Université de Montréal

IFT6269 - Probabilistic Graphical Models

Homework 5

Author
Patrice Béchard
20019173

December 12, 2017

1 Cautionary tale about importance sampling

Suppose that we wish to estimate the normalizing constant Z_p for an un-normalized Gaussian $\tilde{p}(x) = \exp\left(-\frac{1}{2\sigma_p^2}x^2\right)$; i.e. we have $p(\cdot) \sim \mathcal{N}(0, \sigma_p^2)$ with $p(x) = \tilde{p}(x)/Z_p$. Given N i.i.d samples $x^{(1)}, \ldots, x^{(N)}$ from a standard normal $q(\cdot) \sim \mathcal{N}(0, 1)$, consider the importance sampling estimate :

$$\hat{Z} = \frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{p}(x^{(i)})}{q(x^{(i)})}$$

(a) We can show that $\hat{Z}$ is an unbiased estimator of Z_p. First let's develop the expression of the expectation value of $\hat{Z}$:

$$\begin{aligned}
\mathbb{E}_q\left[\hat{Z}\right] &= \mathbb{E}_q\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{p}(x^{(i)})}{q(x^{(i)})}\right] \\
&= \frac{1}{N}\mathbb{E}_q\left[\sum_{i=1}^{N}\frac{\tilde{p}(x^{(i)})}{q(x^{(i)})}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_q\left[\frac{\tilde{p}(x^{(i)})}{q(x^{(i)})}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\int_x \frac{\tilde{p}(x^{(i)})}{\cancel{q(x^{(i)})}}\cancel{q(x^{(i)})}\mathrm{d}x\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{\int_x \tilde{p}(x^{(i)})\mathrm{d}x}_{Z_p}\right) \qquad \text{because } \int_x p(x)\mathrm{d}x = \int_x \frac{\tilde{p}(x)}{Z_p}\mathrm{d}x = 1 \\
&= \frac{1}{\cancel{N}}(\cancel{N}Z_p) \\
&= Z_p
\end{aligned}$$

We know that the formula to compute the bias of an estimator is :

$$\mathrm{Bias}\left[\hat{\theta}\right] = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

For our case, we thus have :

$$\mathrm{Bias}\left[\hat{Z}\right] = \mathbb{E}\left[\hat{Z}\right] - Z_p = Z_p - Z_p = 0$$

We thus have shown that $\hat{Z}$ is an unbiased estimator of Z_p. ■

(b) Letting $f(x) := \tilde{p}(x)/q(x)$, we can show that $\mathrm{Var}\left[\hat{Z}\right] = \frac{1}{N}\mathrm{Var}[f(x)]$ whenever $\mathrm{Var}[f(x)]$ is finite. To do so, let's develop the expression for the variance of $\hat{Z}$:

$$
\begin{aligned}
\mathrm{Var}\Big[\hat{Z}\Big] &= \mathbb{E}\Big[\Big(\hat{Z} - \mathbb{E}\Big[\hat{Z}\Big]\Big)^2\Big] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} f\left(x^{(i)}\right) - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} f\left(x^{(i)}\right)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} f\left(x^{(i)}\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[f\left(x^{(i)}\right)\Big]\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\left(f\left(x^{(i)}\right) - \mathbb{E}\Big[f\left(x^{(i)}\right)\Big]\right)\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_{i,j=1}^{N}\left\langle f\left(x^{(i)}\right) - \mathbb{E}\Big[f\left(x^{(i)}\right)\Big], f\left(x^{(j)}\right) - \mathbb{E}\Big[f\left(x^{(j)}\right)\Big]\right\rangle\right] \\
&= \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}\Big[\left\langle f\left(x^{(i)}\right) - \mathbb{E}\Big[f\left(x^{(i)}\right)\Big], f\left(x^{(j)}\right) - \mathbb{E}\Big[f\left(x^{(j)}\right)\Big]\right\rangle\Big] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\underbrace{\mathbb{E}\Big[\Big(f\left(x^{(i)}\right) - \mathbb{E}\Big[f\left(x^{(i)}\right)\Big]\Big)^2\Big]}_{\mathrm{Var}[f(X)]} \\
&= \frac{1}{N^2} N\ \mathrm{Var}[f(X)] \\
&= \frac{1}{N}\mathrm{Var}[f(X)] \qquad \blacksquare
\end{aligned}
$$

