Assignment 1E, 2025

DD2434 Machine Learning, Advanced Course

Bruno Carchia
carchia@kth.se

Riccardo Alfonso Cerrone
cerrone@kth.se

November 23, 2025

Group # 60

1.1 Dependencies in a Directed Graphical Model

1.1.1 Yes.

1.1.2 No.

1.1.3 No.

1.1.4 No.

1.1.5 Yes.

1.1.6 No.

1.2 Inference in Bayesian Networks

1.2.7 Fish Classification

Suppose the fish was caught on December 20, with

$$\mathbb{P}(X_1) = (0.5, 0, 0, 0.5)$$

corresponding to the four seasons (winter, spring, summer, autumn). The lightness has not been measured, but it is known that the fish is thin.

We want to compute the posterior distributions

$$\mathbb{P}(X_2 \mid X_4 = \text{thin})$$

in order to classify the fish as salmon or sea bass.

Using the Bayes rule, we have

$$\mathbb{P}(X_2 \mid X_4 = \text{thin}) = \frac{\mathbb{P}(X_4 = \text{thin} \mid X_2) \cdot \mathbb{P}(X_2)}{\mathbb{P}(X_4 = \text{thin})}$$

Where:

- $\mathbb{P}(X_4 = \text{thin} \mid X_2)$ is given by the corresponding CPT
- $\mathbb{P}(X_2) = \sum_{x_1} \mathbb{P}(X_2, X_1 = x_1) = \sum_{x_1} \mathbb{P}(X_2 \mid X_1 = x_1) \cdot \mathbb{P}(X_1 = x_1)$ can be computed using the given CPT and the prior on X_1
- $\mathbb{P}(X_4 = \text{thin})$ is a normalizing constant and for this reason we can ignore it in the computation.

At the end, we have:

$$\mathbb{P}(X_2 \mid X_4 = \text{thin}) \propto \mathbb{P}(X_4 = \text{thin} \mid X_2) \cdot \sum_{x_1} \mathbb{P}(X_2 \mid X_1 = x_1) \cdot \mathbb{P}(X_1 = x_1)$$

The second step is to compute the above expression for both values of X_2 (salmon and sea bass) and then compare the results to classify the fish.

$$\begin{aligned} \mathbb{P}(X_2 = \text{salmon} \mid X_4 = \text{thin}) &\propto 0.65 \cdot (0.5 \cdot 0.88 + 0 \cdot 0.32 + 0 \cdot 0.42 + 0.5 \cdot 0.78) \\ &\propto 0.65 \cdot (0.44 + 0 + 0 + 0.39) \\ &\propto 0.65 \cdot 0.83 \\ &\propto 0.5395 \end{aligned}$$

$$\begin{aligned} \mathbb{P}(X_2 = \text{sea bass} \mid X_4 = \text{thin}) &\propto 0.06 \cdot (0.5 \cdot 0.12 + 0 \cdot 0.68 + 0 \cdot 0.58 + 0.5 \cdot 0.22) \\ &\propto 0.06 \cdot (0.06 + 0 + 0 + 0.11) \\ &\propto 0.65 \cdot 0.17 \\ &\propto 0.0102 \end{aligned}$$

For the classification we compare the two results. As said before, we can ignore the normalizing constant because it is the same for both, so it can be seen as a good evaluation score.

$$0.5395 > 0.0102$$

so we classify the fish as **salmon**.

1.2.8 Season Inference

Suppose the fish is thin $X_4 = t$ and of medium lightness $X_3 = m$, we want to understand which season is more likely to have caught the fish X_1.

In other words we want to compute the following posterior distribution

$$\mathbb{P}(X_1 = x_1 \mid X_3 = \mathrm{m}, X_4 = \mathrm{t})$$

Using the Bayes rule we have

$$\mathbb{P}(X_1 = x_1 \mid X_3 = \mathrm{m}, X_4 = \mathrm{t}) = \frac{\mathbb{P}(X_1 = x_1, X_3 = \mathrm{m}, X_4 = \mathrm{t})}{\mathbb{P}(X_3 = \mathrm{m}, X_4 = \mathrm{t})}$$

where the denominator can be seen as a normalizing constant that for our purposes we can ignore. Consequently, we have the following

$$\mathbb{P}(X_1 = x_1 \mid X_3 = \mathrm{medium}, X_4 = \mathrm{thin}) \propto \mathbb{P}(X_1 = x_1, X_3 = \mathrm{m}, X_4 = \mathrm{t})$$

In order to apply the data given by the problem, we apply the chain rule of probability to the right hand side

$$\mathbb{P}(X_1 = x_1, X_3 = \mathrm{m}, X_4 = \mathrm{t}) = \mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(X_3 = \mathrm{m} \mid X_1 = x_1) \cdot \mathbb{P}(X_4 = \mathrm{t} \mid X_3 = \mathrm{m}, X_1 = x_1)$$

Now we can compute each term:

Term 1 - $\mathbb{P}(X_1 = x_1)$**:** is given by the prior.

