

## Homework assignment 2

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

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

We assume that the covariance matrix  $\Sigma_k$  for each component is decomposed as

$$\Sigma_k = \lambda_k I_d + v_k v_k^\top, \quad (2)$$

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

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

Assume the following priors,

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

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

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

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  $k$ th component. The joint likelihood is then written as,

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

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

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

and explain how to sample from it.

- (b) (5 points) Derive the conditional distribution

$$p(\pi|X, Z, \theta \setminus \{\pi\}), \quad (8)$$

and explain how to sample from it.

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

$$q(\mu'_k, \lambda'_k, v'_k | \mu_k, \lambda_k, v_k) = \mathcal{N}(\mu'_k; \mu_k, \sigma_q^2 I_d) \log \mathcal{N}(\lambda'_k; \log \lambda_k, \sigma_q^2) \mathcal{N}(v'_k; v_k, \sigma_q^2 I). \quad (9)$$

Compute the acceptance probability for updating  $(\mu_k, \lambda_k, v_k)$ .

- (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  $\pi$  to work with

$$p(X, Z, \phi) = \int p(X, Z, \theta) d\pi, \quad (10)$$

where

$$\phi = \{\mu_k, \lambda_k, v_k\}_{k=1}^K. \quad (11)$$

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  $\lambda_1$ .