(c) To find for which values of σ_p^2 is the variance $\mathrm{Var}[f(x)]$ finite, let's develop the expression for $f(x)$:

$$
\begin{aligned}
f(x) = \frac{\tilde{p}(x)}{q(x)} &= \frac{\exp\left(-\frac{1}{2\sigma_p^2}x^2\right)}{\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}x^2\right)} \\
&= (2\pi)^{d/2}\exp\left(-\frac{1}{2\sigma_p^2}x^2 + \frac{1}{2}x^2\right) \\
&= (2\pi)^{d/2}\exp\left(-\frac{1}{2}x^2\left(\frac{1}{\sigma_p^2} - 1\right)\right)
\end{aligned}
$$

We notice that $f(x)$ is in fact an un-normalized Gaussian itself with variance $\left(\frac{1}{\sigma_p^2} - 1\right)^{-2}$. The value of the variance diverges when $\sigma_p^2 = 1$. The variance of $f(x)$ is thus finite for all values of ${\sigma_p}^2 \in \mathbb{R}^+ \setminus \{1\}$.

2 Gibbs sampling and mean field variational inference

We consider the Ising model with binary variables $X_s \in \{0, 1\}$ and a factorization of the form:

$$p(x; \eta) = \frac{1}{Z_p} \exp\left(\sum_{s \in V} \eta_s x_s + \sum_{\{s,t\} \in E} \eta_{st} x_s x_t\right) \tag{1}$$

We also consider the 7×7 2D grid as shown here :

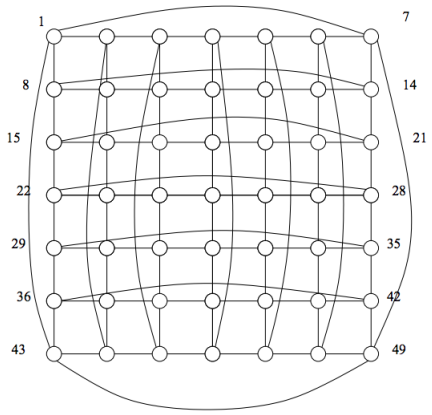

Figure 1: The undirected graphical model considered.

We will consider approximate inference methods to approximate the node marginal moments $\mu_s := p(X_s = 1)$ in this model.

(a) First, we will use Gibbs sampling to estimate the node marginal moments $\mu_s := p(X_s = 1)$. To do so, we have to derive the Gibbs sampling updates for this model. We know that the distribution from which we sample each x_s is given by :

$$\begin{aligned} p(x_i = 1 \mid x_{\neg i}) &\propto p\left(x_i, x_{\neg i}\right) \\ &= \exp\left(\eta_i \underbrace{x_i}_{=1} + \sum_{j \in N(i)} \eta_{ij} \underbrace{x_i}_{=1} x_j + \underbrace{\text{rest}}_{=0}\right) \qquad N(i) : \text{neighbors of node i} \\ &= \exp\left(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j\right) \end{aligned}$$

We re-normalize to get the conditional :

$$\begin{aligned} p(x_i = 1 \mid x_{\neg i}) &= \frac{\exp\left(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j\right)}{\underbrace{1}_{p(x_i = 0 \mid x_{\neg i})} + \exp\left(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j\right)} \\ &= \frac{1}{1 + \exp\left(-(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j)\right)} \end{aligned}$$

$$= \text{sigmoid}\left(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j\right)$$

This is our update for this model. We will only have to generate a random number from Unif$[0, 1]$ and see if its value is smaller than $p(x_i = 1 \mid x_{\neg i})$. If so, the node takes the value of 1 for this iteration, else it takes the value of 0. The class `IsingGibbs` in the file `hwk5.ipynb` is the implementation of the algorithm with cyclic sequential traversal of the nodes with given parameters. The method `_gibbs_sampling_epoch` does the update over each node of the model once. The method `gibbs_sampling` calls the `_gibbs_sampling_epoch` method 1000 times for a burn-in period, then calls it 5000 more times, collecting a sample vector of the state of the model each time. At the end of the iterations, it uses the 5000 samples to form Monte Carlo estimates $\hat{\mu}_s$ of the moments $\mathbb{E}[X_s]$ at each node. Figure 2 present the estimates obtained for each node using a colormap to present the data. We repeat this whole process 10 times and use the estimates $\hat{\mu}_s$ to estimate the empirical standard deviation at each node. Figure 3 presents the estimates the same way we did for figure 2. Numerical values of each figure are presented in the appendix.

