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?

   (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. Let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be a test function, show that

$$\mathbb{E}[\varphi(X_1)|\boldsymbol{\theta}] = \int_{\mathbb{R}} \varphi(x) p_{\boldsymbol{\theta}}(x) dx.$$

   What do you deduce for the posterior distribution $\Pi(\boldsymbol{\theta}|\mathbf{X})$ of this hierarchical Bayesian scheme?

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

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

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

   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., $\text{Distribution}((\boldsymbol{\theta}, \mathbf{Z})|\mathbf{X})$.

   vi. Is the sequence of random variables $(\boldsymbol{\theta}_n)$ 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 $i \in \{1, \ldots, N\}$, 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 $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$. 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.

(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.

(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 $\mathbb{E}_{q(\tau_i)}[\tau_i]$.

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

Hint: The parameter a_i will depend on ν, and b_i will depend on ν and $\mathbb{E}_{q(\boldsymbol{\beta})}[(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2]$.

(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.

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, we propose to 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 $\mathbf{w}$.

(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 $\mathcal{N}(\mu_j, s_j^2)$.

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

Hint: Express your answer in terms of the j-th column of the design matrix $\mathbf{x}_j \in \mathbb{R}^N$, the target $\mathbf{y}$, the parameters α and β, and the expectations $\mu_k = \mathbb{E}_{q_k}[w_k]$ for $k \neq j$.

(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?