

# Advanced Probabilistic Methods

Marshal Sinaga - marshal.sinaga@aalto.fi

Last update: March 1, 2024

This note aims to cover a few examples of probabilistic machine-learning methods. The primary reference is [Bayesian Reasoning and Machine Learning](#) by David Barber and Pekka Martinen (CS-E4820). Ideally, this note will be updated regularly until April 16, 2024.

## 1 Variational Inference

### 1.1 Variational Bayes for Simple Model

Suppose we have  $N$  independent observations  $\mathbf{x} = (x_1, \dots, x_N)$  from a two-component mixture of univariate Gaussian distributions.

$$p(x_n|\theta) = (1 - \tau)N(x_n|0, 1) + \tau N(x_n|\theta, 1) \quad (1)$$

that is with probability  $1 - \tau$  the observation  $x_n$  is generated from the first component  $N(x_n|0, 1)$ , and with probability  $\tau$  from the second component  $N(x_n|\theta, 1)$ . The model 1 has two unknown parameters  $(\tau, \theta)$ , the mixture coefficient and the mean of the second component.

The goal is to carry out a full Bayesian analysis via mean-field variational Bayesian approximation. We place the following priors on the unknown parameters.

$$\begin{aligned} \tau &\sim \text{Beta}(\alpha_0, \alpha_0) \\ \theta &\sim N(0, \beta_0^{-1}) \end{aligned}$$

We formulate the model using latent variables  $\mathbf{z} = (z_1, \dots, z_N)$ , which explicitly specify the component responsible for generating observation  $x_n$ . In detail,

$$z_n = (z_{n1}, z_{n2})^\top = \begin{cases} (1, 0)^\top & x_n \text{ is from } N(x_n|0, 1) \\ (0, 1)^\top & x_n \text{ is from } N(x_n|\theta, 1) \end{cases}$$

and place a prior on the latent variables

$$p(\mathbf{z}|\tau) = \prod_{n=1}^N \tau^{z_{n2}} (1 - \tau)^{z_{n1}}$$

The likelihood in the latent variable model is given by

$$p(\mathbf{x}|\mathbf{z}, \theta) = \prod_{n=1}^N N(x_n|0, 1)^{z_{n1}} N(x_n|\theta, 1)^{z_{n2}}$$

The joint distribution of all observed ( $\mathbf{x}$ ) and unobserved variables ( $\mathbf{z}, \tau, \theta$ ) factories as follows

$$p(\mathbf{x}, \mathbf{z}, \tau, \theta) = p(\tau)p(\theta)p(\mathbf{z}|\tau)p(\mathbf{x}|\mathbf{z}, \theta)$$

and the log joint distribution can correspondingly written as

$$\log p(\mathbf{x}, \mathbf{z}, \tau, \theta) = \log p(\tau) + \log p(\theta) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z}, \theta)$$

We approximate the posterior distribution  $p(\mathbf{z}, \tau, \theta|\mathbf{x})$  using the factorized variational distribution  $q(\mathbf{z})q(\theta)q(\tau)$

**Update factor  $q(\mathbf{z})$**  To compute the updated distribution  $q^*(\mathbf{z})$ , we first compute the expectation of the log of the joint distribution over all other unknowns in the model.

$$\begin{aligned} \log q^*(\mathbf{z}) &= \mathbb{E}_{\tau, \theta}[\log p(\mathbf{x}, \mathbf{z}, \tau, \theta)] \\ &= \mathbb{E}_{\tau}[\log p(\mathbf{z}|\tau)] + \mathbb{E}_{\theta}[\log p(\mathbf{x}|\mathbf{z}, \theta)] + \text{const} \\ &= \mathbb{E}_{\tau}\left[\sum_{n=1}^N z_{n2} \log \tau + z_{n1} \log(1 - \tau)\right] + \mathbb{E}_{\theta}\left[\sum_{n=1}^N z_{n1} \log N(x_n|0, 1) + z_{n2} \log N(x_n|\theta, 1)\right] + \text{const} \\ &= \sum_{n=1}^N z_{n2} \mathbb{E}_{\tau}[\log \tau] + z_{n1} \mathbb{E}_{\tau}[\log(1 - \tau)] + \sum_{n=1}^N z_{n1} \log N(x_n|0, 1) + z_{n2} \mathbb{E}_{\theta}[\log N(x_n|\theta, 1)] + \text{const} \\ &= \sum_{n=1}^N z_{n1} \left( \mathbb{E}_{\tau}[\log(1 - \tau)] - \frac{1}{2} \log 2\pi - \frac{1}{2} x_n^2 \right) + \sum_{n=1}^N z_{n2} \left( \mathbb{E}_{\tau}[\log \tau] - \frac{1}{2} \log 2\pi - \frac{1}{2} \mathbb{E}_{\theta}[(x_n - \theta)^2] \right) + \text{const} \\ &= \sum_{n=1}^N z_{n1} \log \rho_{n1} + z_{n2} \log \rho_{n2} + \text{const} \end{aligned} \tag{2}$$

