SDS 383C - Statistical Modeling 1: Homework 4

Rahul Nandakumar
Graduate Student, ORIE Program (E-ID: rn9355)

November 9, 2022

1.

Solution: Given a normal likelihood model, the likelihood function is given by

$$\begin{aligned} f(y \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-1}{2\sigma^2}(y-\mu)^2\right) \\ p(y_{1:n} \mid \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-1}{2\sigma^2}(y_i-\mu)^2\right) \\ &\propto (\sigma^2)^{-n/2}\exp\left(\frac{-1}{2\sigma^2}\big((n-1)s^2 + n(\bar{y}-\mu)^2\big)\right) \end{aligned}$$

The pdf of an NIG Distribution is given by,

$$f(x, \sigma^2 \mid \mu, \lambda, \alpha, \beta) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}\sigma}\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left(-\frac{2\beta + \lambda(x-\mu)^2}{2\sigma^2}\right)$$

The a-priori NIG prior on μ, σ^2 is given by,

$$p(\mu, \sigma^2) = \text{NIG}(\mu_0, \sigma_0^2\ \kappa_0, v_0, \sigma_0^2)$$

Here, comparing to the original pdf of NIG Distribution, we can see that

$$\begin{aligned} \alpha &= \frac{v_0 + 1}{2} \\ \beta &= \frac{v_0\sigma_0^2}{2} \\ x &= \mu \\ \mu &= \mu_0 \\ \lambda &= \kappa_0 \end{aligned}$$

Now, according to the description of our variables, the prior can be written as follows.

$$\begin{aligned} p(\mu, \sigma^2) &\propto (\sigma^2)^{-(\frac{v_0}{2}+1+\frac{1}{2})}\exp\left(\frac{-1}{2\sigma^2}(v_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2)\right) \\ &\propto \text{Inv-Gamma}(\sigma^2 \mid v_0/2, v_0\sigma_0^2/2) \times \text{Normal}(\mu \mid \mu_0, \sigma^2/\kappa_0) \end{aligned}$$

Since we are able to write this pdf as as a product of 2 densities, where μ is dependent on σ^2, i.e., of the form

$$p(\mu, \sigma^2) = p(\sigma^2 \mid \ldots) \times p(\mu \mid \sigma^2, \ldots) \neq p(\mu) \times p(\sigma^2)$$

We can conclude that the a-priori μ and σ^2 are dependent variables. Now, to prove that tehy are uncorrelated, We know that the covariance of random variables X and Y can be calculated by the followign relationship,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

The law of iterated expectations for general NIG families gives us the relation,

$$\mathbb{E}(\mu.\sigma^2) = \mathbb{E}_{\sigma^2}\mathbb{E}_{\mu|\sigma^2}(\mu.\sigma^2 \mid \sigma^2) \tag{1}$$

Now, $\text{Cov}(\mu, \sigma^2)$ can be calculated as,

$$\begin{aligned}
\text{Cov}(\mu, \sigma^2) &= \mathbb{E}[\mu.\sigma^2] - \mathbb{E}(\mu)E(\sigma^2) \\
&= \mathbb{E}_{\sigma^2}\mathbb{E}_{\mu|\sigma^2}(\mu.\sigma^2 \mid \sigma^2) - \mathbb{E}(\mu)\mathbb{E}(\sigma^2);\ \text{Since from (1)} \\
&= \mathbb{E}_{\sigma^2}\mathbb{E}_{\mu|\sigma^2}(\mu.\sigma^2 \mid \sigma^2) - \mu_0.\frac{v_0\sigma_0^2}{v_0 - 1};\ \text{Since } \mathbb{E}(\sigma^2) = \frac{\beta}{\alpha - 1} \\
&= \int_0^\infty \left(\int_{-\infty}^\infty \mu\sigma^2(\sigma^2)^{-(\frac{v_0}{2}+1+\frac{1}{2})}\exp\left(\frac{-1}{2\sigma^2}(v_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2)\right)d\mu\right)d\sigma^2 \\
&\qquad - \mu_0.\frac{v_0\sigma_0^2}{v_0 - 1} \\
&= \mu_0.\frac{v_0\sigma_0^2}{v_0 - 1} - \mu_0.\frac{v_0\sigma_0^2}{v_0 - 1};\ \text{Evaluating first term using a definite integral calculator} \\
&= 0
\end{aligned}$$