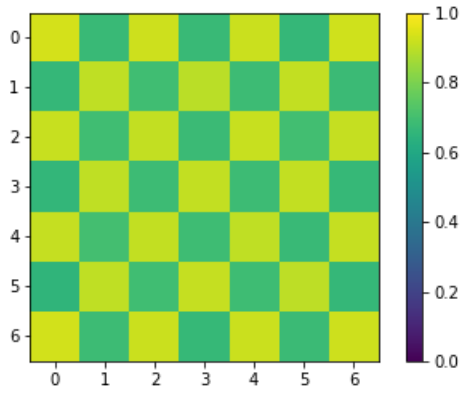

Figure 2: Estimates $\hat{\mu}_s$ of the moments $\mathbb{E}[X_s]$ at each node of the Ising model using Gibbs sampling.

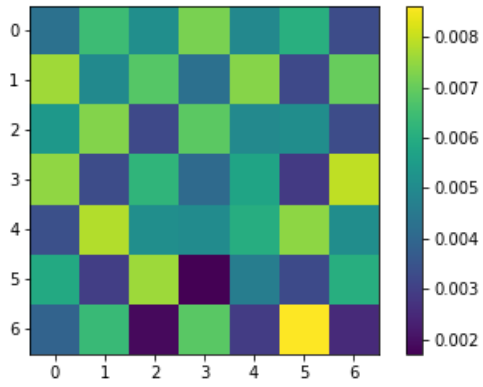

Figure 3: Empirical standard deviation of the estimates $\hat{\mu}_s$ of the moments $\mathbb{E}[X_s]$ at each node of the Ising model using Gibbs sampling.

(b) We will now use the naive mean field approximation to estimate the node marginal moments $\mu_s := p(X_s = 1)$.

First, let's derive the naive mean field updates, based on a fully factorized approximation, where we use the notation $q(X_s = 1) = \tau_s$. We first develop the log-likelihood of $p(Z \mid \eta)$:

$$\begin{aligned}
\log p(z \mid \eta) &= \log \left[\sum_x p(x, z \mid \eta) \right] \\
&= \log \left[\sum_x q(x) \frac{p(x, z \mid \eta)}{q(x)} \right] \\
&= \log \left[\mathbb{E}_q \left[\frac{p(x, z \mid \eta)}{q(x)} \right] \right] \\
&\geq \mathbb{E}_q \left[\log \left[\frac{p(x, z \mid \eta}{q(x)} \right] \right] && \text{by Jensen's inequality} \\
&= \mathcal{L}(q, \eta)
\end{aligned}$$

We have now defined an evidence lower bound, which we now use to substract from the original log probability :

$$\begin{aligned}
\log p(z \mid \eta) - \mathcal{L}(q, \eta) &= \log p(z \mid \eta) - \mathbb{E}_q \left[\log \left[\frac{p(x, z \mid \eta)}{q(x)} \right] \right] \\
&= \mathbb{E}_q[\log p(z \mid \eta)] - \mathbb{E}_q \left[\log \left[\frac{p(x, z \mid \eta)}{q(x)} \right] \right] \\
&= \mathbb{E}_q \left[\log \frac{q(x) p(z \mid \eta)}{p(x, z \mid \eta)} \right] \\
&= \mathbb{E}_q \left[\log \frac{q(x)}{p(x \mid z, \eta)} \right] \\
&= \mathrm{KL}\left(q(x) || p(x \mid z)\right)
\end{aligned}$$

Our goal is to find q such that the KL divergence is minimized. The mean field approximation is that we assume $q(x) = q(x_1, \ldots, x_m)$ factorizes as $q(x) = \prod_{s \in V} q(x_s)$. We also know that the distribution of p is given by equation 1. To update a certain q_i, we fix all other q_j for $j \neq i$, and we optimize the KL divergence with respect to q_i.

$$\begin{aligned}
\mathrm{KL}\left(q || p\right) &= \mathbb{E}_q \left[\log \frac{q(x)}{p(x|z)} \right] \\
&= \mathbb{E}_q[\log q(x) - \log p(x|z)] \\
&= \sum_x q(x) \log q(x) - \mathbb{E}_q \left[\log \left(\frac{1}{Z_p} \exp \left(\sum_{s \in V} \eta_s x_s + \sum_{\{s,t\} \in E} \eta_{st} x_s x_t \right) \right) \right] \\
&= \sum_{s \in V} \left(\sum_x q(x_s) \log q(x_s) \right) \underbrace{- \mathbb{E}_q \left[\sum_{s \in V} \eta_s x_s + \sum_{\{s,t\} \in E} \eta_{st} x_s x_t \right]}_{*} + \underbrace{\mathbb{E}_q[\log Z_p]}_{\log Z_p} \qquad (2)
\end{aligned}$$

