Solutions to Problem Sheet 7

1. **Mixture model and latent variables**

Let $(a,b) \in (\mathbb{R}_+^*)^2$, $(m_0, m_1) \in \mathbb{R}^2$, $(\sigma_0, \sigma_1) \in (\mathbb{R}_+^*)^2$ be known and fixed. Consider the following Bayesian scheme: the prior distribution on $\boldsymbol{\theta}$ is $\Pi = \text{Beta}(a,b)$. Given $\boldsymbol{\theta} = \theta$, the observations $\mathbf{X} = (X_1, \ldots, X_N)$ are i.i.d. according to the Gaussian mixture with density

$$p_\theta(x) = (1-\theta)\phi_0(x) + \theta\phi_1(x) = (1-\theta) \times \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-m_0)^2}{2\sigma_0^2}} + \theta \times \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}}.$$

(a) Determine the posterior distribution, i.e., $\Pi(\boldsymbol{\theta}|\mathbf{X})$. Is it a standard distribution?

Solution: Denoting as usual $\pi(\theta|\mathbf{X})$ as the density of the posterior distribution, we have

$$\begin{aligned}\pi(\theta|\mathbf{X}) &\propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{]0,1[}(\theta) \prod_{j=1}^N p_\theta(X_j) \\ &= \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{]0,1[}(\theta) \prod_{j=1}^N \left\{(1-\theta)\phi_0(X_j) + \theta\phi_1(X_j)\right\},\end{aligned}$$

which is not a standard distribution. In particular, it is not a Beta distribution. In other words, the family of Beta distributions is not conjugate for the Gaussian mixture model.

(b) Consider the following hierarchical Bayesian scheme: the prior distribution on $\boldsymbol{\theta}$ is still $\Pi = \text{Beta}(a,b)$. Then, given $\boldsymbol{\theta} = \theta$, the variables $\mathbf{Z} = (Z_1, \ldots, Z_N) \in \{0,1\}^N$ are i.i.d. according to the Bernoulli distribution $\text{Ber}(\theta)$. Finally, given $\mathbf{Z} = \mathbf{z}$ and $\boldsymbol{\theta} = \theta$, the observations $\mathbf{X} = (X_1, \ldots, X_N)$ are independent with

$$\pi(X_j|\mathbf{Z} = \mathbf{z}, \boldsymbol{\theta} = \theta) = \pi(X_j|Z_j = z_j, \boldsymbol{\theta} = \theta) = \pi(X_j|Z_j = z_j) = \mathcal{N}(m_{z_j}, \sigma_{z_j}^2).$$

The random variables Z_j are called latent variables (because they are unobserved).

i. Prove that, given $\boldsymbol{\theta}$, the observations X_j all have the same distribution (corresponding to the Gaussian mixture model from Question 1) and are independent. What do you deduce for the posterior distribution $\Pi(\boldsymbol{\theta}|\mathbf{X})$ of this hierarchical Bayesian scheme?

Solution: The joint distribution of $(\mathbf{Z}, \mathbf{X}) \mid \boldsymbol{\theta}$ is given by $\pi(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})\pi(\mathbf{Z} \mid \boldsymbol{\theta})$, which is also equal to

$$\prod_{i=1}^n \pi(X_j|Z_j) \prod_{i=1}^n \pi(Z_j \mid \boldsymbol{\theta}),$$

using the conditional independence of the observations $X_1, \ldots, X_n$ given $\mathbf{Z}$ and $\boldsymbol{\theta}$. Integrating over the latent variables $Z_1, \ldots, Z_N \mid \boldsymbol{\theta} \overset{i.i.d.}{\sim} \text{Ber}(\boldsymbol{\theta})$, we obtain that the density of $\mathbf{X} \mid \boldsymbol{\theta}$ is given by

$$\begin{aligned}
&\prod_{i=1}^{n} [\pi(X_j|Z_j = 0)\pi(Z_j = 0 \mid \boldsymbol{\theta}) + \pi(X_j|Z_j = 1)\pi(Z_j = 1 \mid \boldsymbol{\theta})] \\
&= \prod_{i=1}^{n} [\pi(X_j|Z_j = 0)(1 - \boldsymbol{\theta}) + \pi(X_j|Z_j = 1)\boldsymbol{\theta}] \\
&= \prod_{i=1}^{n} [\phi_0(X_i)(1 - \boldsymbol{\theta}) + \phi_1(X_i)\boldsymbol{\theta}] = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(X_i),
\end{aligned}$$

proving that observations X_j all have the same conditional distribution $p_{\boldsymbol{\theta}}$ and are comnditionally independent. Hence, the posterior distribution $\Pi(\boldsymbol{\theta}|\mathbf{X})$ is the same as in Question 1.