Since the covariance equals zero, we can conclude that μ, σ^2 are uncorrelated. Let us now consider the posterior for μ, σ^2. The posterior is evaluated as,

$$\begin{aligned}
p(\mu, \sigma^2 \mid y_{1:n}) &\propto p(y_{1:n}) \times p(\mu, \sigma^2) \\
&= (\sigma^2)^{-(\frac{v_0}{2}+1+\frac{1}{2})}\exp\left(\frac{-1}{2\sigma^2}(v_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2)\right)\times \\
&\quad (\sigma^2)^{-n/2}\exp\left(\frac{-1}{2\sigma^2}\big((n-1)s^2 + n(\bar{y} - \mu)^2\big)\right) \\
&\propto (\sigma^2)^{-\left(\frac{v_0+n}{2}+1+\frac{1}{2}\right)}\exp\left(\frac{-1}{2\sigma^2}\big(v_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{n+\kappa_0}(\bar{y} - \mu_0)^2 + \right. \\
&\qquad \left. (\kappa_0 + n)(\mu - \mu_n)^2\big)\right)
\end{aligned}$$

This looks like the pdf of a Normal-Inverse Gamma Distribution with the following parameters

$$\begin{aligned}
\alpha &= v_n = v_0 + n \\
\beta &= \sigma_n^2 = \frac{1}{v_n}\left(v_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{n+\kappa_0}(\bar{y} - \mu_0)^2\right) \\
x &= \mu \\
\mu &= \mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n} \\
\lambda &= \kappa_n = \kappa_0 + n
\end{aligned}$$

Therefore, we can evaluate the posterior on μ, σ^2 as,

$$p(\mu, \sigma^2 \mid y_{1:n}) = \text{NIG}(\mu_n, \sigma_n^2/\kappa_n, v_n, \sigma^2/n)$$

Since this belongs to the same family of distribution as the prior, we can conclude that μ, σ^2 but uncorrelated.

2.

Solution: Given, a normal likelihood model, the likelihood function is given by

$$\begin{aligned} f(y \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-1}{2\sigma^2}(y-\mu)^2\right) \\ p(y_{1:n} \mid \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-1}{2\sigma^2}(y_i-\mu)^2\right) \\ &\propto (\sigma^2)^{-n/2}\exp\left(\frac{-1}{2\sigma^2}\big((n-1)s^2 + n(\bar{y}-\mu)^2\big)\right) \end{aligned}$$

The non-informative prior for μ, σ^2 is given as,

$$p(\mu, \sigma^2) \propto \sigma^{-2}$$

Thus, the posterior can be evaluated as,

$$\begin{aligned} p(\mu, \sigma^2 \mid y_{1:n}) &\propto p(\mu, \sigma^2) \times p(y_{1:n} \mid \mu, \sigma^2) \\ &= (\sigma^2)^{-n/2}\exp\left(\frac{-1}{2\sigma^2}\big((n-1)s^2 + n(\bar{y}-\mu)^2\big)\right) \times \sigma^{-2} \\ &= (\sigma^2)^{-(n/2+1)}\exp\left(\frac{-1}{2\sigma^2}\big((n-1)s^2 + n(\bar{y}-\mu)^2\big)\right) \end{aligned}$$

This is the pdf of an inverse gamma distribution with shape parameter $n/2$ and scale parameter $\big((n-1)s^2 + n(\bar{y}-\mu)^2\big)$. From this function, we can conclude that μ and σ^2 are dependent. Now, we know the relation

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Thus,

$$\mathrm{Cov}(\mu, \sigma^2) = \mathbb{E}[\mu.\sigma^2] - \mathbb{E}[\mu]\mathbb{E}[\sigma^2]$$