Where we have defined  $\rho_{n1}$  and  $\rho_{n2}$  for all  $n$  as follows

$$\log \rho_{n1} = \mathbb{E}_{\tau}[\log(1 - \tau)] - \frac{1}{2} \log 2\pi - \frac{1}{2} x_n^2 \tag{3}$$

$$\log \rho_{n2} = \mathbb{E}_{\tau}[\log \tau] - \frac{1}{2} \log 2\pi - \frac{1}{2} \mathbb{E}_{\theta}[(x_n - \theta)^2] \tag{4}$$

By exponentiating both sides of Equation 2, we obtain

$$q^*(\mathbf{z}) \propto \prod_{n=1}^N \prod_{k=1}^2 \rho_{nk}^{z_{nk}}$$

which can be normalized to make a proper distribution

$$q^*(\mathbf{z}) = \prod_{n=1}^N \prod_{k=1}^2 r_{nk}^{z_{nk}}$$

where

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^2 \rho_{nj}}$$

Note that computing  $r_{nk}$  requires  $\mathbb{E}_\tau[\log \tau]$ ,  $\mathbb{E}_\tau[\log(1 - \tau)]$ , and  $\mathbb{E}_\theta[(x_n - \theta)^2]$ , where the expectations are computed over the distribution  $q(\tau)$  and  $q(\theta)$ , which will be derived next.

**Update factor  $q(\tau)$**

$$\begin{aligned} \log q^*(\tau) &= \mathbb{E}_{\mathbf{z}, \theta}[\log p(\mathbf{x}, \mathbf{z}, \tau, \theta)] \\ &= \log p(\tau) + \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{z}|\tau)] + \text{const} \\ &= \log p(\tau) + \sum_{n=1}^N \mathbb{E}_{z_n}[z_{n2}] \log \tau + \mathbb{E}_{z_n}[z_{n1}] \log(1 - \tau) + \text{const} \\ &= \log p(\tau) + \sum_{n=1}^N r_{n2} \log \tau + r_{n1} \log(1 - \tau) + \text{const} \\ &= \log \tau^{\alpha_0 - 1} + \log(1 - \tau)^{\alpha_0 - 1} + \sum_{n=1}^N \log \tau^{r_{n2}} + \log(1 - \tau)^{r_{n1}} + \text{const} \quad (5) \end{aligned}$$

$$= \log \tau^{\sum_{n=1}^N r_{n2} + \alpha_0 - 1} + \log(1 - \tau)^{\sum_{n=1}^N r_{n1} + \alpha_0 - 1} + \text{const} \quad (6)$$

We exponentiate and recognize the exponentiated form as

$$q^*(\tau) = \text{Beta}(\tau | N_2 + \alpha_0, N_1 + \alpha_0)$$

We exponentiate and recognize the exponentiated form as

$$q^*(\tau) = \text{Beta}(\tau | N_2 + \alpha_0, N_1 + \alpha_0)$$

i.e.,  $\tau$  has  $\text{Beta}(a, b)$  with  $a = N_2 + \alpha_0$  and  $b = N_1 + \alpha_0$ , where  $N_k = \sum_{n=1}^N r_{nk}$  for  $k = 1, 2$ . Using this distribution, we get the following formulas for the terms required when updating  $q(\mathbf{z})$

$$\mathbb{E}_\tau[\log \tau] = \psi(N_2 + \alpha_0) - \psi(N_1 + N_2 + 2\alpha_0) \quad (7)$$

$$\mathbb{E}_\tau[\log(1 - \tau)] = \psi(N_1 + \alpha_0) - \psi(N_1 + N_2 + 2\alpha_0) \quad (8)$$

where  $\psi$  is the digamma function. Formulas above follow from the basic property of Beta distribution and the fact that if  $\tau \sim \text{Beta}(a, b)$  then  $1 - \tau \sim \text{Beta}(b, a)$