ii. What is the distribution of $\boldsymbol{\theta}$ given $\mathbf{Z}$? What is the distribution of $\boldsymbol{\theta}$ given $(\mathbf{Z}, \mathbf{X})$?

Solution: We are dealing with a classical Bayesian scheme where the prior on $\boldsymbol{\theta}$ is a Beta distribution and, given $\boldsymbol{\theta}$, the Z_j are i.i.d. Bernoulli with parameter θ. Thus, the law of $\boldsymbol{\theta}$ given $\mathbf{Z}$ admits a density:

$$\begin{aligned}
\pi(\theta|\mathbf{Z}) &\propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{]0,1[}(\theta) \prod_{j=1}^{N} \theta^{Z_j}(1-\theta)^{1-Z_j} \\
&= \theta^{a+N\bar{Z}_N-1}(1-\theta)^{b+N-N\bar{Z}_N-1}\mathbf{1}_{]0,1[}(\theta),
\end{aligned}$$

which means that

$$\Pi(\boldsymbol{\theta}|\mathbf{Z}) = \text{Beta}(a + N\bar{Z}_N, b + N - N\bar{Z}_N).$$

For the distribution of $\boldsymbol{\theta}$ given $(\mathbf{Z}, \mathbf{X})$, let $f(\theta, \mathbf{z}, \mathbf{x})$ denote the joint density. Then,

$$\pi(\theta|\mathbf{z}, \mathbf{x}) = f(\theta|\mathbf{z}, \mathbf{x}) \propto f(\theta, \mathbf{z}, \mathbf{x}) = f(\theta)f(\mathbf{z}|\theta)f(\mathbf{x}|\mathbf{z}, \theta).$$

By construction, $f(\mathbf{x}|\mathbf{z}, \theta) = f(\mathbf{x}|\mathbf{z})$, which does not depend on θ, so we have

$$\pi(\theta|\mathbf{z}, \mathbf{x}) \propto f(\theta)f(\mathbf{z}|\theta) \propto \pi(\theta|\mathbf{z}).$$

That is, $\pi(\boldsymbol{\theta}|\mathbf{Z}, \mathbf{X}) = \pi(\boldsymbol{\theta}|\mathbf{Z})$. Intuitively, if the latent variables Z_j are known, the observations X_j provide no additional information about $\boldsymbol{\theta}$.

iii. What is the distribution of Z_1 given $(\boldsymbol{\theta}, \mathbf{X})$?

Solution: By applying Bayes' theorem directly to the joint density of the observations $\mathbf{X}$ and the latent variable Z_1 given $\boldsymbol{\theta}$, we compute the conditional probability

for $Z_1 = 1$. We first expand the likelihood $p(\mathbf{X} \mid Z_1 = 1, \boldsymbol{\theta})$ using the chain rule of probability.
Given $\boldsymbol{\theta}$, the tuples (X_j, Z_j) are conditionally independent.Therefore, this justifies simplifying $p(X_i \mid X_{j<i}, Z_1 = 1, \boldsymbol{\theta})$ to $p(X_i \mid \boldsymbol{\theta})$ below. We apply the same logic to expand the marginal likelihood $p(\mathbf{X} \mid \boldsymbol{\theta})$ in the denominator using the law of total probability over $Z_1 \in \{0, 1\}$.