Term 2 - $\mathbb{P}(X_3 = \mathbf{m} \mid X_1 = x_1)$**:** it can be computed as follows knowing that for X_3 , X_2 is enough and X_1 is independent of X_3 given X_2

$$\begin{aligned}
\mathbb{P}(X_3 = \mathrm{m} \mid X_1 = x_1) &= \sum_{x_2} \mathbb{P}(X_3 = \mathrm{m},\, X_2 = x_2 \mid X_1 = x_1) \\
&= \sum_{x_2} \mathbb{P}(X_3 = \mathrm{m} \mid X_1 = x_1,\, X_2 = x_2)\, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) \\
&= \sum_{x_2} \mathbb{P}(X_3 = \mathrm{m} \mid X_2 = x_2)\, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1)
\end{aligned}$$

Disclaimer: the derivation of Term 2 is not strictly necessary for the computations that follow.

Term 3 - $\mathbb{P}(X_4 = \mathbf{t} \mid X_3 = \mathbf{m}, X_1 = x_1)$: it needs to be computed as follows knowing that for X_4, X_2 is enough and X_1/X_3 is independent from X_4 given X_2:

$$\begin{aligned}
\mathbb{P}(X_4 = \mathrm{t} \mid X_3 = \mathrm{m}, X_1 = x_1) &= \sum_{x_2} \mathbb{P}(X_4 = \mathrm{t},\, X_2 = x_2 \mid X_3 = \mathrm{m}, X_1 = x_1) \\
&= \sum_{x_2} \mathbb{P}(X_4 = \mathrm{t} \mid X_2 = x_2,\, X_3 = \mathrm{m},\, X_1 = x_1)\, \mathbb{P}(X_2 = x_2 \mid X_3 = \mathrm{m}, X_1 = x_1) \\
&= \sum_{x_2} \mathbb{P}(X_4 = \mathrm{t} \mid X_2 = x_2)\, \mathbb{P}(X_2 = x_2 \mid X_3 = \mathrm{m}, X_1 = x_1) \\
&= \sum_{x_2} \mathbb{P}(X_4 = \mathrm{t} \mid X_2 = x_2)\, \frac{\mathbb{P}(X_3 = \mathrm{m} \mid X_2 = x_2)\, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1)}{\mathbb{P}(X_3 = \mathrm{m} \mid X_1 = x_1)}
\end{aligned}$$

We can build the final expression for the approximation of $\mathbb{P}(X_1 = x_1 \mid X_3 = \mathrm{m}, X_4 = \mathrm{t})$ substituting the three terms in the equation where $\mathbb{P}(X_3 = \mathrm{m} \mid X_1 = x_1)$ will cancel out:

$$\mathbb{P}(X_1 = x_1) \cdot \sum_{x_2} \mathbb{P}(X_3 = \mathrm{m} \mid X_2 = x_2) \cdot \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) \cdot \mathbb{P}(X_4 = \mathrm{t} \mid X_2 = x_2)$$

At this point we can compute the above expression for each value of X_1 (the four seasons) and then compare the results to find the most likely season.

- For $X_1 =$ winter:

$$\begin{aligned}
&0.25 \cdot (0.34 \cdot 0.88 \cdot 0.65 + 0.12 \cdot 0.12 \cdot 0.06) \\
&= 0.048836
\end{aligned}$$

- For $X_1 =$ spring:

$$\begin{aligned}
&0.25 \cdot (0.34 \cdot 0.32 \cdot 0.65 + 0.12 \cdot 0.68 \cdot 0.06) \\
&= 0.018904
\end{aligned}$$

- For $X_1 =$ summer:

$$\begin{aligned}
&0.25 \cdot (0.34 \cdot 0.42 \cdot 0.65 + 0.12 \cdot 0.58 \cdot 0.06) \\
&= 0.024249
\end{aligned}$$

- For $X_1 =$ autumn:

$$\begin{aligned}
&0.25 \cdot (0.34 \cdot 0.78 \cdot 0.65 + 0.12 \cdot 0.22 \cdot 0.06) \\
&= 0.043491
\end{aligned}$$

Comparing the results we have that the most likely season is **winter**.