**Update factor  $q(\theta)$**

$$\begin{aligned}
\log q^*(\theta) &= \mathbb{E}_{\tau, \mathbf{z}}[\log p(\mathbf{x}, \mathbf{z}, \tau, \theta)] \\
&= \log p(\theta) + \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{x}|\mathbf{z}, \theta)] + \text{const} \\
&= -\frac{1}{2} \log \beta_0^{-1} - \frac{\beta_0}{2} \theta^2 + \mathbb{E}_{\mathbf{z}} \left[ \sum_{n=1}^N z_{n1} \left( -\frac{1}{2} x_n^2 \right) + z_{n2} \left( -\frac{1}{2} (x_n - \theta)^2 \right) \right] + \text{const} \\
&= -\frac{1}{2} \log \beta_0^{-1} - \frac{\beta_0}{2} \theta^2 + \sum_{n=1}^N \mathbb{E}_{z_n}[z_{n1}] \left( -\frac{1}{2} x_n^2 \right) + \mathbb{E}_{z_n}[z_{n2}] \left( -\frac{1}{2} (x_n - \theta)^2 \right) + \text{const} \\
&= -\frac{1}{2} \log \beta_0^{-1} - \frac{\beta_0}{2} \theta^2 + \sum_{n=1}^N r_{n1} \left( -\frac{1}{2} x_n^2 \right) + r_{n2} \left( -\frac{1}{2} (x_n - \theta)^2 \right) + \text{const} \\
&= -\frac{\beta_0}{2} \theta^2 + \sum_{n=1}^N -\frac{r_{n2}}{2} (x_n^2 - 2x_n\theta + \theta^2) + \text{const} \\
&= -\frac{1}{2} \left( \left( \beta_0 + \sum_{n=1}^N r_{n2} \right) \theta^2 + \sum_{n=1}^N r_{n2} x_n^2 - 2\theta \sum_{n=1}^N r_{n2} x_n \right) + \text{const} \\
&= -\frac{\beta_0 + \sum_{n=1}^N r_{n2}}{2} \left( \theta - \frac{1}{\beta_0 + \sum_{n=1}^N r_{n2}} \sum_{n=1}^N r_{n2} x_n \right)^2 + \text{const} \tag{9}
\end{aligned}$$

Again, we exponentiate both sides of 9 and recognize this as

$$q^*(\theta) = N(\theta | m_2, \beta_2^{-1}) \tag{10}$$

with

$$\beta_2 = \beta_0 + N_2 \quad \text{and} \quad m_2 = \beta_2^{-1} N_2 \bar{x}_2$$

where we have defined

$$\bar{x}_2 = \frac{1}{N_2} \sum_{n=1}^N r_{n2} x_n$$

We can use the distribution 10 to compute  $\mathbb{E}_\theta[(x_n - \theta)^2]$ , needed when updating  $q(\mathbf{z})$ :

$$\begin{aligned}
\mathbb{E}_\theta[(x_n - \theta)^2] &= \mathbb{E}_\theta[(x_n - m_2 + m_2 - \theta)^2] \\
&= (x_n - m_2)^2 + 2(x_n - m_2)\mathbb{E}[m_2 - \theta] + \mathbb{E}[(m_2 - \theta)^2] \\
&= (x_n - m_2)^2 + \beta_2^{-1} \tag{11}
\end{aligned}$$

The overall VB algorithm is obtained by cycling through updating:

- The responsibilities  $r_{nk}$  using formulas 3, 4, 5
- The terms 11 needed when computing the responsibilities
- The term 7 and 8 needed when computing the responsibilities

## 1.2 Derivation of ELBO for the Simple Model

Recall that variational inference is based on the decomposition.

$$\log p(x) = \mathcal{L}(q) + \text{KL}[q|p]$$

where  $q(\mathbf{Z})$  is any approximation to the posterior distribution  $p(\mathbf{Z}|\mathbf{X})$  of the unobserved variables  $\mathbf{Z}$  in the model, given the observed variables  $\mathbf{X}$ . The goal of the variational inference algorithm is to maximize the evidence lower bound (ELBO)  $\mathcal{L}(q)$ , or equivalently minimize the KL-divergence  $\text{KL}[q|p]$  between the approximation and the true posterior. Here, we show how to compute the ELBO for the "simple model" derived earlier. Briefly, the model is

$$p(x_n|\theta, \tau) = (1 - \tau)N(x_n|0, 1) + \tau N(x_n|\theta, 1) \quad n = 1, \dots, N$$

The latent variable representation is given by

$$p(\mathbf{x}|\mathbf{z}, \theta) = \prod_{n=1}^N N(x_n|0, 1)^{z_{n1}} N(x_n|\theta, 1)^{z_{n2}} \quad (12)$$

and

$$p(\mathbf{z}|\tau) = \prod_{n=1}^N \tau^{z_{n2}} (1 - \tau)^{z_{n1}} \quad (13)$$