Let's develop the (*) part independently :

$$\mathbb{E}_q\left[\sum_{s\in V}\eta_s x_s + \sum_{\{s,t\}\in E}\eta_{st}x_s x_t\right] = \mathbb{E}_{q_i}\left[\sum_{s\in V}\eta_s \mathbb{E}_{q_{\neg i}}[x_s] + \sum_{\{s,t\}\in E}\eta_{st}\mathbb{E}_{q_{\neg i}}[x_s x_t]\right]$$
$$= \mathbb{E}_{q_i}\left[\eta_i x_i + \sum_{j\neq i}\eta_j \mathbb{E}_{q_{\neg i}}[x_j] + \sum_{j\in N(i)}\eta_{ij}\mathbb{E}_{q_{\neg i}}[x_i x_j] + \underbrace{\text{rest}}_{\text{no } x_i}\right]$$
$$= \mathbb{E}_{q_i}\left[\eta_i x_i + \sum_{j\neq i}\eta_j \tau_j + \sum_{j\in N(i)}\eta_{ij} x_i \tau_j + \underbrace{\text{rest}}_{\text{no } x_i}\right]$$

We can now put back the (*) part in the KL calculation and derivate with respect to q_i :

$$\frac{\partial}{\partial q_i}\text{KL}\left(q||p\right) = \frac{\partial}{\partial q_i}\left[\sum_{s\in V}\left(\sum_x q(x_s)\log q(x_s)\right) - \mathbb{E}_{q_i}\left[\eta_i x_i + \sum_{j\neq i}\eta_j\tau_j + \sum_{j\in N(i)}\eta_{ij}x_i\tau_j + \text{rest}\right] + \log Z_p\right]$$
$$= \log q(x_i) + 1 - \eta_i x_i - x_i\sum_{j\in N(i)}\eta_{ij}\tau_j$$
$$= 0$$

We can now solve for $q(x_i)$:

$$q^{(t+1)}(x_i) \propto \exp\left(\eta_i x_i + x_i\sum_{j\in N(i)}\eta_{ij}\tau_j^{(t)}\right)$$

And finally, since x_i can take two values and that we only want the update for $q(x_i = 1) = \tau_i$, we find the value we are looking for by normalizing :

$$\tau_i^{(t+1)} = \frac{\exp\left(\eta_i(1) + (1)\sum_{j\in N(i)}\eta_{ij}\tau_j^{(t)}\right)}{\exp\left(\eta_i(0) + (0)\sum_{j\in N(i)}\eta_{ij}\tau_j^{(t)}\right) + \exp\left(\eta_i(1) + (1)\sum_{j\in N(i)}\eta_{ij}\tau_j^{(t)}\right)}$$
$$= \frac{\exp\left(\eta_i + \sum_{j\in N(i)}\eta_{ij}\tau_j^{(t)}\right)}{1 + \exp\left(\eta_i + \sum_{j\in N(i)}\eta_{ij}\tau_j^{(t)}\right)}$$
$$= \text{sigmoid}\left(\eta_i + \sum_{j\in N(i)}\eta_{ij}\tau_j^{(t)}\right)$$

Let's now derive the expression of $\text{KL}\left(q||p\right) - \log(Z_p)$, which we will use to monitor the progress of the model until convergence by plotting this value as a function of the number of epochs. We have already done the majority of the work previously, so let's start from 2:

$$\begin{aligned}\mathrm{KL}\left(q||p\right) - \log Z_p &= \sum_{s\in V}\left(\sum_x q(x_s)\log q(x_s)\right) - \mathbb{E}_q\left[\sum_{s\in V}\eta_s x_s + \sum_{\{s,t\}\in E}\eta_{st}x_s x_t\right] + \cancel{\log Z_p} - \cancel{\log Z_p} \\ &= \sum_{s\in V}\left(\tau_s\log\tau_s + (1-\tau_s)\log(1-\tau_s)\right) - \sum_{s\in V}\eta_s\tau_s - \sum_{\{s,t\}\in E}\eta_{st}\tau_s\tau_t\end{aligned}$$

Figure 4 presents the evolution for $\mathrm{KL}\left(q||p\right) - \log Z_p$ as a function of the number of epochs for 10 different random initializations of the grid.