1.3 CAVI

1.3.9 Datasets

The function `generate_data` is used to produce synthetic observations drawn from a univariate Normal distribution with mean μ and precision τ. Since the precision is defined as $\tau = 1/\sigma^2$, the corresponding standard deviation is computed as $\sigma = 1/\sqrt{\tau}$. A fixed random seed ensures reproducibility of the generated datasets. We create three datasets with increasing sample sizes ($N = 10, 100, 1000$), as requested in the assignment. The datasets are then visualized using histograms, highlighting the reduction in sampling variability for larger sample sizes.

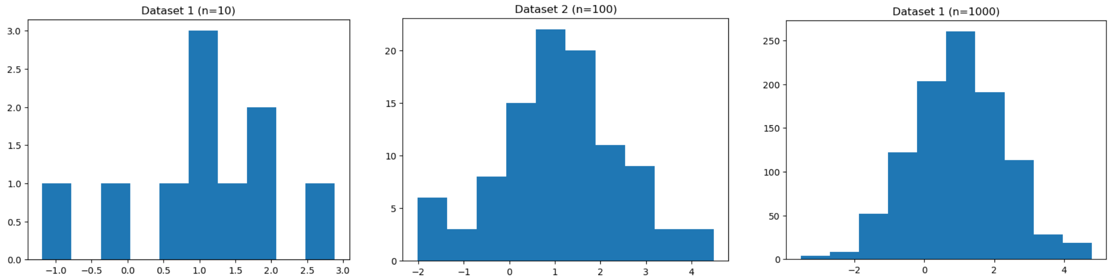

Figure 1: $N = 10$

Figure 2: $N = 100$

Figure 3: $N = 1000$

Figure 4: Histograms of the three generated datasets displayed side by side.

1.3.10 ML estimates

The function `ML_est` computes the Maximum Likelihood (ML) estimates of the parameters of a univariate Normal distribution given an observed dataset D. The ML estimate of the mean is obtained as the sample average,

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_i,$$

while the ML estimate of the variance corresponds to the empirical variance computed with `ddof = 0`, i.e.,

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{\mathrm{ML}})^2.$$

Since the precision parameter is defined as $\tau = 1/\sigma^2$, the ML estimate of the precision is given by

$$\tau_{\mathrm{ML}} = \frac{1}{\sigma^2_{\mathrm{ML}}}.$$

The function returns the pair $(\mu_{\mathrm{ML}}, \tau_{\mathrm{ML}})$, which serve as point estimates of the true mean and precision underlying the data-generating process.

We got the following results:

- $N = 10$: (1.0845817205152941, 0.884224414821371)
- $N = 100$: (1.1123121218037397, 0.5346648053646902)
- $N = 1000$: (0.9794138084900784, 0.5683337478172712)

1.3.11 Exact posterior derivation

We know that the model is defined in this assignment is defined as follows

$$X_n \mid \mu, \tau \sim \mathcal{N}\left(\mu, \frac{1}{\tau}\right), \qquad (\mu, \tau) \sim \mathcal{N}\Gamma(\mu_0, \lambda_0, \alpha_0, \beta_0)$$

Using the Bayes rule we can write the posterior as:

$$\mathbb{P}(\theta \mid D) = \frac{\mathbb{P}(D \mid \theta) \cdot \mathbb{P}(\theta)}{\mathbb{P}(D)} \propto \mathbb{P}(D \mid \theta) \cdot \mathbb{P}(\theta)$$

where:

- $\mathbb{P}(\theta) = \mathbb{P}(\mu, \tau)$ is the prior distribution

  Knowing that:

$$\mathbb{P}(\tau) \sim \text{Gamma}(\alpha_0, \beta_0) \qquad \mathbb{P}(\mu \mid \tau) \sim \mathcal{N}\left(\mu_0, \frac{1}{\lambda_0 \tau}\right)$$

$$\mathbb{P}(\mu, \tau) = \mathbb{P}(\tau) \cdot \mathbb{P}(\mu \mid \tau) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \sqrt{\frac{\lambda_0 \tau}{2\pi}} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau} e^{-\frac{\lambda_0 \tau}{2}(\mu - \mu_0)^2}$$

- $\mathbb{P}(D \mid \theta)$ is the likelihood of the data

$$\mathbb{P}(D \mid \theta) = \prod_{n=1}^{N} f_\theta(x_n) = \prod_{n=1}^{N} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x_n - \mu)^2}$$