Using the law of iterated expectations,

$$\mathbb{E}(\mu.\sigma^2) = \mathbb{E}_{\sigma^2}\mathbb{E}_{\mu|\sigma^2}(\mu.\sigma^2 \mid \sigma^2)$$

we can arrive at the relation,

$$\begin{aligned} \mathrm{Cov}(\mu, \sigma^2) &= \mathbb{E}_{\sigma^2}\mathbb{E}_{\mu|\sigma^2}(\mu.\sigma^2 \mid \sigma^2) - \mathbb{E}[\mu]\mathbb{E}[\sigma^2] \\ &= \int_0^\infty \left(\int_{-\infty}^\infty \mu\sigma^2(\sigma^2)^{-(n/2+1)}\exp\left(\frac{-1}{2\sigma^2}\big((n-1)s^2 + n(\bar{y}-\mu)^2\big)\right)d\mu\right)d\sigma^2 \\ &\qquad - \mathbb{E}[\mu]\mathbb{E}[\sigma^2] \\ &= 0 \end{aligned}$$

Therefore, we can conclude that μ and σ^2 are dependent but uncorrelated. *Hence Proved. (a)*

To find out the marginal posteriors, we first consider the following relations.

$$\begin{aligned} p(\mu \mid y_{1:n}) &= \int p(\mu, \sigma^2 | y_{1:n}) d\sigma^2 \\ &= \int_0^\infty (\sigma^2)^{-(n/2+1)} \exp\left(\frac{-1}{2\sigma^2} \big((n-1)s^2 + n(\bar{y} - \mu)^2 \big) \right) d\sigma^2 \end{aligned}$$

Let

$$\begin{aligned} v_0 &= n \\ \sigma_0 &= (n-1)s^2/n + (\bar{y} - \mu)^2 \end{aligned}$$

Thus,

$$\begin{aligned} p(\mu \mid y_{1:n}) &\propto \int_0^\infty (\sigma^2)^{-(n/2+1)} \exp\left(\frac{-1}{2\sigma^2} \big((n-1)s^2 + n(\bar{y} - \mu)^2 \big) \right) d\sigma^2 \\ &\propto (\sigma_0^2)^{-n/2} \int_0^\infty \frac{(n/2)^{-n/2}}{\Gamma(n/2)} (\sigma_0^2)^{n/2} (\sigma^2)^{-n/2+1} \exp\left(- \frac{n\sigma_0^2}{2\sigma^2} \right) d\sigma^2 \\ &\propto ((n-1)s^2/n + (\bar{y} - \mu)^2)^{-n/2} \times 1; \end{aligned}$$

Since the integrand is the pdf of a Inverse - Gamma distribution with parameters v_0 and σ_0, the value of the integral is 1

$$\propto \left(1 + \frac{1}{n+1} \left(\frac{\mu - \bar{y}}{s/\sqrt{n}} \right)^2 \right)^{-\frac{(n-1)+1}{2}}$$

We recognize this as the pdf of a scaled and shifted t distribution with $n-1$ degrees of freedom. Therefore,

$$p(\mu \mid y_{1:n}) \sim t_{n-1}(\bar{y}, s^2/n)$$

Considering the second relation,

$$\begin{aligned} p(\sigma^2 \mid y_{1:n}) &= \int_{-\infty}^\infty p(\mu, \sigma^2 | y_{1:n}) d\mu \\ &= \int_{-\infty}^\infty (\sigma^2)^{-(n/2+1)} \exp\left(\frac{-1}{2\sigma^2} \big((n-1)s^2 + n(\bar{y} - \mu)^2 \big) \right) d\mu \\ &= (\sigma^2)^{-n/2+1} \exp\left(\frac{(n-1)s^2}{2\sigma^2} \right) \int_0^\infty \exp\left(\frac{-n}{2\sigma^2} (\bar{y} - \mu)^2 \right) d\mu; \end{aligned}$$

