AI 701 Bayesian machine learning, Fall 2020

Homework assignment 2

1. (100 + 40 points) Consider the following Mixture of Gaussians (MOG) model defined on $\mathbb{R}^d$.

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1. \tag{1}$$

We assume that the covariance matrix Σ_k for each component is decomposed as

$$\Sigma_k = \lambda_k I_d + v_k v_k^\top, \tag{2}$$

where $\lambda_k > 0$, $v \in \mathbb{R}^d$, and I_d is the $d \times d$ identity matrix. The set of parameters θ is defined as

$$\theta = \{\pi \in \mathbb{R}^K, \{\mu_k \in \mathbb{R}^d, \lambda_k \in \mathbb{R}_+, v_k \in \mathbb{R}^d\}_{k=1}^K\}. \tag{3}$$

Assume the following priors,

$$p(\theta) = \mathrm{Dir}(\pi; \mathbb{1}_K) \prod_{k=1}^{K} \mathcal{N}(\mu_k; \mathbb{0}_d, 5.0 \cdot I_d) \cdot \log\mathcal{N}(\lambda_k; 0.1, 0.1) \cdot \mathcal{N}(v_k; \mathbb{0}_d, 0.25 \cdot I_d), \tag{4}$$

where $\mathbb{1}_K = \underbrace{[1, \dots, 1]}_{K}^\top$, $\mathbb{0}_d = \underbrace{[0, \dots, 0]}_{d}^\top$ and $\log\mathcal{N}(x; \mu, \sigma^2)$ is the log-normal distribution with density

$$\log\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(\log x - \mu)^2}{2\sigma^2} \right). \tag{5}$$

Let $X = \{x_i\}_{i=1}^n$ be a set of observed data. For the ease of implementation, we introduce a set of latent variables $Z = \{z_i\}_{i=1}^n$ where $z_i \in \{1, \dots, K\}$. $z_i = k$ indicates that the observation x_i was generated from kth component. The joint likelihood is then written as,

$$p(X, Z, \theta) = p(\theta) \cdot \prod_{i=1}^{n} \prod_{k=1}^{K} \Big(\pi_k \mathcal{N}(x_i|\mu_k, \lambda_k I_d + v_k v_k^\top) \Big)^{\mathbb{1}_{\{z_i = k\}}}, \tag{6}$$

where $\mathbb{1}_{\{z_i=k\}} = 1$ only if $z_i = k$ and zero otherwise. Your goal is to implement a sampler conducting the posterior inference for $p(\theta, Z|X)$.

(a) (5 points) Derive the conditional distribution

$$p(z_i|Z \setminus \{z_i\}, X, \theta), \tag{7}$$

and explain how to sample from it.

Solution: If we collect the terms related to z_i from the joint likelihood $p(X, Z, \theta)$, we get

$$p(z_i|Z \setminus \{z_i\}, X, \theta) \propto \prod_{k=1}^{K} \Big(\pi_k \mathcal{N}(x_i|\mu_k, \lambda_k I_d + v_k v_k^\top) \Big)^{\mathbb{1}_{\{z_i = k\}}}. \tag{8}$$

Hence, we see that the distribution is a categorical distribution with probability

$$p(z_i = k | Z \setminus \{z_i\}, X, \theta) = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \lambda_k I_d + v_k v_k^\top)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}(x_i | \mu_\ell, \lambda_\ell I_d + v_\ell v_\ell^\top)}. \tag{9}$$

We can sample from this using typical categorical sampling method.

(b) (5 points) Derive the conditional distribution

$$p(\pi | X, Z, \theta \setminus \{\pi\}), \tag{10}$$

and explain how to sample from it.

Solution: The terms containing π are

$$p(\pi | X, Z, \theta \setminus \{\pi\}) \propto p(\pi) \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{\mathbb{1}_{\{z_i = k\}}} = \prod_{k=1}^{k} \pi_k^{1 + n_k - 1}, \tag{11}$$

where $n_k := \sum_{i=1}^{n} \mathbb{1}_{\{z_i = k\}}$. Hence, we get that

$$p(\pi | X, Z, \theta \setminus \{\pi\}) = \mathrm{Dir}(\pi; \mathbb{1}_K + [n_1, \ldots, n_K]^\top). \tag{12}$$

We can easily sample from this using standard sampling technique for Dirichlet distribution (e.g., sampling gamma random variables and normalizing them).

(c) (10 points) The posterior for (μ_k, λ_k, v_k) is not easily simulated via Gibbs sampling, so we will use the Metropolis-Hastings algorithm. Consider the following random-work proposal distribution,

$$q(\mu_k', \lambda_k', v_k' | \mu_k, \lambda_k, v_k) = \mathcal{N}(\mu_k'; \mu_k, \sigma_q^2 I_d) \log \mathcal{N}(\lambda_k'; \log \lambda_k, \sigma_q^2) \mathcal{N}(v_k'; v_k, \sigma_q^2 I_d). \tag{13}$$

Compute the acceptance probability for updating (μ_k, λ_k, v_k).

