DD2434, Assignment 1A, 2024

Zyad Haddad, Erik Wallinder-Mähler

November 20, 2024

1 Assignment 1A

1.1 Dependencies in a Directed Graphical Model

1.1.1 $w_{n,g} \perp w_{n,g+1}|s_n$?

$w_{n,g} \perp w_{n,g+1}|s_n$ is **false** ($w_{n,g} \leftarrow z_n \rightarrow w_{n,g+1}$, z_n is not observed).

1.1.2 $l_n \perp w_{n,g}|x_{n,g}$?

$l_n \perp w_{n,g}|x_{n,g}$ is **false**. ($x_{n,g}$ is a descendant of $y_{n,g}$ which means that the V-structure does not give independence.)

1.1.3 $z_n \perp x_{n,g}|w_{n,g}, h_{n,g}$?

This is **true**.
(While $w_{n,g}$ and $h_{n,g}$ are observed, for example $w_{n,g+1}$ and $h_{n,g+1}$ are not. These do not have paths to $x_{n,g}$ however. There is a path from $w_{n,g+1}$ to $x_{n,g}$ through l_n, but no descendants of l_n are observed, so there is a independence-giving V-structure there.)

1.1.4 $z_1^n \perp z_M^n|C_n, A_{1:I,1:J}^{1:K}$?

This is **false**. (There is a clear path $z_1^n \rightarrow z_2^n \rightarrow \cdots \rightarrow z_{M-1}^n \rightarrow z_M^n$ that does not cross any observed variables.)

1.1.5 $X_1^n \perp X_M^n|X_2^n, C^n$?

This is **true**.
(Because X_2^n is observed, X_1^n can no longer influence X_M^n through $e_{i,r}^k$. Likewise, the paths through z_1^n run into this same issue. It also clearly cannot influence through C^n, due to it being observed.)

1.1.6 $C^n \perp C^{n+1}|z_{1:M}^n, X_{1:M}^n$?

This is **false**. (Both C^n and C^{n+1} depend on π, which is not observed.)

1.2 CAVI

1.2.7 Implement a function that generates data points for the given model. Set $\mu = 1$, $\tau = 0.5$ and generate datasets with size $N = 10, 100, 1000$. Plot the histogram for each of 3 datasets you generated.

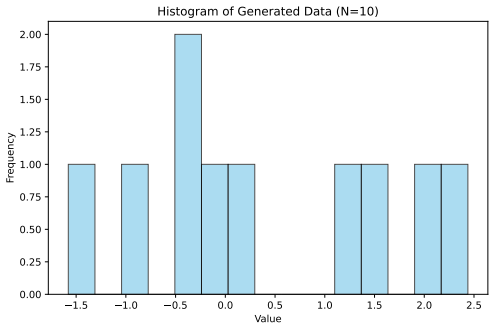

Figure 1: Histogram of data points for $N = 10$

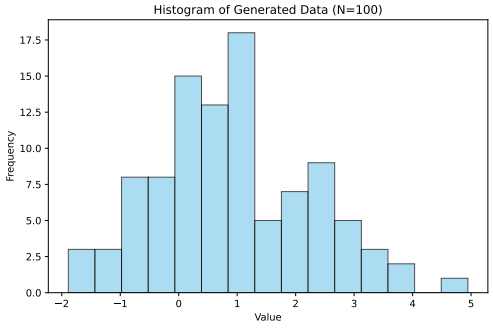

Figure 2: Histogram of data points for $N = 100$

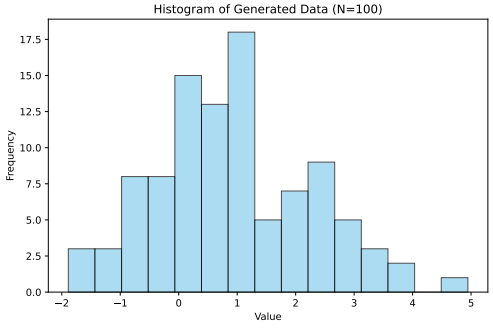

Figure 3: Histogram of data points for $N = 100$

The function is implemented as `generate_data`.

1.2.8 Find ML estimates of the variables μ and τ.

The ML estimates for μ and τ are the mean of the data, and the inverse variance of the data respectively. ML_est was implemented to return these values.