We can now write the log-posterior ignoring the terms that are constants with respect to μ and τ

$$\log \mathbb{P}(\mu, \tau \mid X) = \log\left(\prod_{n=1}^{N} P(x_n|\mu, \tau) P(\mu, \tau)\right) = \sum_{n=1}^{N} \log P(x_n|\mu, \tau) + \log P(\mu, \tau)$$

$$\propto \sum_{n=1}^{N} \left[\frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{\tau}{2}(x_n - \mu)^2\right] + \left[\frac{1}{2}\log\left(\frac{\lambda_0 \tau}{2\pi}\right) + (\alpha_0 - 1)\log\tau - \beta_0\tau - \frac{\lambda_0 \tau}{2}(\mu - \mu_0)^2\right]$$

$$\propto \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2 + \left(\alpha_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\lambda_0\tau}{2}(\mu - \mu_0)^2$$

$$\propto \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}({x_n}^2 - 2x_n\mu + \mu^2) + \left(\alpha_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\lambda_0\tau}{2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)$$

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\tau}{2}\left[\mu^2(N+\lambda_0) - 2\mu\left(\sum_{n=1}^{N} x_n + \lambda_0\mu_0\right) + \sum_{n=1}^{N} {x_n}^2 + \lambda_0\mu_0^2\right]$$

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\tau}{2}(N+\lambda_0)\left[\mu^2 - \frac{2\mu}{N+\lambda_0}\left(\sum_{n=1}^{N} x_n + \lambda_0\mu_0\right) + \frac{1}{N+\lambda_0}\left(\sum_{n=1}^{N} {x_n}^2 + \lambda_0\mu_0^2\right)\right]$$

We can define C as:

$$C = \frac{1}{N+\lambda_0}\sum_{n=1}^{N}\left({x_n}^2 + \lambda_0\mu_0^2\right)$$

Continuing from the previous expression we have:

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\tau}{2}(N+\lambda_0)[\mu^2 - \frac{2\mu}{N+\lambda_0}(\sum_{n=1}^{N} x_n + \lambda_0\mu_0) + C]$$

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right)\log\tau - \tau\left(\beta_0 + \frac{C}{2}(N+\lambda_0)\right) - \frac{\tau}{2}(N+\lambda_0)\left[\mu^2 - \frac{2\mu}{N+\lambda_0}\left(\sum_{n=1}^{N} x_n + \lambda_0\mu_0\right)\right]$$

We can define D as:

$$D = \frac{2\mu}{N+\lambda_0}\left(\sum_{n=1}^{N} x_n + \lambda_0\mu_0\right)$$

And in order to complete the square we can add and subtract $\frac{D^2\tau(N+\lambda_0)}{2}$

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right)\log\tau - \tau\left(\beta_0 + \frac{C}{2}(N+\lambda_0)\right) - \frac{\tau}{2}(N+\lambda_0)\left[\mu^2 - \frac{2\mu}{N+\lambda_0}D\right] + \frac{D^2\tau(N+\lambda_0)}{2} - \frac{D^2\tau(N+\lambda_0}{2}$$

$$= \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right)\log\tau - \tau\left(\beta_0 + \frac{C}{2}\left(N+\lambda_0\right) - \frac{N+\lambda_0}{2}D^2\right) - \frac{\tau}{2}(N+\lambda_0)[\mu^2 - 2\mu D + D^2]$$

$$= \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right)\log\tau - \tau\left(\beta_0 + \frac{C}{2}(N+\lambda_0) - \frac{N+\lambda_0}{2}D^2\right) - \frac{\tau}{2}(N+\lambda_0)[\mu - D]^2$$

We can rewrite it as:

$$\log\mathbb{P}(\mu,\tau \mid X) \propto \left(\alpha_0^* - \frac{1}{2}\right)\log\tau - \beta_0^*\tau - \frac{\tau\lambda_0^*}{2}(\mu - \mu_0^*)^2$$

We recognize the kernel of a Normal-Gamma distribution, so the posterior is distributed as:

$$\mathbb{P}(\mu,\tau \mid X) \sim \mathcal{N}\Gamma(\mu_0^*,\, \lambda_0^*,\, \alpha_0^*,\, \beta_0^*)$$

with updated parameters:

$$\alpha_0^* = \alpha_0 + \frac{N}{2}$$

$$\lambda_0^* = N + \lambda_0$$

$$\mu_0^* = D = \frac{\lambda_0\mu_0 + \sum x_n}{\lambda_0 + N}$$

$$\beta_0^* = \beta_0 + \frac{N + \lambda_0}{2}(C - D^2) = \beta_0 + \frac{1}{2}\left[\sum x_n^2 + \lambda_0\mu_0^2 - \lambda_0^*\mu_0^{*2}\right]$$

1.3.12 VI algorithm implementation

Derivation of CAVI updates

Below we recall the standard CAVI procedure: for each latent variable, we isolate all terms in the joint density that depend on it, take the expectation with respect to the other variational factor, and recognise the resulting expression as a known member of the exponential family. This yields closed-form updates for both $q(\tau)$ and $q(\mu)$ in the Normal–Gamma model.

