CSE 291-D: Homework 3

Hao-en Sung [A53204772] (wrangle1005@gmail.com)
Department of Computer Science
University of California, San Diego
San Diego, CA 92092

Problem 1

(a)

The expectation of a mixture of K Gaussians can be derived as follows.

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{x}] &= \int_{-\infty}^{\infty} x \cdot \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \boldsymbol{\Sigma}_k) dx \\
&= \sum_{k=1}^{K} \pi_k \cdot \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} dx \\
&= \sum_{k=1}^{K} \pi_k \cdot \int_{-\infty}^{\infty} (x+\mu_k) \frac{1}{\sqrt{\pi}} e^{-x^2} dx \\
&= \sum_{k=1}^{K} \pi_k \cdot \left(\int_{-\infty}^{\infty} x \frac{1}{\sqrt{\pi}} e^{-x^2} dx + \int_{-\infty}^{\infty} \mu_k \frac{1}{\sqrt{\pi}} e^{-x^2} dx \right) \\
&= \sum_{k=1}^{K} \pi_k \mu_k \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-x^2} dx \\
&= \sum_{k=1}^{K} \pi_k \mu_k
\end{aligned}
$$

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

(b)

The convolution of a mixture of K Gaussians can be derived as follows.

$$\begin{aligned}
\text{cov}[\boldsymbol{x}] &= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\intercal] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\intercal \\
&= \sum_{k=1}^{K} \pi_k \cdot \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} dx - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\intercal \\
&= \sum_{k=1}^{K} \pi_k \cdot \int_{-\infty}^{\infty} (x+\mu_k)^2 \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{x^2}{2\sigma_k^2}} dx - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\intercal \\
&= \sum_{k=1}^{K} \pi_k \cdot \left(\int_{-\infty}^{\infty} x x^\intercal \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{x^2}{2\sigma_k^2}} dx + \int_{-\infty}^{\infty} 2x\mu_k^\intercal \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{x^2}{2\sigma_k^2}} dx \right. \\
&\qquad\qquad \left. + \int_{-\infty}^{\infty} \mu_k^\intercal \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{x^2}{2\sigma_k^2}} dx \right) - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\intercal \\
&= \sum_{k=1}^{K} \pi_k \cdot \left(2\sigma_k\sigma_k^\intercal \cdot \int_{-\infty}^{\infty} x x^\intercal \frac{1}{\sqrt{\pi}} e^{-x^2} dx + \mu_k\mu_k^\intercal \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-x^2} dx \right) - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\intercal \\
&= \left[\sum_{k=1}^{K} \pi_k \cdot (\boldsymbol{\Sigma}_k + \mu_k\mu_k^\intercal) \right] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\intercal
\end{aligned}$$

Problem 2

(a)

The figure is shown as follows.

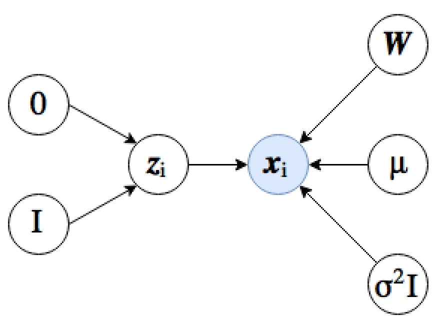

Figure 1: Directed Graphical Model Diagram

(b)

Since the prior and likelihood for PCA are $P(\boldsymbol{z}_i|0, I)$ and $P(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{W}^\intercal, \mu, \sigma^2 I)$, respectively, according to some given formulas derivation in (b), I can have the following notation replacement.

$$\begin{aligned}
\boldsymbol{x} &:= \boldsymbol{z}_i \\
\mu &:= 0 \\
\Lambda^{-1} &:= I \\
\boldsymbol{y} &:= \boldsymbol{z}_i \\
\boldsymbol{A} &:= \boldsymbol{W} \\
\boldsymbol{b} &:= \mu \\
\boldsymbol{L}^{-1} &:= \sigma^2 I
\end{aligned}$$

Then, I can derive the posterior distribution as follows.

$$
\begin{aligned}
p(\boldsymbol{z}_i, \boldsymbol{x}_i, \mu, \boldsymbol{W}, \sigma^2) &= \mathcal{N}(\boldsymbol{z}_i | (I + \boldsymbol{W}^\intercal \frac{I}{\sigma^2} \boldsymbol{W})^{-1} \cdot [\boldsymbol{W}^\intercal \frac{I}{\sigma^2}(\boldsymbol{x}_i - \mu) + 0 \cdot I], (I + \boldsymbol{W}^\intercal \frac{I}{\sigma^2} \boldsymbol{W})^{-1}) \\
&= \mathcal{N}(M^{-1}\boldsymbol{W}^\intercal(\boldsymbol{x}_i - \mu), \sigma^2(I + \boldsymbol{W}^\intercal\boldsymbol{W})^{-1}) \\
&= \mathcal{N}(\boldsymbol{M}^{-1}\boldsymbol{W}^\intercal(\boldsymbol{x}_i - \mu), \sigma^2\boldsymbol{M}^{-1}), \boldsymbol{M} = \boldsymbol{W}^\intercal\boldsymbol{W} + \sigma^2 I
\end{aligned}
$$

