PGM-Assignment #3

Note: Please use the convention from the sildes to draw graphical representations in these problems.

1. **Mixture Model.** Basic mixture distributin can be formulized by: (n indexes the $n-$th observation.)

$$P(\boldsymbol{x}[n]|\boldsymbol{\theta}) = \sum_{k=1}^{K} P(z[n] = k) p_k(\boldsymbol{x}[n]|\boldsymbol{\theta})$$

The variable $z_i \in \{1, 2, .., K\}$, represents a latent state, and $p_k(\boldsymbol{x}[n]|\boldsymbol{\theta}) = p(\boldsymbol{x}[n]|z[n] = k, \boldsymbol{\theta})$ is called the $k-$th **base distribution** for the observations.

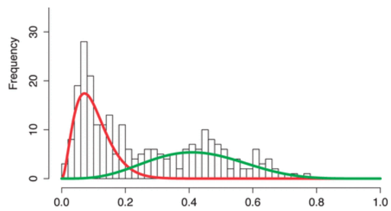
(1) In the class, we have discussed Gaussian Mixture Model (GMM). Here, we consider another example, **Poisson Mixture Model**. In this model, K Poisson distributions $Poisson(\lambda_k)$ are mixed with the proportions $\pi_1, ..., \pi_K$ ($\sum_{k=1}^{K} \pi_k = 1$).

Please draw a graphical representation of this model. (*Hint:* discriminate between parameters and variables.)

(2) Now suppose we've known all parameters: $\boldsymbol{\pi} = \{\pi_k, k = 1, \cdots, K\}$ and $\boldsymbol{\lambda} = \{\lambda_k, k = 1, \cdots, K\}$. For an observation x, please calculate $P(z = k|x), k = 1, \cdots, K$.

(3) **Base distribution.** Different local base distributions could help us model different kinds of data.

Now we have a batch of observations $x[n] \in R,\ n = 1, \cdots, N$, which are real numbers located at the interval $(0, 1)$. We plot the histogram of these observations:



If we model it with a mixture model (as the red/green lines), can you give an appropriate base distribution? You can choose one from Gaussian / Poisson / Uniform / Beta / Binomial.

(4) (Optional) **Number of mixture components.** All of the above problems consider a **finite** mixture model. So how to determine the K? This problem has no general standards, but there exists some solutions. We hope you give $2 \sim 3$ possible solutions.

2. **Latent variable models (LVMs).** The mixture model may be the simplest form of LVMs. We will discuss more in this problem set.

(1) **Factor analysis(FA).** FA assumes a latent variable $\boldsymbol{z}[n] \in R^L$, whose prior is a Gaussian:

$$p(\boldsymbol{z}[n]) = \mathcal{N}(\boldsymbol{z}[n]|\mathbf{0}, \boldsymbol{I})$$

where $\boldsymbol{I}$ is an identity matrix, and an observation $\boldsymbol{x}[n] \in R^D (D \gg L)$, whose conditional distribution is also a Gaussian with the mean defined by a **linear** function of $\boldsymbol{z}[n]$:

$$p(\boldsymbol{x}[n]|\boldsymbol{z}[n]) = \mathcal{N}(\boldsymbol{x}[n]|\boldsymbol{W}\boldsymbol{z}[n] + \boldsymbol{\mu}, \Psi)$$

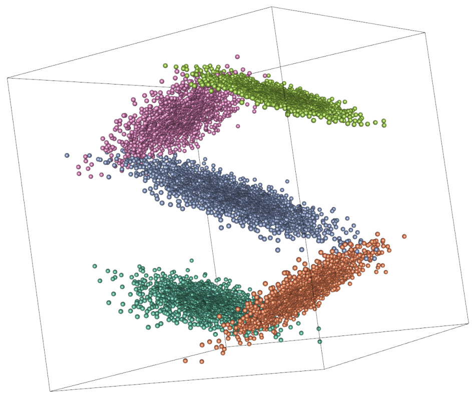
where Ψ is forced to be diagonal. (That is to say, x_i and x_j are independent given $\boldsymbol{z}$.)

(a) Please draw a graphical representation of this model.

(b) Please figure out the number of independent parameters of the FA model, and compare it with that of the general multivariate Gaussian distribution.

(c) (Optional) Please list some possible applications of FA.