Update for $q^*(\tau)$. To compute $q^*(\tau)$, we collect all terms in $\log \mathbb{P}(\tau, D)$ that depend on τ and take the expectation with respect to $q(\mu)$. The resulting expression contains only a log-term in τ and a linear term in τ, which matches the canonical form of a Gamma distribution. This directly identifies the updated shape and rate parameters, leading to the expression below.

$$\begin{aligned}
\log q^*(\tau) &= \mathbb{E}_{q(\mu)}\left[\log \mathbb{P}(\tau, D)\right] \\
&= \mathbb{E}_{q(\mu)}\left[\left(a_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\lambda_0}{2}(\mu - \mu_0)^2 + \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right] \\
&= \big(a_0 - \frac{1}{2} + \frac{N}{2}\big)\log\tau \; - \; \tau\left(\beta_0 + \frac{\lambda_0}{2}\mathbb{E}_{q(\mu)}[(\mu - \mu_0)^2] + \frac{1}{2}\mathbb{E}_{q(\mu)}\left[\sum_{i=1}^{N}(x_i - \mu)^2\right]\right) \\
&= \log\tau\,\big(a_0 + \frac{N}{2} - \frac{1}{2}\big) - \tau\left(\beta_0 + \frac{\lambda_0}{2}\mathbb{E}_{q(\mu)}[(\mu - \mu_0)^2] + \frac{1}{2}\mathbb{E}_{q(\mu)}\left[\sum_{i=1}^{N}(x_i - \mu)^2\right]\right)
\end{aligned}$$

$$q^*(\tau) \propto \tau^{\,a_0 + \frac{N}{2} - \frac{1}{2} - 1}\;\exp\!\left(-\tau\left[\beta_0 + \frac{\lambda_0}{2}\mathbb{E}_{q(\mu)}\left[(\mu - \mu_0)^2\right] + \frac{1}{2}\mathbb{E}_{q(\mu)}\left[\sum_{i=1}^{N}(x_i - \mu)^2\right]\right]\right)$$

$$q(\tau) \sim \mathrm{Gam}(\alpha_N, \beta_N) \tag{1}$$

where

$$\alpha_N = \alpha_0 + \frac{N+1}{2} \tag{2}$$

$$\beta_N = b_0 + \frac{\mathbb{E}_{q(\mu)}}{2}\left[\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_i-\mu)^2\right] \tag{3}$$

Update for $q^*(\mu)$**.** For the update of $q(\mu)$, we extract all terms in $\log \mathbb{P}(\mu, D)$ that are functions of μ and take the expectation with respect to $q(\tau)$. The dependence on μ is entirely quadratic, so the optimal factor must be Gaussian. Identifying the quadratic coefficient and the linear term immediately yields the updated posterior mean and precision.

$$\begin{aligned}\log q^*(\mu) &= \mathbb{E}_{q(\tau)}\big[\log \mathbb{P}(\mu, D)\big] \\ &= \mathbb{E}_{q(\tau)}\Big[-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2 - \frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2\Big] \\ &= -\frac{\mathbb{E}_{q(\tau)}[\tau]}{2}\Big[\lambda_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\Big].\end{aligned}$$

$$q(\mu) \sim \mathcal{N}(\mu_N, \lambda_N) \tag{4}$$

where

$$\mu_N = \frac{\lambda_0\mu_0 + \sum x_n}{\lambda_0 + N} \tag{5}$$

$$\lambda_N = \mathbb{E}_{q(\tau)}[\tau](\lambda_0 + N) \tag{6}$$

Derivation of the ELBO

We now derive the Evidence Lower Bound by expanding each term in $\mathbb{E}_q[\log p(D, \mu, \tau)]$ and $\mathbb{E}_q[\log q(\mu, \tau)]$. All expectations will be computed using the variational distributions $q(\mu)$ in (4) and $q(\tau)$ in (1). In particular, note that:

$$\mathbb{E}_q[\mu] = \mu_N \quad \text{from (5)}, \qquad \mathbb{E}_q[\mu^2] = \frac{1}{\lambda_N} + \mu_N^2 \quad \text{from (4)},$$