(c)

Considering the PCA likelihood for all dataset, I have

$$
\begin{aligned}
P(\boldsymbol{x}|\mu, \boldsymbol{W}, \sigma^2) &= \prod_i P(\boldsymbol{x}_i|\mu, \boldsymbol{W}, \sigma^2) \\
\log P(\boldsymbol{x}|\mu, \boldsymbol{W}, \sigma^2) &= \sum_i \log P(\boldsymbol{x}_i|\mu, \boldsymbol{W}, \sigma^2) \\
&= \sum_i \log \mathcal{N}(\boldsymbol{x}_i|\mu, \boldsymbol{W}\boldsymbol{W}^\intercal + \sigma^2 I) \\
&= \sum_i \left[\log \left(\frac{1}{\sqrt{2\pi(\boldsymbol{W}\boldsymbol{W}^\intercal + \sigma^2 I)}} \right) - \frac{(\boldsymbol{x}_i - \mu)^2}{2\pi(\boldsymbol{W}\boldsymbol{W}^\intercal + \sigma^2 I)} \right] \\
\frac{\partial \log P(\boldsymbol{x}|\mu, \boldsymbol{W}, \sigma^2)}{\partial \mu} &= \sum_i \frac{\boldsymbol{x}_i - \mu}{2\pi(\boldsymbol{W}\boldsymbol{W}^\intercal + \sigma^2 I)}
\end{aligned}
$$

To maximize μ, one can set the derivative of log likelihood for PCA model to zero, which indicates

$$
\begin{aligned}
0 &= \sum_i \frac{-(\boldsymbol{x}_i - \mu)}{2\pi(\boldsymbol{W}\boldsymbol{W}^\intercal + \sigma^2 I)} \\
\mu &= \frac{1}{n} \sum_i \boldsymbol{x}_i.
\end{aligned}
$$

(d)

For $p(\tilde{\boldsymbol{z}}|\theta)$, I have

$$
\begin{aligned}
p(\tilde{\boldsymbol{z}}|\theta) &= p(\tilde{\boldsymbol{z}}|0, I) \\
&= \frac{1}{\sqrt{2\pi I}} e^{-\frac{\tilde{\boldsymbol{z}}^\intercal \tilde{\boldsymbol{z}}}{2\pi I}} \\
&= \frac{1}{\sqrt{2\pi I}} e^{-\frac{\boldsymbol{z}^\intercal R^\intercal R \boldsymbol{z}}{2\pi I}} \\
&= \frac{1}{\sqrt{2\pi I}} e^{-\frac{\boldsymbol{z}^\intercal \boldsymbol{z}}{2\pi I}} \\
&= p(\boldsymbol{z}|\theta).
\end{aligned}
$$

For $P(\boldsymbol{x}|\tilde{\boldsymbol{z}}, \theta)$, I have

$$
\begin{aligned}
P(\boldsymbol{x}|\tilde{\boldsymbol{z}}, \theta) &= P(\boldsymbol{x}|\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{W}}, \mu, \sigma^2 I) \\
&= \frac{1}{\sqrt{2\pi\sigma^2 I}} e^{-\frac{(\boldsymbol{x} - \tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} - \mu)^2}{2\pi\sigma^2 I}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2 I}} e^{-\frac{(\boldsymbol{x} - \boldsymbol{W}R^\intercal R\boldsymbol{z} - \mu)^2}{2\pi\sigma^2 I}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2 I}} e^{-\frac{(\boldsymbol{x} - \boldsymbol{W}\boldsymbol{z} - \mu)^2}{2\pi\sigma^2 I}} \\
&= P(\boldsymbol{x}|\boldsymbol{z}, \theta).
\end{aligned}
$$

For $P(\boldsymbol{x}, \tilde{\boldsymbol{z}}|\theta)$, I have

$$\begin{aligned}P(\boldsymbol{x}, \tilde{\boldsymbol{z}}|\theta) &= p(\tilde{\boldsymbol{z}}|\theta) \cdot P(\boldsymbol{x}|\tilde{\boldsymbol{z}}, \theta) \\ &= p(\boldsymbol{z}|\theta) \cdot P(\boldsymbol{x}|\boldsymbol{z}, \theta) \\ &= P(\boldsymbol{x}, \boldsymbol{z}|\theta).\end{aligned}$$