$$\begin{aligned}
P(Z_1 = 1 \mid \boldsymbol{\theta}, \mathbf{X}) &= \frac{p(\mathbf{X} \mid Z_1 = 1, \boldsymbol{\theta}) P(Z_1 = 1 \mid \boldsymbol{\theta})}{p(\mathbf{X} \mid \boldsymbol{\theta})} \\
&= \frac{p(X_1 \mid Z_1 = 1, \boldsymbol{\theta}) \left(\prod_{i=2}^{N} p(X_i \mid X_{j<i}, Z_1 = 1, \boldsymbol{\theta}) \right) P(Z_1 = 1 \mid \boldsymbol{\theta})}{p(\mathbf{X} \mid \boldsymbol{\theta})} \\
&= \frac{p(X_1 \mid Z_1 = 1, \boldsymbol{\theta}) \left(\prod_{i=2}^{N} p(X_i \mid \boldsymbol{\theta}) \right) \theta}{p(X_1 \mid Z_1 = 0, \boldsymbol{\theta}) \left(\prod_{i=2}^{N} p(X_i \mid \boldsymbol{\theta}) \right) (1 - \theta) + p(X_1 \mid Z_1 = 1, \boldsymbol{\theta}) \left(\prod_{i=2}^{N} p(X_i \mid \boldsymbol{\theta}) \right) \theta} \\
&= \frac{\theta \phi_1(X_1)}{(1 - \theta)\phi_0(X_1) + \theta \phi_1(X_1)}.
\end{aligned}$$

Since $Z_1 \in \{0, 1\}$, it follows a Bernoulli distribution with the derived parameter:

$$Z_1 \mid \boldsymbol{\theta}, \mathbf{X} \sim \text{Ber}\left(\frac{\theta \phi_1(X_1)}{(1 - \theta)\phi_0(X_1) + \theta \phi_1(X_1)} \right).$$

iv. What is the distribution of Z_1 given $(\boldsymbol{\theta}, \mathbf{X}, Z_2, \ldots, Z_N)$?

Solution: For the same reasons as previously, we have:

$$Z_1 | \boldsymbol{\theta}, \mathbf{X}, Z_2, \ldots, Z_N \sim Z_1 | \boldsymbol{\theta}, \mathbf{X} \sim Z_1 | \boldsymbol{\theta}, X_1.$$

v. Deduce a Gibbs sampler allowing to construct, given $\mathbf{X}$, a Markov chain $((\boldsymbol{\theta}_n, \mathbf{Z}_n))_{n \geq 0}$ admitting as a stationary distribution the posterior distribution of the pair $(\boldsymbol{\theta}, \mathbf{Z})$, i.e., Distribution$((\boldsymbol{\theta}, \mathbf{Z}) | \mathbf{X})$.

Solution: We start from an arbitrary initial condition $(\boldsymbol{\theta}_0, \mathbf{Z}_0)$. Given $(\boldsymbol{\theta}_n, \mathbf{Z}_n) = (\theta, \mathbf{z})$, a step of the sequential Gibbs sampler functions as follows:

- Simulate $\theta' \sim \text{Beta}(a + N\bar{Z}_N, b + N - N\bar{Z}_N)$.
- Simulate $Z_1' \sim \text{Ber}\left(\frac{\theta' \phi_1(X_1)}{(1-\theta')\phi_0(X_1) + \theta' \phi_1(X_1)} \right)$.
- $\ldots$
- Simulate $Z_N' \sim \text{Ber}\left(\frac{\theta' \phi_1(X_N)}{(1-\theta')\phi_0(X_N) + \theta' \phi_1(X_N)} \right)$.
- Set $(\boldsymbol{\theta}_{n+1}, \mathbf{Z}_{n+1}) = (\theta', \mathbf{z}')$.

vi. Is the sequence of random variables $(\boldsymbol{\theta}_n)$ a Markov chain?

Solution: Yes, and this is a general result for the Gibbs sampler. In our particular case, $\boldsymbol{\theta}_{n+1}$ is obtained by simulation of a Beta law whose parameters depend on $\mathbf{Z}_n$, and the vector $\mathbf{Z}_n$ is itself obtained by simulations of Bernoulli variables whose parameters depend on $\boldsymbol{\theta}_n$. Consequently, one can write $\boldsymbol{\theta}_{n+1} = f(\boldsymbol{\theta}_n, U_{n+1})$ for an i.i.d. sequence U_{n+1}, proving that $(\boldsymbol{\theta}_n)$ is a Markov chain.

2. **Variational Inference for Robust Regression**

In standard linear regression, the assumption of Gaussian noise makes the model highly sensitive to outliers. To overcome this issue, a Student-t distribution is often used to model the noise, which corresponds to robust regression. This model can be elegantly formulated using latent variables.