and, analogously for the Gamma distribution in (1),

$$\mathbb{E}_q[\tau] = \frac{\alpha_N}{\beta_N}, \qquad \mathbb{E}_q[\log \tau] = \psi(\alpha_N) - \log \beta_N.$$

where ψ is the digamma function.
We start from the ELBO definition:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_q\left[\log \frac{\mathbb{P}(D,\mu,\tau)}{q(\mu,\tau)}\right]\\ &= \mathbb{E}_q[\log \mathbb{P}(D,\mu,\tau)] - \mathbb{E}_q[\log q(\mu,\tau)]\\ &= \mathbb{E}_q[\log \mathbb{P}(\mu \mid \lambda_0,\mu_0,\tau)] + \mathbb{E}_q[\log \mathbb{P}(\tau \mid \alpha_0,b_0)] + \mathbb{E}_q[\log \mathbb{P}(D \mid \mu,\tau)]\\ &\quad - \mathbb{E}_q[\log q(\mu)] - \mathbb{E}_q[\log q(\tau)]\end{aligned}$$

The joint log-density separates into prior terms for μ and τ plus the likelihood term for the dataset D.

The following sections compute each coloured term separately.

Term 1: $\mathbb{E}_q[\log p(\mu \mid \lambda_0,\mu_0,\tau)]$ We use the Normal prior over μ, whose log-density depends on $\log\tau$ and on the quadratic term $(\mu-\mu_0)^2$. Substituting the expressions and applying expectations using the moments of $q(\mu)$ from (4), we obtain:

$$\begin{aligned}\mathbb{E}_q[\log \mathbb{P}(\mu \mid \lambda_0,\mu_0,\tau)] &= \mathbb{E}_q\left[\frac{1}{2}\log\tau + \frac{1}{2}\log\frac{\lambda_0}{2\pi} - \frac{1}{2}\lambda_0\tau(\mu-\mu_0)^2\right]\\ &= \frac{1}{2}\log\frac{\lambda_0}{2\pi} + \frac{1}{2}\mathbb{E}_q[\log\tau] - \frac{1}{2}\lambda_0\mathbb{E}_q[\tau]\Big(\mathbb{E}_q[\mu^2] - 2\mu_0\mathbb{E}_q[\mu] + \mu_0^2\Big)\end{aligned}$$

Term 2: $\mathbb{E}_q[\log p(\tau \mid \alpha_0,b_0)]$ Using the Gamma prior, whose log-density is linear in $\log\tau$ and τ, and substituting expectations using the parameters from (2)-(3):

$$\begin{aligned}\mathbb{E}_q[\log \mathbb{P}(\tau \mid \alpha_0,b_0)] &= \mathbb{E}_q\big[\alpha_0\log b_0 - \log\Gamma(\alpha_0) + (\alpha_0-1)\log\tau - b_0\tau\big]\\ &= \alpha_0\log b_0 - \log\Gamma(\alpha_0) + (\alpha_0-1)\mathbb{E}_q[\log\tau] - b_0\mathbb{E}_q[\tau]\end{aligned}$$

Term 3: $\mathbb{E}_q[\log p(D \mid \mu,\tau)]$ The likelihood factorizes over the observations and contains terms depending on $(x_n-\mu)^2$. Using $\mathbb{E}_q[(x_n-\mu)^2]$ computed from (4), we obtain:

$$
\begin{aligned}
\mathbb{E}_q[\log \mathbb{P}(D \mid \mu, \tau)] &= \mathbb{E}_q\left[\log\left(\prod_{n=1}^{N} \mathbb{P}(x_n \mid \mu, \tau)\right)\right] = \mathbb{E}_q\left[\sum_{n=1}^{N} \log \mathbb{P}(x_n \mid \mu, \tau)\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_q\left[\log\left(\sqrt{\frac{\tau}{2\pi}}\, e^{-\frac{\tau}{2}(x_n-\mu)^2}\right)\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_q\left[\frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{\tau}{2}(x_n-\mu)^2\right] \\
&= \frac{N}{2}\mathbb{E}_q[\log\tau] - \frac{N}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_q[\tau]\sum_{n=1}^{N}\mathbb{E}_q\left[(x_n-\mu)^2\right] \\
&= \frac{N}{2}\mathbb{E}_q[\log\tau] - \frac{N}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_q[\tau]\left(\sum_{n=1}^{N} x_n^2 - 2\mathbb{E}_q[\mu]\sum_{n=1}^{N} x_n + N\mathbb{E}_q[\mu^2]\right)
\end{aligned}
$$