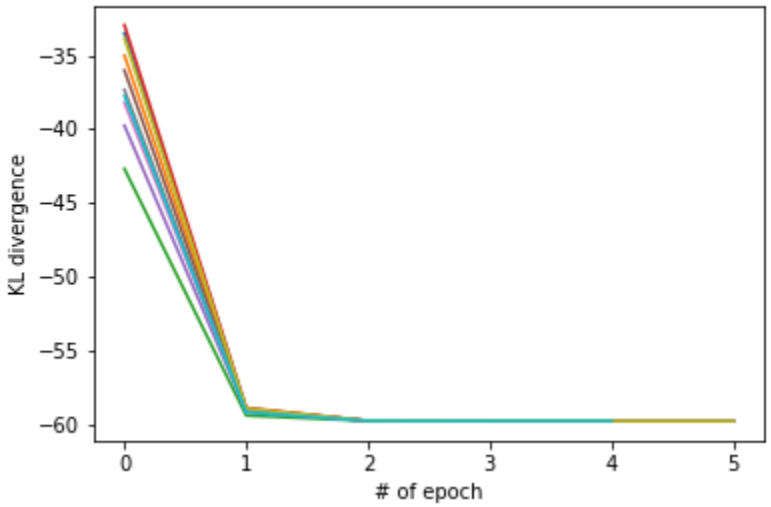

Figure 4: Progress of expression $\mathrm{KL}\left(q||p\right) - \log Z_p$ as a function of the number of epochs for 10 different random initializations of the grid using the mean field approximation for variational inference of the node marginal moments in the model.

We see that the KL divergence always converges to extremely similar values after between 2 and 5 iterations, which is much faster that the Gibbs sampling method. We can once again plot the estimates for each node using a colormap to present the data and repeat the process 10 times and use the estimates $\hat{\tau}$ to estimate the empirical standard deviation at each node. Figures 5 and 6 presents both cases, respectively. Numerical values of each figure are presented in the appendix.

We see that the estimates obtained for $\hat{\tau}$ are extremely similar to the estimates obtained for $\hat{\mu}$ for each node. Moreover, we notice that the standard deviation obtained with the latter method is smaller than the former, meaning that the method using mean field approximation yields more consistent results than method using Gibbs sampling. The mean field is thus a good approximation for our problem and does not get stuck in different local minima, which is different from some other models, where this can be a problem. Computing the L1 distance between the mean field estimated moments $\hat{\tau}_s$ and the Gibbs estimates $\hat{\mu}_s$, we get a really small value of 0.00778, meaning that both method yield really similar results.

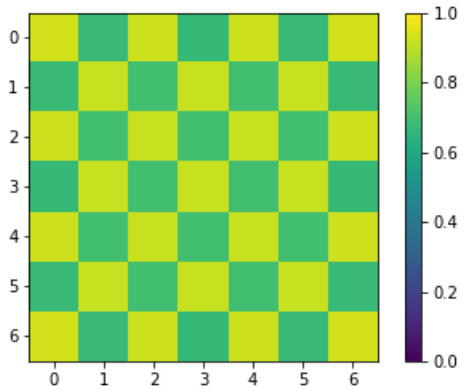

Figure 5: Estimates $\hat{\mu}_s$ of the moments $\mathbb{E}[X_s]$ at each node of the Ising model using variational inference.

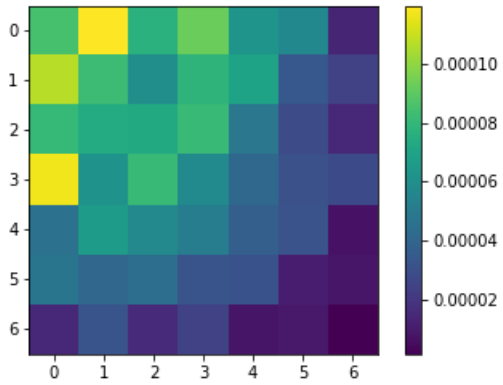

Figure 6: Empirical standard deviation of the estimates $\hat{\mu}_s$ of the moments $\mathbb{E}[X_s]$ at each node of the Ising model using variational inference.