Consider a dataset with covariates $\mathbf{X} \in \mathbb{R}^{N\times p}$ and continuous responses $\mathbf{y} \in \mathbb{R}^N$ (where $\mathbf{x}_i \in \mathbb{R}^p$ denotes the i-th row of $\mathbf{X}$). We assume the following generative hierarchical model:

- **Prior on the weights:** $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$, with $\boldsymbol{\Sigma}_0$ a fixed symmetric positive-definite covariance matrix.
- **Latent precision variables:** For each observation $i \in \{1, \dots, N\}$, we introduce a local precision $\tau_i \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$, where $\nu > 0$ is the known and fixed number of degrees of freedom.
- **Conditional likelihood:** Given $\boldsymbol{\beta}$ and τ_i, the observation y_i follows a normal distribution:

$$p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \tau_i) = \mathcal{N}\left(y_i; \mathbf{x}_i^T\boldsymbol{\beta}, \frac{1}{\tau_i}\right).$$

(a) Write down the expression, up to an additive constant, of the joint log-density $\log p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{X})$, where $\boldsymbol{\tau} = (\tau_1, \dots, \tau_N)$. Briefly explain why the exact posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}, \mathbf{X})$ (and its marginal $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$) is analytically intractable.

Solution: The joint log-density expands as:

$$\log p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{X}) = \log p(\boldsymbol{\beta}) + \sum_{i=1}^N \log p(\tau_i) + \sum_{i=1}^N \log p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \tau_i).$$

Dropping constants with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, we get:

$$\log p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{X}) \propto -\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta} + \sum_{i=1}^N \left(\left(\frac{\nu}{2} - 1\right)\log\tau_i - \frac{\nu}{2}\tau_i\right) + \sum_{i=1}^N \left(\frac{1}{2}\log\tau_i - \frac{\tau_i}{2}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right).$$

The exact posterior is intractable because the interaction term $\tau_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$ closely couples $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. Integrating over either variable does not yield a standard closed-form distribution, meaning the normalization constant cannot be computed analytically.

(b) We want to approximate the true posterior with a factorized variational distribution (mean-field approach) of the form:

$$q(\boldsymbol{\beta}, \boldsymbol{\tau}) = q(\boldsymbol{\beta}) \prod_{i=1}^{N} q(\tau_i).$$

Recall the general expression of the optimal pseudo-density $q^*(\boldsymbol{\beta})$ derived from the Coordinate Ascent Variational Inference (CAVI) algorithm as a function of the expectation over the other latent variables.

Solution: The general CAVI update states that the optimal variational distribution for a set of variables is proportional to the exponentiated expected log-joint density, where the expectation is taken over all other variables. Thus:

$$q^*(\boldsymbol{\beta}) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\tau})}[\log p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{X})]\right).$$

(c) Applying this formula, show that the optimal variational distribution $q^*(\boldsymbol{\beta})$ is a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$. Explicitly express the precision matrix $\boldsymbol{\Sigma}_N^{-1}$ and the mean vector $\boldsymbol{\mu}_N$ as a function of $\boldsymbol{\Sigma}_0$, $\mathbf{X}$, $\mathbf{y}$, and the expectations $\mathbb{E}_{q(\tau_i)}[\tau_i]$.

Solution: Using the joint log-density expression and isolating terms that depend on $\boldsymbol{\beta}$:

$$\begin{aligned}
\log q^*(\boldsymbol{\beta}) &= \mathbb{E}_{q(\boldsymbol{\tau})}\left[-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta} - \frac{1}{2}\sum_{i=1}^{N}\tau_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right] + \text{const} \\
&= -\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta} - \frac{1}{2}\sum_{i=1}^{N}\mathbb{E}[\tau_i](y_i^2 - 2y_i\mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\beta}) + \text{const} \\
&= -\frac{1}{2}\boldsymbol{\beta}^T\left(\boldsymbol{\Sigma}_0^{-1} + \sum_{i=1}^{N}\mathbb{E}[\tau_i]\mathbf{x}_i\mathbf{x}_i^T\right)\boldsymbol{\beta} + \boldsymbol{\beta}^T\left(\sum_{i=1}^{N}\mathbb{E}[\tau_i]y_i\mathbf{x}_i\right) + \text{const}.
\end{aligned}$$