Term 4: $\mathbb{E}_q[\log q(\mu)]$ The entropy term of the Gaussian $q(\mu)$ in (4) yields:

$$
\begin{aligned}
\mathbb{E}_q[\log q(\mu)] &= \mathbb{E}_q\left[\log\left(\sqrt{\frac{\lambda_N}{2\pi}}\, e^{-\frac{\lambda_N}{2}(\mu-\mu_N)^2}\right)\right] \\
&= \mathbb{E}_q\left[\frac{1}{2}\log\frac{\lambda_N}{2\pi} - \frac{\lambda_N}{2}(\mu-\mu_N)^2\right] \\
&= \frac{1}{2}\log\frac{\lambda_N}{2\pi} - \frac{\lambda_N}{2}\underbrace{\mathbb{E}_q\left[(\mu-\mu_N)^2\right]}_{\mathrm{Var}_q[\mu]=\frac{1}{\lambda_N}} \\
&= \frac{1}{2}\log\frac{\lambda_N}{2\pi} - \frac{\lambda_N}{2}\cdot\frac{1}{\lambda_N} \\
&= \frac{1}{2}\left(\log\frac{\lambda_N}{2\pi} - 1\right)
\end{aligned}
$$

Term 5: $\mathbb{E}_q[\log q(\tau)]$ Finally, for the Gamma variational distribution in (1):

$$
\begin{aligned}
\mathbb{E}_q[\log q(\tau)] &= \mathbb{E}_q\left[\alpha_N\log\beta_N - \log\Gamma(\alpha_N) + (\alpha_N-1)\log\tau - \beta_N\tau\right] \\
&= \alpha_N\log\beta_N - \log\Gamma(\alpha_N) + (\alpha_N-1)\mathbb{E}_q[\log\tau] - \beta_N\mathbb{E}_q[\tau]
\end{aligned}
$$

This completes the derivation of the ELBO, expressed entirely in terms of the variational parameters μ_N, λ_N, α_N, β_N, which correspond directly to the CAVI updates given in (4)-(3), and μ_0, λ_0, α_0, b_0.

Conclusions

The CAVI algorithm was successfully implemented for the Normal-Gamma conjugate model and evaluated on the second datasets of varying sizes (N=100). The results demonstrate:

- **Convergence behavior**: The ELBO plot (Figure 5) shows rapid convergence within approximately 5-6 iterations, starting from an initial value of around -0.070 and stabilizing at approximately -0.035. The steep increase in the first iteration followed by marginal improvements indicates that the algorithm efficiently identifies the optimal variational parameters. This fast convergence is characteristic of conjugate variational inference, where coordinate updates have closed-form solutions.

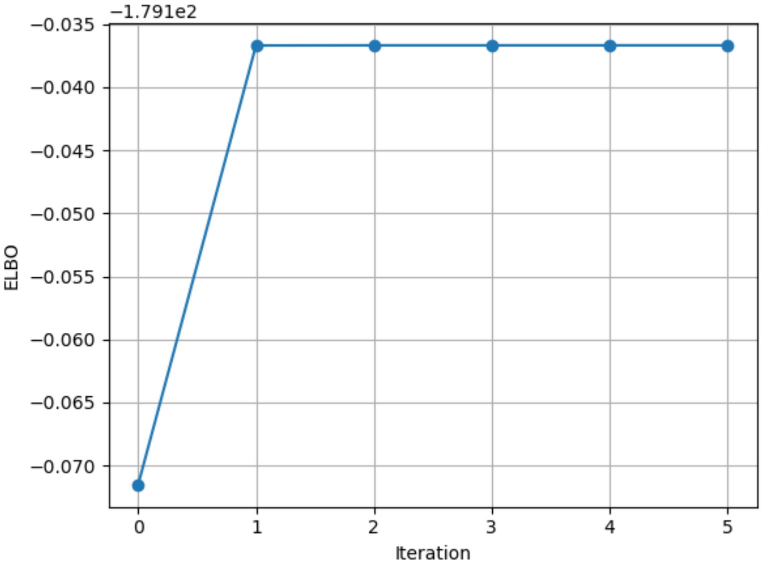

Figure 5: ELBO convergence over CAVI iterations for N=100 dataset.