Appendix

Results using Gibbs sampling

The numerical values of the 7x7 matrix of the estimated moments using Gibbs sampling are presented below:

```
9.3440e-01 6.7820e-01 9.2440e-01 6.7900e-01 9.2100e-01 6.6420e-01 9.2940e-01
6.6440e-01 9.0440e-01 6.9380e-01 8.9820e-01 6.8660e-01 9.0800e-01 6.8000e-01
9.1540e-01 6.9420e-01 9.0660e-01 6.8200e-01 9.1480e-01 6.9920e-01 9.1380e-01
6.6280e-01 9.0580e-01 6.8740e-01 9.0880e-01 6.8440e-01 9.0720e-01 6.6860e-01
9.1180e-01 7.0120e-01 9.0640e-01 6.8880e-01 9.0400e-01 6.7620e-01 9.1360e-01
6.5260e-01 9.0520e-01 6.9380e-01 9.1020e-01 6.8920e-01 8.9940e-01 6.6480e-01
9.3020e-01 6.8440e-01 9.1840e-01 6.7080e-01 9.1860e-01 6.8260e-01 9.2460e-01
```

The numerical values of the 7x7 matrix of the empirical standard deviation of the estimate at each node using Gibbs sampling are presented below:

```
4.2815e-03 6.4605e-03 5.1366e-03 7.2240e-03 4.9400e-03 6.0626e-03 3.3027e-03
7.6515e-03 4.9928e-03 6.7953e-03 4.2521e-03 7.3663e-03 3.2190e-03 7.0105e-03
5.4073e-03 7.3483e-03 3.2361e-03 6.8842e-03 4.9745e-03 5.0939e-03 3.3084e-03
7.4914e-03 3.3120e-03 6.2316e-03 4.0881e-03 5.7397e-03 2.8786e-03 7.9765e-03
3.3802e-03 7.8437e-03 5.1124e-03 5.0344e-03 6.0236e-03 7.4613e-03 5.0973e-03
5.9223e-03 2.9735e-03 7.6286e-03 1.6994e-03 4.6299e-03 3.2573e-03 6.0560e-03
3.8936e-03 6.3951e-03 1.8779e-03 6.8596e-03 2.9247e-03 8.6307e-03 2.5330e-03
```

Results using mean field approximation for variational inference

The numerical values of the 7x7 matrix of the estimated moments using mean field approximation for variational inference are presented below:

```
9.3099e-01 6.7247e-01 9.2268e-01 6.7156e-01 9.2278e-01 6.7261e-01 9.3108e-01
6.7259e-01 9.1429e-01 6.9713e-01 9.1529e-01 6.9716e-01 9.1446e-01 6.7270e-01
9.2268e-01 6.9709e-01 9.1625e-01 6.9665e-01 9.1628e-01 6.9725e-01 9.2286e-01
6.7148e-01 9.1528e-01 6.9663e-01 9.1618e-01 6.9671e-01 9.1532e-01 6.7176e-01
9.2278e-01 6.9714e-01 9.1620e-01 6.9669e-01 9.1626e-01 6.9723e-01 9.2288e-01
6.7261e-01 9.1439e-01 6.9721e-01 9.1532e-01 6.9723e-01 9.1449e-01 6.7272e-01
9.3107e-01 6.7266e-01 9.2285e-01 6.7177e-01 9.2288e-01 6.7272e-01 9.3111e-01
```

The numerical values of the 7x7 matrix of the empirical standard deviation of the estimate at each node using mean field approximation for variational inference are presented below:

```
8.5176e-05 1.1977e-04 7.7009e-05 9.2897e-05 6.2700e-05 5.6627e-05 1.3618e-05
1.0708e-04 8.2448e-05 5.9993e-05 7.8298e-05 7.0057e-05 3.4120e-05 2.4267e-05
8.0537e-05 7.4273e-05 7.3517e-05 8.1563e-05 4.8222e-05 2.8810e-05 1.4649e-05
1.1710e-04 6.2241e-05 8.1682e-05 5.7588e-05 4.1591e-05 3.0611e-05 2.7805e-05
4.5902e-05 6.6432e-05 5.7375e-05 5.1351e-05 3.7406e-05 3.1951e-05 7.1190e-06
4.7557e-05 4.0587e-05 4.4093e-05 3.2121e-05 3.1387e-05 1.0726e-05 8.4529e-06
1.4792e-05 3.2175e-05 1.5793e-05 2.4717e-05 8.2956e-06 9.6515e-06 1.4491e-06
```