Priors are specified as follows

$$\begin{aligned} p(\tau) &= \text{Beta}(\tau|\alpha_0, \beta_0) \propto \tau^{\alpha_0-1} (1 - \tau)^{\beta_0-1} \\ p(\theta) &= N(\theta|0, \beta_0^{-1}) \propto \exp\left(-\frac{\beta_0}{2}\theta^2\right) \end{aligned}$$

The logarithm of the joint distribution can be written as:

$$\log p(\mathbf{x}, \mathbf{z}, \tau, \theta) = \log p(\tau) + \log p(\theta) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z}, \theta) \quad (14)$$

We assume the mean-field approximation.

$$p(\mathbf{z}, \tau, \theta|\mathbf{x}) \approx q(\tau)q(\theta) \prod_n q(z_n) \quad (15)$$

Assume that currently, we have factors.

$$q(z_n|r_{n1}, r_{n2}) = \text{Categorical}(z_n|r_{n1}, r_{n2}) = r_{n1}^{z_{n1}} r_{n2}^{z_{n2}} \quad (16)$$

$$q(\tau) = \text{Beta}(\tau|\alpha_\tau, \beta_\tau) \quad (17)$$

$$q(\theta) = N(\theta|m_2, \beta_2^{-1}) \quad (18)$$

where  $r_{n1}, r_{n2}, n = 1, \dots, N, \alpha_\tau, \beta_\tau, m_2, \beta_2$  are so-called variational parameters, i.e., parameters that specify the exact distribution of the factor. The general formula of ELBO is given by

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}[\log q(\mathbf{Z})]
\end{aligned} \tag{19}$$

where  $\mathbf{Z}$  is a generic notation that includes all unobservables. We then rewrite ELBO as follows:

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_{q(\tau)q(\theta)q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}, \tau, \theta)] - \mathbb{E}_{q(\tau)q(\theta)q(\mathbf{z})}[\log q(\tau)q(\theta)q(\mathbf{z})] \\
&= \mathbb{E}_{q(\tau)q(\theta)q(\mathbf{z})}[\log p(\tau) + \log p(\theta) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z}, \theta)] \\
&\quad - \mathbb{E}_{q(\tau)q(\theta)q(\mathbf{z})}[\log q(\tau) + \log q(\theta) + \log q(\mathbf{z})] \\
&= \mathbb{E}_{q(\tau)}[\log p(\tau)] + \mathbb{E}_{q(\theta)}[\log p(\theta) + \mathbb{E}_{q(\tau)q(\mathbf{z})}[\log p(\mathbf{z}|\tau)]] + \mathbb{E}_{q(\mathbf{z})q(\theta)}[\log p(\mathbf{x}|\mathbf{z}, \theta)] \\
&\quad - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\tau)}[\log q(\tau)] - \mathbb{E}_{q(\theta)}[\log q(\theta)]
\end{aligned} \tag{20}$$

As with the simple model, all seven terms in formula 20 can be computed analytically when conjugate priors are used. Below, we consider each of these terms. The ELBO can be computed simply by plugging each derived term into Equation 20. In these derivations, we will occasionally discard some terms that do not depend on the variational parameters, as our purpose of deriving the ELBO is to monitor the convergence of the VB algorithm, and those terms are constant across the iterations.

**1st term in Equation 20:**

$$\begin{aligned}
\mathbb{E}_{q(\tau)}[\log p(\tau)] &= \mathbb{E}_{q(\tau)}[(\alpha_0 - 1) \log \tau + (\alpha_0 - 1) \log(1 - \tau)] \\
&= (\alpha_0 - 1) \mathbb{E}_{q(\tau)}[\log \tau] + (\alpha_0 - 1) \mathbb{E}_{q(\tau)}[\log(1 - \tau)] \\
&= (\alpha_0 - 1) [\psi(\alpha_\tau) - \psi(\alpha_\tau + \beta_\tau)] + (\alpha_0 - 1) [\psi(\beta_\tau) - \psi(\alpha_\tau + \beta_\tau)]
\end{aligned}$$

**2nd term in Equation 20:**

$$\begin{aligned}
\mathbb{E}_{q(\theta)}[\log p(\theta)] &= \mathbb{E}_{q(\theta)} \left[ -\frac{\beta_0}{2} \theta^2 \right] \\
&= -\frac{\beta_0}{2} (\mathbb{V}[\theta] + \mathbb{E}[\theta]^2) \\
&= -\frac{\beta_0}{2} (\beta_2^{-1} + m_2^2)
\end{aligned}$$