- **Accuracy of approximation**: The contour plots reveal that the CAVI approximate posterior (green dashed contours) closely matches the exact posterior distribution (blue solid contours) (Figure 6 - 7) for both the mean parameter μ and precision parameter τ. The numerical comparison confirms this:

$$\text{Mean } (\mu)\text{: Exact} = 1.1122, \text{ CAVI} = 1.1122, \ \Delta\mu \approx 0.000008$$
$$\text{Precision } (\tau)\text{: Exact} = 0.5339, \text{ CAVI} = 0.5391, \ \Delta\tau \approx 0.005$$

The approximation is nearly perfect for the mean parameter, while showing a small discrepancy in the precision parameter. This slight difference in τ is expected because the mean-field assumption in variational inference (factorizing

$q(\mu, \tau) = q(\mu)q(\tau)$) introduces some approximation error, particularly in capturing the correlation structure between parameters.

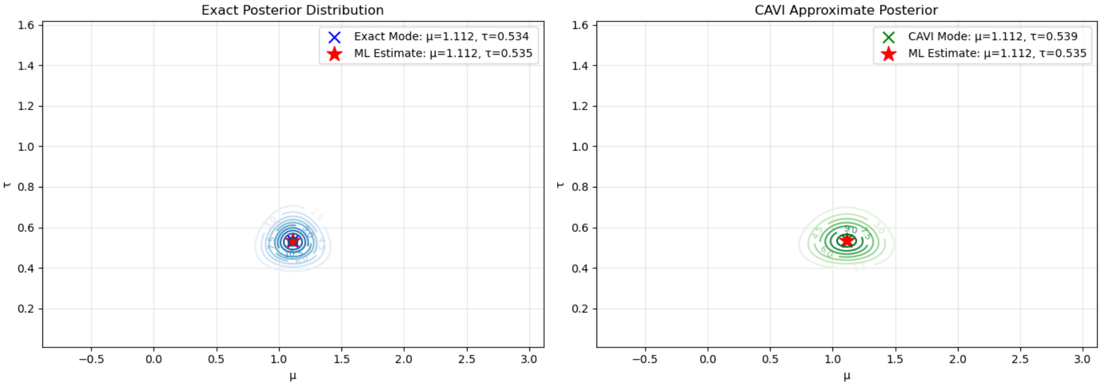

Figure 6: Showing exact posterior (blue solid) and CAVI approximation (green dashed) contours individually for N=100 dataset.

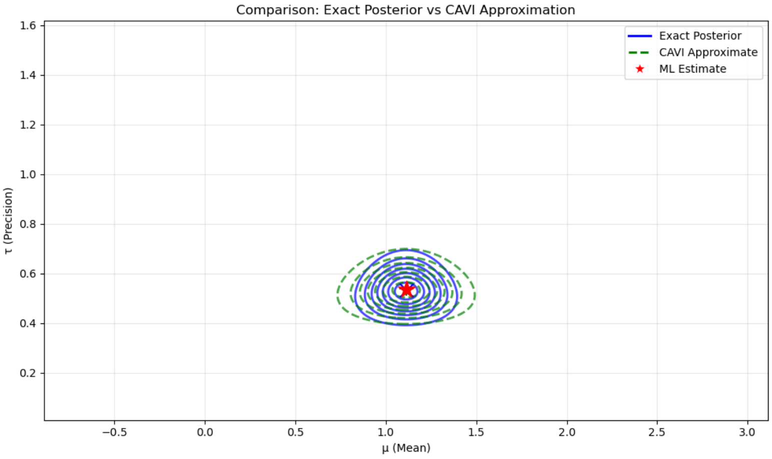

Figure 7: Contour plots comparing exact posterior (blue solid) and CAVI approximation (green dashed) for N=100 dataset. Red stars indicate ML estimates.

- **Comparison with ML estimate**: The maximum likelihood estimates ($\mu_{ML} = 1.1123$, $\tau_{ML} = 0.5347$) are marked with red stars on the contour plots and lie very close to both the exact and approximate posterior modes. This alignment is expected for large sample sizes, where the posterior becomes increasingly concentrated around the ML estimate. However, the Bayesian approaches (both exact and CAVI) provide full distributional information rather than point estimates, quantifying uncertainty through the posterior distribution's spread.
- **Final comment**: The CAVI algorithm provides an excellent approximation to the exact posterior for this conjugate Normal-Gamma model. The mean-field factoriza-

tion assumption introduces minimal error, particularly for the location parameter μ. The rapid convergence and high accuracy make CAVI a computationally efficient alternative to exact posterior computation, especially valuable for models where exact inference becomes intractable. For the dataset analyzed (N=100), all three approaches (exact posterior, CAVI, and ML) yield consistent parameter estimates, validating both the implementation and the effectiveness of variational inference for this problem.

Appendix - Code

See the following pages.