1.2.9 What is the exact posterior? (Show your derivations.)

To derive the posterior, we first begin by defining the likelihood and prior:

$$\text{Likelihood: } p(X|\mu,\tau) = (\frac{\tau}{2\pi})^{N/2}\exp\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\} \text{ - As defined in Bishop 10.21} \tag{1}$$

With the given conjugate prior distributions for μ and τ in Bishop 10.22 and 10.23:

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1}) = \frac{\sqrt{\tau\lambda_0}}{\sqrt{2\pi}}\exp\{-\frac{\lambda_0\tau(x-\mu)}{2}\} \\ p(\tau) &= \text{Gam}(\tau|a_0,b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}\exp\{-b_0\tau\} \end{aligned} \tag{2}$$

The posterior is defined through Bayes' theorem, as:

$$p(\mu,\tau|X) \propto p(X|\mu,\tau)p(\mu|\tau)p(\tau), \tag{3}$$

Combining the terms results in the expression:

$$p(\mu,\tau|X) = (\frac{\tau}{2\pi})^{N/2}\exp\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\} \times \frac{\sqrt{\tau\lambda_0}}{\sqrt{2\pi}}\exp\{-\frac{\lambda_0\tau(\mu_0-\mu)^2}{2}\} \times \frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}\exp\{-b_0\tau\} \tag{4}$$

Splitting the terms by μ and τ:

$$= (\frac{\tau}{2\pi})^{N/2}\exp\{-\tau/2\}\exp\{\sum_{n=1}^{N}(x_n-\mu)^2\}\frac{\sqrt{\tau\lambda_0}}{\sqrt{2\pi}}\exp\{\tau\}\exp\{-\frac{\lambda_0}{2}(\mu-\mu_0)^2\}\frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}\exp\{-b_0\tau\} \tag{5}$$

The sum and the squared term can be expanded as:

$$\begin{aligned} \sum_{n=1}^{N}(x_n-\mu)^2 &= \sum_{n=1}^{N}x_n^2 - 2\mu\sum_{n=1}^{N}x_n + N\mu^2 \\ \lambda_0(\mu_0-\mu)^2 &= \lambda_0\mu_0^2 - 2\lambda_0\mu_0\mu + \lambda_0\mu^2 \end{aligned} \tag{6}$$

Together the terms create

$$\begin{aligned} \sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu_0-\mu)^2 &= (N+\lambda_0)\mu^2 - 2(\sum_{n=1}^{N}x_n + \lambda_0\mu_0)\mu + \sum_{n=1}^{N}x_n^2 + \lambda_0\mu_0^2 \\ &= (N+\lambda_0)(\mu^2 - \frac{\sum_{n=1}^{N}x_n + \lambda_0\mu_0}{N+\lambda_0} + \frac{\sum_{n=1}^{N}x_n^2 + \lambda_0\mu_0^2}{N+\lambda_0}) \end{aligned} \tag{7}$$

We want to be able to express these sums in a more suitable form. This can be done by completing the square $(\mu-\mu^*)$:

$$\begin{aligned} \text{We define } \mu^* \text{ as: } \mu^* &= \frac{\sum_{n=1}^{N}x_n + \lambda_0\mu_0}{N+\lambda_0} \\ \text{(Eq 7 with } N+\lambda_0 \text{ factored out)} &= \mu^2 - 2\mu^*\mu + \frac{\sum_{n=1}^{N}x_n^2 + \lambda_0\mu_0^2}{N+\lambda_0} = \\ \mu^2 - 2\mu^*\mu + (\mu^*)^2 - (\mu^*)^2 + \frac{\sum_{n=1}^{N}x_n^2 + \lambda_0\mu_0^2}{N+\lambda_0} &= \\ (\mu-\mu^*)^2 - (\mu^*)^2 + \frac{\sum_{n=1}^{N}x_n^2 + \lambda_0\mu_0^2}{N+\lambda_0} &= \\ (\mu-\mu^*)^2 + \frac{(N+\lambda_0)\left(\sum_{n=1}^{N}x_n^2 + \lambda_0\mu_0^2\right) - \left(\sum_{n=1}^{N}x_n + \lambda_0\mu_0\right)^2}{(N+\lambda_0)^2} & \end{aligned} \tag{8}$$