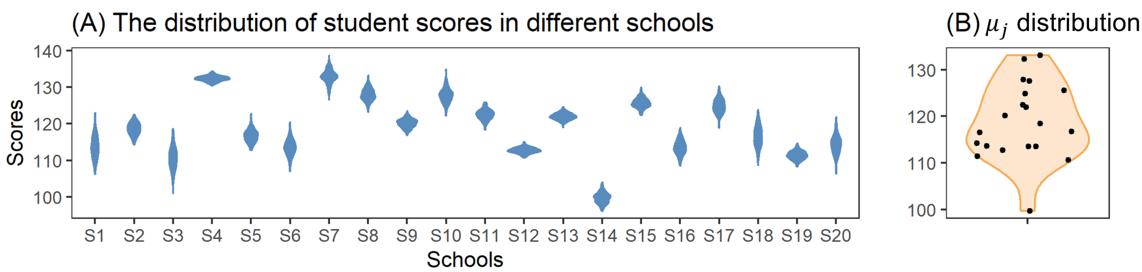
(2) Consider the following data points. Maybe you have known we can regard the factor analysis model as a dimenion reduction method, observing it uses a lower-dimension hyerplane to approximate original data points. But in tha above figure, it's hard to model these data points using FA, becuase no **single** planes can fit them. But we can model them as a finite **mixture** of planes. Please design and give a detailed description of an appropriate model and its local probability form.



3. **Hierarchical Bayesian model.** Hierarchical Bayesian model is a special and widely used type of Bayesian networks, which can model the information from multiple levels. Now we consider a simple case: applying a hierarchical model to the students' college entrance examination scores from different schools in a city.

Suppose we have a dataset including the students' scores from m schools. Let y_{ij} denote the score of student i in school j ($j = 1, ..., m$; $i = 1, ..., n_j$; n_j is the number of students in school j, $\sum_{j=1}^{m} n_j = n$). We assume that the scores y_{ij} are continuous and obey Gaussian distribution.

Following is the visualization of the scores' distribution.



Now we consider a **two-stage hierarchical model**.

Firstly, considering that <u>different schools have different educational resources</u>, we assume that each school has a specific mean μ_j and precision τ_j ($\tau_j = 1/\sigma_j^2$, σ_j is the standard deviation) (**first stage**)(Fig.A). So the distribution of the score of student i in school j can be denoted as

$$y_{ij}|\mu_j, \tau_j \sim \mathcal{N}(\mu_j, 1/\tau_j)$$

which means different school has different μ_j value.

Secondly, different schools also <u>share some similarities</u> due to the same examination questions, which means the μ_j are not independent. So we add the **second stage** to model this similarity (Fig.B). In detail, we set the prior distribution of the mean μ_j as

$$\mu_j|\mu, \tau_\mu \sim \mathcal{N}(\mu, 1/\tau_\mu)$$

which means all μ_j share a common mean μ and precision τ_μ.

For the variables τ_j, μ, and τ_μ, we set their prior distributions as

$$\tau_j \sim \Gamma(a, b),\ \ j = 1, ..., m$$
$$\mu \sim \mathcal{N}(\xi, \lambda)$$
$$\tau_\mu \sim \Gamma(c, d)$$

where a, b, c, d, ξ, λ are hyperparameters.

(1) Please draw the corresponding graphical model.

(2) (Optional) Suppose we also have some features of these students, such as their simulation test scores, which can be denoted as $\boldsymbol{x}_{ij}$. And we think the mean μ_j has a generalized linear relationship with the student feature vector $\boldsymbol{x}_{ij}$, which is $\mu_j = g(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j)$ (g is a link function, and $\boldsymbol{\beta}_j$ is the weight vector). So the model can be re-denoted as

$$y_{ij} | \boldsymbol{x}_{ij}, \boldsymbol{\beta}_j, \tau_j \sim \mathcal{N}(g(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j), 1/\tau_j), \quad j = 1, ..., m$$
$$\boldsymbol{\beta}_j | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\xi}, \boldsymbol{\Psi})$$
$$\tau_j \sim \Gamma(a, b)$$

where $a, b, \boldsymbol{\Sigma}, \boldsymbol{\xi}, \boldsymbol{\Psi}$ are hyperparameters.

(Optional) Please draw the corresponding graphical model.