This is the exponent of a multivariate Gaussian. Matching the coefficients of $\boldsymbol{\beta}$, we define a diagonal weight matrix $\mathbf{W} = \text{diag}(\mathbb{E}[\tau_1], \ldots, \mathbb{E}[\tau_N])$. Then:

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T\mathbf{W}\mathbf{X}$$

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N\mathbf{X}^T\mathbf{W}\mathbf{y} = \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.$$

(d) Similarly, determine the form of the optimal variational distribution $q^*(\tau_i)$ for $i \in \{1, \ldots, N\}$. Show that it is a Gamma(a_i, b_i) distribution and give the expressions for its shape parameter a_i and rate parameter b_i.

Solution: Applying the CAVI update for τ_i, we isolate terms in the joint log-density

that depend on τ_i:

$$\log q^*(\tau_i) = \mathbb{E}_{q(\boldsymbol{\beta})}\left[\left(\frac{\nu}{2}-1\right)\log\tau_i - \frac{\nu}{2}\tau_i + \frac{1}{2}\log\tau_i - \frac{\tau_i}{2}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right] + \text{const}$$
$$= \left(\frac{\nu+1}{2}-1\right)\log\tau_i - \tau_i\left(\frac{\nu}{2} + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\beta})}[(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2]\right) + \text{const.}$$

We recognize the log-density of a Gamma distribution $\text{Gamma}(a_i, b_i)$, where:

$$a_i = \frac{\nu+1}{2}$$

$$b_i = \frac{\nu}{2} + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\beta})}[(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2].$$

(e) Recall that for a variable $Z \sim \text{Gamma}(a, b)$, its expectation is $\mathbb{E}[Z] = a/b$, and for a random vector $\mathbf{V} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the expectation of a scalar quadratic form is $\mathbb{E}[(\mathbf{c}^T\mathbf{V})^2] = (\mathbf{c}^T\boldsymbol{\mu})^2 + \mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}$. Using these facts, summarize in a few lines the steps of an iterative algorithm to jointly optimize the parameters of $q(\boldsymbol{\beta})$ and $q(\tau_i)$ until convergence.

Solution: 1. Initialize the expected precisions $\mathbb{E}[\tau_i] = 1$ for all $i \in \{1, \ldots, N\}$.
2. **Update $q(\boldsymbol{\beta})$:** Compute the covariance matrix $\boldsymbol{\Sigma}_N$ and mean vector $\boldsymbol{\mu}_N$ using the current values of $\mathbb{E}[\tau_i]$ via the formulas derived in part (c).
3. **Update $q(\tau_i)$:** For each observation i, compute the expected squared residual:

$$E_i = \mathbb{E}_{q(\boldsymbol{\beta})}[(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2] = (y_i - \mathbf{x}_i^T\boldsymbol{\mu}_N)^2 + \mathbf{x}_i^T\boldsymbol{\Sigma}_N\mathbf{x}_i.$$

Then, update the expectation $\mathbb{E}[\tau_i] = \frac{a_i}{b_i} = \frac{\nu+1}{\nu+E_i}$.
4. Repeat steps 2 and 3 iteratively until the ELBO converges or the parameters stabilize.

3. **Mean-Field Variational Inference for Bayesian Linear Regression**

Consider a standard Bayesian linear regression model. We are given a dataset with a design matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ and a target vector $\mathbf{y} \in \mathbb{R}^N$. We assume the following generative model:

- **Prior:** $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_p)$, where $\alpha > 0$ is a known precision parameter.
- **Likelihood:** $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}_N)$, where $\beta > 0$ is the known noise precision.

We know that the exact posterior $p(\mathbf{w} \mid \mathbf{y}, \mathbf{X})$ is a multivariate Gaussian. However, for pedagogical purposes, we will approximate it using a fully factorized mean-field variational distribution:

$$q(\mathbf{w}) = \prod_{j=1}^{p} q_j(w_j)$$

(a) The exact posterior is $p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \mathbf{S})$. Give the explicit expression for the precision matrix $\mathbf{S}^{-1}$. By inspecting $\mathbf{S}^{-1}$, explain briefly why the exact posterior does not naturally factorize across the components of $\mathbf{w}$.

