|\mu, \tau) = \prod_{i=1}^N \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(X_i - \mu)^2} \quad (2)$$

where we assume  $X_i|\mu, \tau \sim N(\mu, \frac{1}{\tau})$ .

The prior distributions are:

$$p(\mu|\tau) = N(\mu_0, \frac{1}{\lambda_0\tau}); \quad p(\tau) = \text{Gamma}(\alpha_0, \beta_0) \quad (3)$$

This gives the joint posterior:

$$p(\mu, \tau|X) \propto \prod_{i=1}^N \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(X_i - \mu)^2} \times \sqrt{\frac{\lambda_0\tau}{2\pi}} e^{-\frac{\lambda_0\tau}{2}(\mu - \mu_0)^2} \times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \tau^{\alpha_0-1} e^{-\beta_0\tau} \quad (4)$$

Simplifying, we get:

$$p(\mu, \tau|X) \propto \left(\frac{\tau}{2\pi}\right)^{\frac{N+1}{2}} e^{-\frac{\tau}{2}(\sum_{i=1}^N (X_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2 + 2\beta_0)} \tau^{\alpha_0 + \frac{N}{2}} \quad (5)$$

After completing the square for terms involving  $\mu$  and simplifying, the posterior distribution of  $\mu$  given  $X$  and  $\tau$  is:

$$p(\mu|X, \tau) \sim N\left(\frac{\sum X_i + \lambda_0\mu_0}{N + \lambda_0}, \frac{1}{(N + \lambda_0)\tau}\right). \quad (6)$$

The implementation in Code is given under 1.2.9 in the appended Jupiter Notebook.

## 1.2.10

We implement the CAVI algorithm for the system described in Bishop 10.24. We introduce the prior parameters as:

$$\mu_0 = 0; \lambda_0 = 0.1; a_0 = 0.1; b_0 = 0.1$$

We choose these values for the prior parameters, as we want a more data driven approach for the sake of this exercise. A small value for the prior precision  $\lambda_0$  and the shape parameter  $a_0$  will make the prior less informative.  $\mu_0$  is set to zero, as our guess for the mean and we set  $b_0$  to a small value to indicate a broader prior.

We run the CAVI algorithm with a tolerance of  $10^{-12}$  and plot the ELBO at each iteration for every dataset. The results are shown in figure Figure 8. We see that the ELBO converges to a local maximum for all datasets while being monotonically increasing.

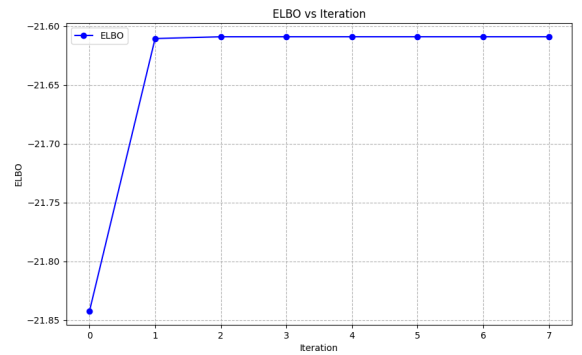


Figure 5: N=10

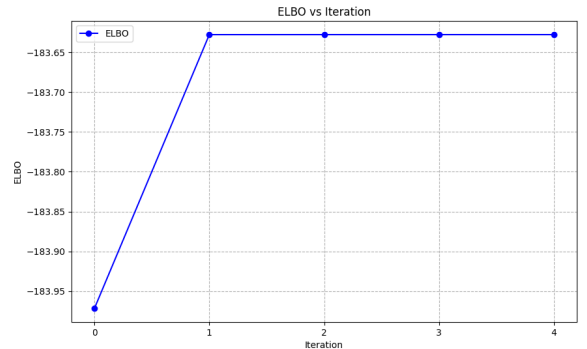


Figure 6: N=100

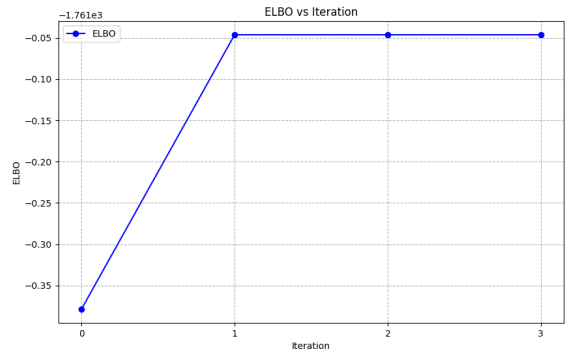


Figure 7: N=1000

Figure 8: ELBO for different datasets

We also plot the posterior mean and variance for  $\mu$  and  $\tau$  for each dataset. The results are shown in figure Figure 12.

We see, that the exact posterior is well approximated by the CAVI algorithm. The posterior mean and variance for  $\mu$  and  $\tau$  converge to the true values as the number of datapoints increases.

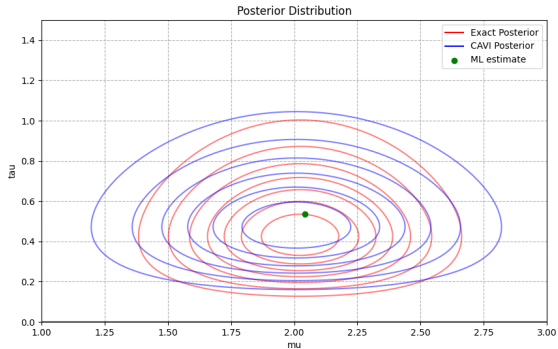


Figure 9: N=10

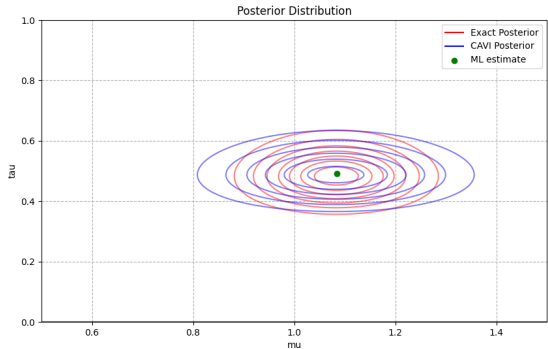


Figure 10: N=100

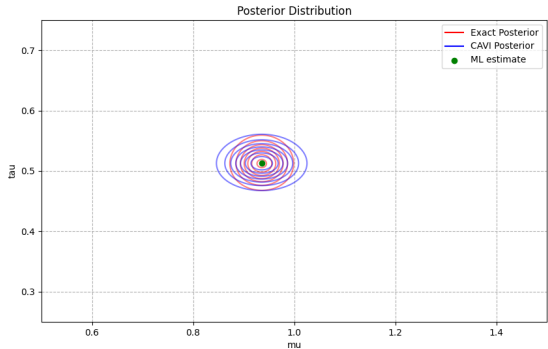


Figure 11: N=1000

Figure 12: Posterior for different datasets