In equation 5, we now have:

$$
\begin{aligned}
(\frac{\tau}{2\pi})^{N/2}\exp\{-\tau/2\}\exp\{\sum_{n=1}^{N}(x_n-\mu)^2\}\frac{\sqrt{\tau\lambda_0}}{\sqrt{2\pi}}\exp\{\tau\}\exp\{-\frac{\lambda_0}{2}(\mu-\mu_0)^2\}\frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}\exp\{-b_0\tau\} = \\
(\frac{\tau}{2\pi})^{N/2}\exp\{-\tau/2\}\frac{\sqrt{\tau\lambda_0}}{\sqrt{2\pi}}\exp\{\tau\}\exp\{(N+\lambda_0)(\mu-\mu^*)^2\} \\
\exp\{\sum_{n=1}^{N}x_n^2+\lambda_0\mu_0^2)\}\exp\{-\frac{\left(\sum_{n=1}^{N}x_n+\lambda_0\mu_0\right)^2}{N+\lambda_0}\} \\
\frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}\exp\{-b_0\tau\}
\end{aligned}
\tag{9}
$$

Through combining these terms, we identify:

$$
\begin{aligned}
\mu^* &= \frac{\sum_{n=1}^{N}x_n+\lambda_0\mu_0}{N+\lambda_0} \\
\lambda^* &= N+\lambda_0 \\
a_N &= a_0+\frac{N}{2} \\
b_N &= b_0+\frac{1}{2}(\sum_{n=1}^{N}x_n^2+\lambda_0\mu_0^2-\lambda^*\mu^{*2})
\end{aligned}
\tag{10}
$$

1.2.10 Implement the VI algorithm for the variational distribution in Equation (10.24) in Bishop. Run the VI algorithm on the datasets. Plot the ELBO results. Compare the inferred variational distribution with the exact posterior and the ML estimate. Visualize the results and discuss your findings.

CAVI is implemented in the notebook, based on the factorized variational approximation from Bishop. This gives the following posterior distributions:

$$\mu_N = \frac{\lambda_0\mu_0+N\bar{x}}{\lambda_0+N} \tag{11}$$

$$\lambda_N = (\lambda_0+N)\mathbb{E}[\tau],\ \text{where } \mathbb{E}[\tau] \text{ is estimated as } \frac{a_N}{b_N} \tag{12}$$

$$a_N = a_0+\frac{N}{2} \tag{13}$$

$$b_N = b_0+\frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2+\lambda_0(\mu-\mu_0)^2\right] \tag{14}$$

Iteration with CAVI and plotting the ELBO:s gave these results over the datasets:

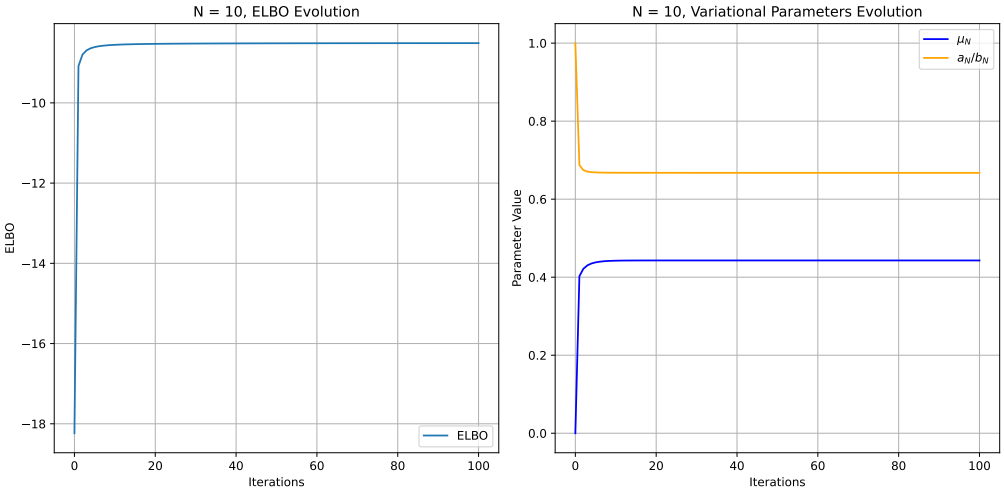

Figure 4: ELBO results and μ_N, τ_N for the dataset of size 10

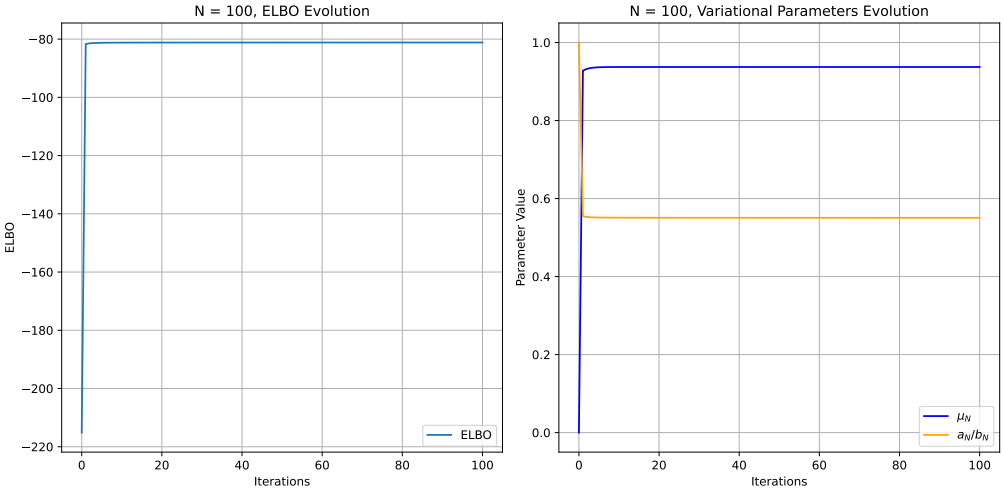

Figure 5: ELBO results and μ_N, τ_N for the dataset of size 100

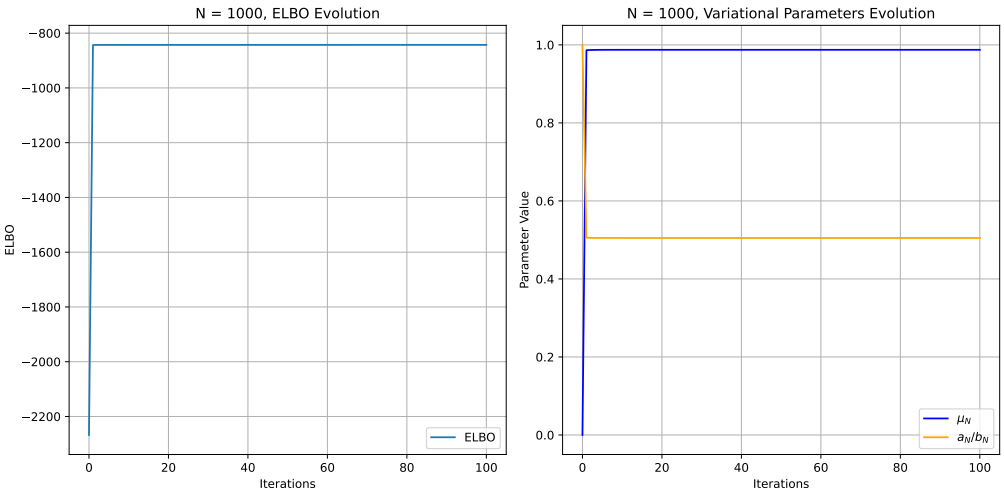

Figure 6: ELBO results and μ_N, τ_N for the dataset of size 1000

Comparing the inferred values μ, τ with the ML estimate and the exact posterior for these datasets:

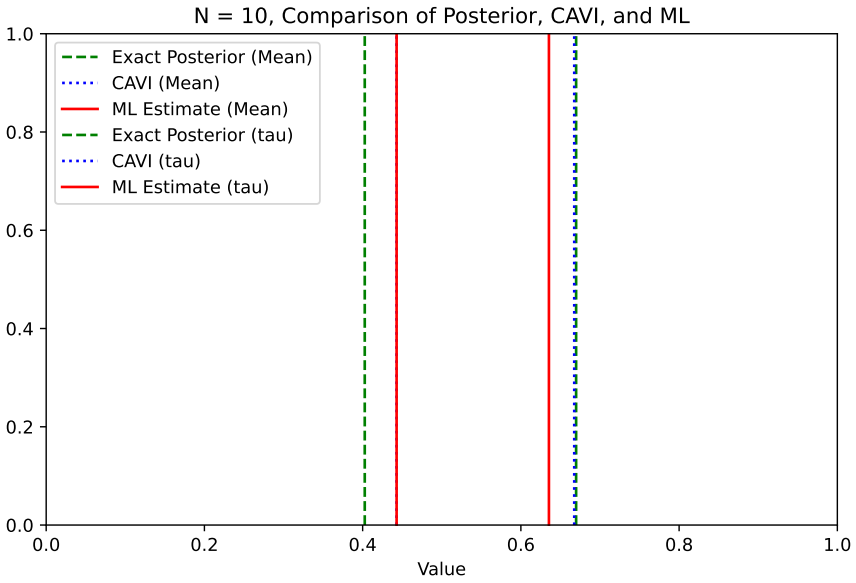

Figure 7: Comparison over the inferred parameters

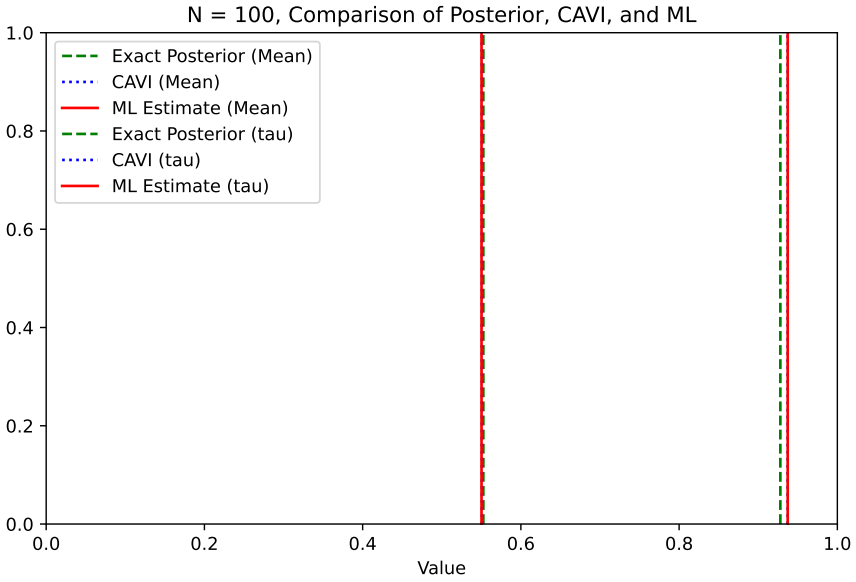

Figure 8: Comparison over the inferred parameters

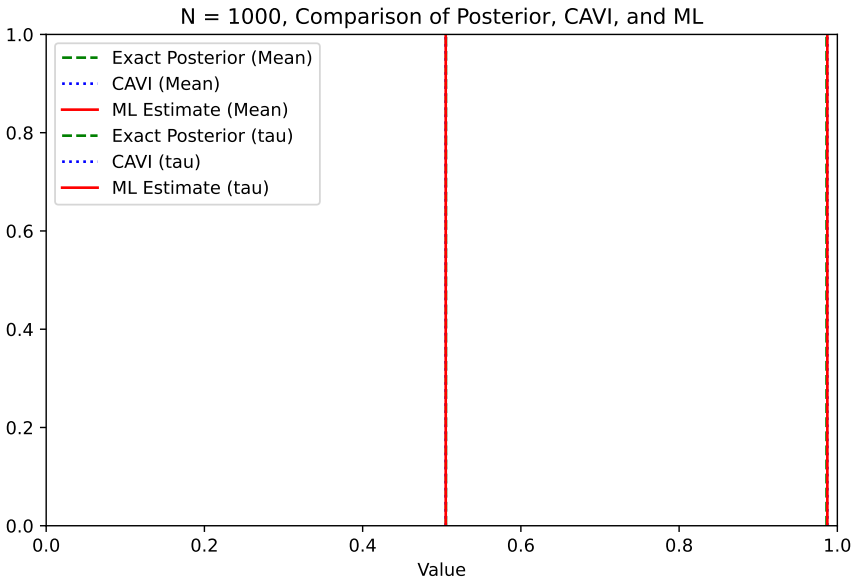

Figure 9: Comparison over the inferred parameters

Analysis

The CAVI algorithm converges very quickly for these datasets, and produces results that are comparable to or closer to the exact posterior compared with the ML estimate, even for the small datasets. Part of this is due to the simple distribution that generated the data, which is just a normal distribution. For the very small datasets, individual points have a large effect, and all methods have a large difference compared with the parameters used to generate the data $\mu = 1, \tau = 0.5$. This difference, as expected, becomes much smaller when the size of the generated data increases, and the parameters converge towards the ones used to generate the datasets.