For $P(\boldsymbol{x}|\theta)$, I have

$$\begin{aligned}P(\boldsymbol{x}|\theta) &= \sum_{z} P(\boldsymbol{x}, \tilde{\boldsymbol{z}}|\theta) \\ &= P(\boldsymbol{x}, \boldsymbol{z}|\theta).\end{aligned}$$

Problem 3

(a)

Model Motivation

In my design, I assume that animal distribution in each state varied greatly from one to another, and thus, different states shall not share parameters with each other. On top of that, I believe there are only N different statuses of animal distribution in one specific state. Thus, node $\boldsymbol{\theta}$ with prior α is a matrix in the shape of $(N, |P| = 9)$, where $\theta^{(Z_i=c)}$ represents the animal distribution of status c; while the value for Z_i is determined by a prior α and previous status Z_{i-1}.

Model Prior and Likelihood

The distribution in this design is listed as follows.

$$\begin{aligned}\theta^{(c)} &\sim \text{Dirichlet}(\alpha), \forall c \\ X_i &= \theta^{Z_i};\end{aligned}$$

while the probability of $Z_i = c'$ from $Z_{i-1} = c$ is given as

$$P(Z_i = c'|Z_{i-1} = c) = M_{c,c'}, \forall i > 1,$$

where M represents the transition of Z_i.

The overall model condition distribution then can be written as

$$P(\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{X}|\alpha) = \prod_{c} P(\theta^{(c)}|\alpha) \cdot P(Z_1) \cdot \prod_{i=2}^{T} P(Z_i|Z_{i-1}) \cdot \prod_{i=2}^{T} \prod_{c} P(X_i|Z_i, \theta^{(c)}).$$

From the above formula, one can tell that it is similar to Hidden Markov Model (HMM), except X_i is now observed as a vector, which is represented by θ^c, instead of a scalar.

Graphical Model Diagram

The plane figure for graphical model is shown as Fig. 2.

(b)

Train Stage

To update this algorithm, I believe Gibbs sampling is a suitable choice.

For the update of $\theta^{(c)}$, it is pretty similar to what I did in the last homework. The parameters for Dirichlet distribution will be proportional to the summation of animal amounts at those c-status time point plus prior α.

The update of Z_i is very similar to the procedure I learned about HMM in class. I can either learn the transition matrix with Gibbs Sampling or forward-backward dynamic programming algorithm.

This same procedure will be applied to every state to learn $|S| = 50$ different models.

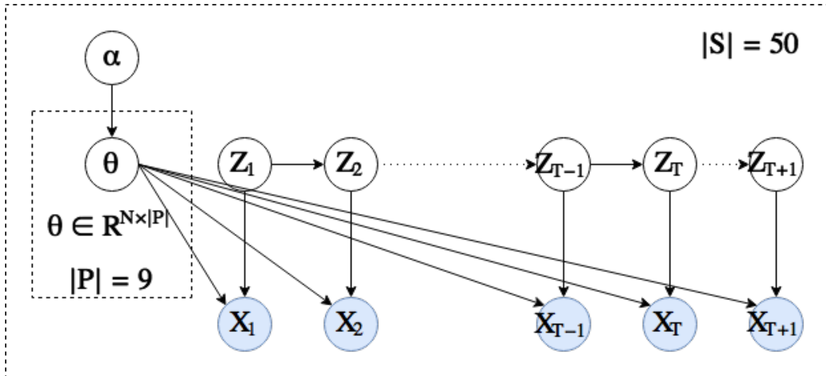

Figure 2: Directed Graphical Model Diagram

Test Stage

Once I learn $\boldsymbol{\theta}$ and Z_1 to Z_T, the probability of $P(Z_{T+1} = c)$ can be calculated. Later, I can estimate X_{T+1} as $\mathbb{E}_{Z_{T+1}}[\theta^{(Z_{T+1})}]$.

This same procedure will be applied to every state to predict $|S| = 50$ different distribution of animals (in numbers).

(c)

Experiment I

[Quantitative] Split the given data into training and validation sets with certain time threshold. After that, I can measure our model performance with Mean Squared Error (MSE) metric.

Experiment II

[Qualitative] Under the assumption that more animals should inhabit in a larger territory, I can regard my predicted animal numbers as a ranking problem sorted by the territory surface and then evaluate the score under Mean Average Precision (MAP) metric.

Experiment III

[Quantitative] In my current model setting, I assume that the number of potential statuses N is far smaller than given number of time points T, i.e. $N << T$. This implies that the number of animals cannot grow proportional to time; otherwise, $N << T$ is meaningless here.

In view of this, I am also interested in only considering the relatively ratio between animals but not the exact number, which might be biased because of model. In other words, I would like to measure the error in terms of the ratio but not the precise number of animals. For this task, I can again utilize Mean Average Precision (MAP) metric to evaluate the model performance.