**3rd term in Equation 20:**

$$\begin{aligned}
\mathbb{E}_{q(\tau)q(\mathbf{z})}[\log p(\mathbf{z}|\tau)] &= \sum_{n=1}^N \mathbb{E}_{q(\tau)q(z_n)}[\log p(z_n|\tau)] \\
&= \sum_{n=1}^N \mathbb{E}_{q(\tau)q(z_n)}[z_{n2} \log \tau + z_{n1} \log(1 - \tau)] \\
&= \sum_{n=1}^N \mathbb{E}_{q(z_n)}[z_{n2}] \mathbb{E}_{q(\tau)}[\log \tau] + \mathbb{E}_{q(z_n)}[z_{n1}] \mathbb{E}_{q(\tau)}[\log(1 - \tau)] \\
&= \sum_{n=1}^N r_{n2}[\psi(\alpha_\tau) - \psi(\alpha_\tau + \beta_\tau)] + r_{n1}[\psi(\beta_\tau) - \psi(\alpha_\tau + \beta_\tau)]
\end{aligned}$$

**4th term in Equation 20:**

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})q(\theta)}[\log p(\mathbf{x}|\mathbf{z}, \theta)] &= \sum_{n=1}^N \mathbb{E}_{q(z_n)q(\theta)} \left[ -\frac{z_{n1}}{2} (\log 2\pi + x_n^2) \frac{z_{n2}}{2} (\log 2\pi + (x_n - \theta)^2) \right] \\
&= \sum_{n=1}^N -\mathbb{E}_{q(z_n)}[z_{n1}] \mathbb{E}_{q(\theta)} \left[ \frac{1}{2} (\log 2\pi + x_n^2) \right] - \mathbb{E}_{q(z_n)}[z_{n2}] \mathbb{E}_{q(\theta)} \left[ \frac{1}{2} (\log 2\pi + (x_n - \theta)^2) \right] \\
&= -\frac{1}{2} \sum_{n=1}^N r_{n1} \log 2\pi - \frac{1}{2} \sum_{n=1}^N r_{n1} x_n^2 - \frac{1}{2} \sum_{n=1}^N r_{n2} \log 2\pi - \frac{1}{2} \sum_{n=1}^N r_{n2} \mathbb{E}_{q(\theta)}[(x_n - \theta)^2] \\
&= -\frac{1}{2} \sum_{n=1}^N (r_{n1} + r_{n2}) \log 2\pi - \frac{1}{2} \sum_{n=1}^N r_{n1} x_n^2 - \frac{1}{2} \sum_{n=1}^N r_{n2} (\mathbb{E}_{q(\theta)}[(x_n - \theta)^2] + \mathbb{V}[x_n - \theta]) \\
&= -\frac{1}{2} \sum_{n=1}^N (r_{n1} + r_{n2}) \log 2\pi - \frac{1}{2} \sum_{n=1}^N r_{n1} x_n^2 - \frac{1}{2} \sum_{n=1}^N r_{n2} ((x_n - \mathbb{E}_{q(\theta)}[\theta])^2 + \mathbb{V}[\theta]) \\
&= -\frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{n=1}^N r_{n1} x_n^2 - \frac{1}{2} \sum_{n=1}^N r_{n2} ((x_n - m_2)^2 + \beta_2^{-1})
\end{aligned}$$

**5th term in Equation 20:**

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] &= \sum_{n=1}^N \mathbb{E}_{q(z_n)}[z_{n1} \log r_{n1} + z_{n2} \log r_{n2}] \\
&= \sum_{n=1}^N r_{n1} \log r_{n1} + r_{n2} \log r_{n2}
\end{aligned}$$

**6th term in Equation 20:**

$$\mathbb{E}_{q(\tau)}[\log q(\tau)] = \log \frac{\Gamma(\alpha_\tau + \beta_\tau)}{\Gamma(\alpha_\tau)\Gamma(\beta_\tau)} + (\alpha_\tau - 1)\psi(\alpha_\tau) + (\beta_\tau - 1)\psi(\beta_\tau) - (\alpha_\tau + \beta_\tau - 2)\psi(\alpha_\tau + \beta_\tau)$$

This is just the negative entropy of  $Beta(\alpha_\tau, \beta_\tau)$

**7th term in Equation 20:**

By the definition of the negative entropy of normal distribution, we obtain

$$\mathbb{E}_{q(\theta)}[\log q(\theta)] = -\frac{1}{2} \log(2\pi e\beta_2^{-1})$$