Solution:

$$
\begin{aligned}
\rho &= \frac{p(X, Z, \theta \setminus \{\mu_k, \lambda_k, v_k\}, \mu_k', \lambda_k', v_k')q(\mu_k, \lambda_k, v_k|\mu_k', \lambda_k', v_k')}{p(X, Z, \theta)q(\mu_k', \lambda_k', v_k'|\mu_k, \lambda_k, v_k)} \\
&= \prod_{z_i=k} \frac{\mathcal{N}(x_i|\mu_k', \lambda_k' I_d + v_k'(v_k')^\top)}{\mathcal{N}(x_i|\mu_k, \lambda_k I_d + v_k v_k^\top)} \\
&\times \frac{\mathcal{N}(\mu_k'; \mathbb{0}_d, 5.0 \cdot I_d) \log \mathcal{N}(\lambda_k'; 0.1, 0.1)\mathcal{N}(v_k'; \mathbb{0}_d, 0.25 \cdot I_d)}{\mathcal{N}(\mu_k; \mathbb{0}_d, 5.0 \cdot I_d) \log \mathcal{N}(\lambda_k; 0.1, 0.1)\mathcal{N}(v_k; \mathbb{0}_d, 0.25 \cdot I_d)} \\
&\times \frac{\mathcal{N}(\mu_k; \mu_k', \sigma_q^2 I_d) \log \mathcal{N}(\lambda_k; \lambda_k', \sigma_q^2)\mathcal{N}(v_k; v_k', \sigma_q^2 I_d)}{\mathcal{N}(\mu_k'; \mu_k, \sigma_q^2 I_d) \log \mathcal{N}(\lambda_k'; \lambda_k, \sigma_q^2)\mathcal{N}(v_k'; v_k, \sigma_q^2 I_d)} \\
&= \prod_{z_i=k} \frac{\mathcal{N}(x_i|\mu_k', \lambda_k' I_d + v_k'(v_k')^\top)}{\mathcal{N}(x_i|\mu_k, \lambda_k I_d + v_k v_k^\top)} \\
&\times \frac{\mathcal{N}(\mu_k'; \mathbb{0}_d, 5.0 \cdot I_d)\lambda_k' \log \mathcal{N}(\lambda_k'; 0.1, 0.1)\mathcal{N}(v_k'; \mathbb{0}_d, 0.25 \cdot I_d)}{\mathcal{N}(\mu_k; \mathbb{0}_d, 5.0 \cdot I_d)\lambda_k \log \mathcal{N}(\lambda_k; 0.1, 0.1)\mathcal{N}(v_k; \mathbb{0}_d, 0.25 \cdot I_d)} \\
&= \prod_{z_i=k} \frac{\mathcal{N}(x_i|\mu_k', \lambda_k' I_d + v_k'(v_k')^\top)}{\mathcal{N}(x_i|\mu_k, \lambda_k I_d + v_k v_k^\top)} \\
&\times \frac{\mathcal{N}(\mu_k'; \mathbb{0}_d, 5.0 \cdot I_d)\mathcal{N}(\log \lambda_k'; 0.1, 0.1)\mathcal{N}(v_k'; \mathbb{0}_d, 0.25 \cdot I_d)}{\mathcal{N}(\mu_k; \mathbb{0}_d, 5.0 \cdot I_d)\mathcal{N}(\log \lambda_k; 0.1, 0.1)\mathcal{N}(v_k; \mathbb{0}_d, 0.25 \cdot I_d)}.
\end{aligned}
\tag{14}
$$

The acceptance probability is then given as $\min\{1, \rho\}$ (You don't have to further expand the densities, it suffices to show that the proposal densities cancel out due to the symmetry).

(d) (80 points) Download the file `X.txt` attached. Set X to be the 2D data ($d = 2$) written in `X.txt`. Write a sampler simulating $p(\theta, Z|X)$ via the Gibbs + Metropolis-Hastings with the sampling strategies described above, while fixing the number of components $K = 3$. You can use any scientific programming language you like. Along with the code, you should submit a report describing your implementation and explaining the result. Especially, your report should convince that your sampler works properly. You may show the trace plots of $\log p(X, Z, \theta)$, the clustering induced by the mixture assignments Z after convergence, or the estimated parameters.

(e) (30 points) (Bonus point) Consider marginalizing out the parameter π to work with

$$p(X, Z, \phi) = \int p(X, Z, \theta)\mathrm{d}\pi, \tag{15}$$

where

$$\phi = \{\mu_k, \lambda_k, v_k\}_{k=1}^K. \tag{16}$$

Derive $p(X, Z, \phi)$ and implement a sampler for $p(\phi, Z|X)$ using the same data (`X.txt`).

(f) (10 points) (Bonus point) Write a code to measure effective sample sizes and report the effective sample sizes for the parameter λ_1.