Solution: By Bayes' rule, the posterior precision matrix is the sum of the prior precision and the data precision:

$$\mathbf{S}^{-1} = \alpha \mathbf{I}_p + \beta \mathbf{X}^T \mathbf{X}.$$

The exact posterior does not factorize across the components w_j because the off-diagonal elements of $\mathbf{S}^{-1}$ (which are given by $\beta \mathbf{x}_j^T \mathbf{x}_k$ for $j \neq k$) are generally non-zero. These non-zero elements introduce correlations between the different weights in the posterior distribution.

(b) The optimal distribution $q_j^*(w_j)$ satisfies:

$$\log q_j^*(w_j) = \mathbb{E}_{q_{-j}}[\log p(\mathbf{y}, \mathbf{w} \mid \mathbf{X})] + \text{const}$$

where q_{-j} denotes the product of all $q_k(w_k)$ for $k \neq j$. By expanding the joint log-density, show that $q_j^*(w_j)$ takes the form of a univariate Gaussian $\mathcal{N}(\mu_j, s_j^2)$.

Solution: The joint log-density is:

$$\log p(\mathbf{y}, \mathbf{w} \mid \mathbf{X}) \propto -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}).$$

Taking the expectation with respect to q_{-j} and keeping only terms that depend on w_j:

$$\begin{aligned}
\log q_j^*(w_j) &= \mathbb{E}_{q_{-j}}\left[-\frac{\alpha}{2}w_j^2 - \frac{\beta}{2}\sum_{i=1}^{N}\left(y_i - \sum_{k=1}^{p} x_{ik}w_k\right)^2\right] + \text{const} \\
&= -\frac{\alpha}{2}w_j^2 - \frac{\beta}{2}\sum_{i=1}^{N}\mathbb{E}_{q_{-j}}\left[-2y_i x_{ij} w_j + x_{ij}^2 w_j^2 + 2x_{ij}w_j \sum_{k\neq j} x_{ik}w_k\right] + \text{const} \\
&= -\frac{1}{2}\left(\alpha + \beta\sum_{i=1}^{N} x_{ij}^2\right) w_j^2 + \beta \sum_{i=1}^{N} x_{ij}\left(y_i - \sum_{k\neq j} x_{ik}\mu_k\right) w_j + \text{const},
\end{aligned}$$

where $\mu_k = \mathbb{E}_{q_k}[w_k]$. Since this expression is quadratic in w_j, $q_j^*(w_j)$ is a univariate Gaussian $\mathcal{N}(\mu_j, s_j^2)$.

(c) Derive the explicit update equations for the variational variance s_j^2 and the variational mean μ_j.

Solution: A univariate Gaussian $\mathcal{N}(\mu_j, s_j^2)$ has the log-density form: $-\frac{1}{2s_j^2}w_j^2 + \frac{\mu_j}{s_j^2}w_j$. Matching coefficients with the expansion from the previous part:

$$\frac{1}{s_j^2} = \alpha + \beta\sum_{i=1}^{N} x_{ij}^2 = \alpha + \beta\|\mathbf{x}_j\|^2 \implies s_j^2 = \frac{1}{\alpha + \beta\|\mathbf{x}_j\|^2}.$$

For the mean:

$$\frac{\mu_j}{s_j^2} = \beta \sum_{i=1}^{N} x_{ij} \left(y_i - \sum_{k \neq j} x_{ik} \mu_k \right) \implies \mu_j = s_j^2 \beta \mathbf{x}_j^T \left(\mathbf{y} - \sum_{k \neq j} \mu_k \mathbf{x}_k \right).$$

(d) Show that the variational variance s_j^2 depends only on the design matrix and hyperparameters, and does not depend on the variational means μ_k. What does this imply about the order of updates in our CAVI algorithm?

Solution: As seen in the formula $s_j^2 = \frac{1}{\alpha + \beta \|\mathbf{x}_j\|^2}$, the variance relies solely on α, β, and the ℓ_2-norm of the j-th column of $\mathbf{X}$. It does not contain any μ_k terms. This implies that the variances do not need to be updated iteratively; they can be pre-computed once before the iterative CAVI algorithm begins, and only the means μ_j need to be updated during the loop.