Evaluating this integral using a definite integral calculator, we get

$$(\sigma^2)^{-n/2+1} \exp\left(- \frac{(n-1)s^2}{2\sigma^2} \right) \sqrt{\frac{2\pi\sigma^2}{n}}$$

$$(\sigma^2)^{-(\frac{n-1}{2}+1)} \exp\left(- \frac{(n-1)s^2}{2\sigma^2} \right)$$

We recognize this as the pdf of an inverse gamma distribution with parameters $\frac{n-1}{2}$, $\frac{(n-1)s^2}{2}$. Therefore,

$$p(\sigma^2 \mid y_{1:n}) \sim \text{Inv-Ga}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2} \right)$$

Now, for the predictive distribution, we arrive at the following expression.

$$\begin{aligned}
p(y_{new} \mid y_{1:n}) &= \int\int p(y_{new} \mid \mu, \sigma^2, y_{1:n}) p(\mu, \sigma^2 \mid y_{1:n}) d\mu d\sigma^2 \\
&\propto \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma} \exp\left(\frac{-1}{2\sigma^2}(y_{new} - \mu)^2\right) \times \\
&\quad (\sigma^2)^{-(n/2+1)} \exp\left(\frac{-1}{2\sigma^2}\left((n-1)s^2 + n(\bar{y} - \mu)^2\right)\right) d\mu d\sigma^2 \\
&\propto \frac{1}{\sqrt{n-1}\sigma} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(n/2)} \left(1 + \frac{1}{n-1}\left(\frac{y_{new} - \bar{y}}{\sigma}\right)^{-n/2}\right) \\
&\propto \frac{1}{\sqrt{n-1}\sigma} \frac{1}{\text{Beta}(n/2, 1/2)} \left(1 + \frac{1}{n-1}\left(\frac{y_{new} - \bar{y}}{\sigma}\right)^{-n/2}\right)
\end{aligned}$$

Where $\sigma^2 = \left(1 + \frac{1}{n}\right)^{1/2} s$. This integral evaluates to the pdf of a scaled and shifted t distribution with $n-1$ degrees of freedom, location parameter $\bar{y}$, scale $\left(1 + \frac{1}{n}\right)^{1/2} s$ i.e.,

$$p(y_{new} \mid y_{1:n}) \sim t_{n-1}(\bar{y}, (1 + 1/n)^{1/2} s)$$

3.

Solution: Let us consider a multinomial likelihood model with K categories. The pmf of a multinomial distribution is given by,

$$p(y \mid \boldsymbol{\pi}, n) = n! \prod_{i=1}^{K} \frac{\pi_i^{x_1}}{x_i!}$$

Let us consider the log likelihood for this pmf. The log likelihood is given by the following expression. From this, we calculate the Fisher Information matrix to arrive at the expression for the Jeffreys' prior.

$$\begin{aligned}
\log(\mathcal{L}(\pi)) &= \log\left(n! \prod_{i=1}^{K} \frac{\pi_i^{x_1}}{x_i!}\right) \\
&= \log(n!) + \sum_{i=1}^{k} x_i \log(\pi_i) - \sum_{i=1}^{K} \log(x_i!) \\
\frac{\partial \mathcal{L}(\pi)}{\partial \pi_i} &= \frac{\partial}{\partial \pi_i}\left(\log(n!) + \sum_{i=1}^{k} x_i \log(\pi_i) - \sum_{i=1}^{K} \log(x_i!)\right) \\
&= \left(0 + \frac{x_i}{\pi_i} + 0\right) \\
\frac{\partial^2 \mathcal{L}(\pi)}{\partial \pi_i^2} &= \frac{-x_i}{\pi_i^2}
\end{aligned}$$

Essentially, the diagonal elements of the Fisher information matrix is given by the Ex-

pected value of this partial differential. i.e.,

$$\begin{aligned} I(\pi) &= \mathbb{E}\bigg(- \frac{\partial^2 \mathcal{L}(\pi)}{\partial \pi_i^2} \bigg) \\ &= \frac{\mathbb{E}(x_i)}{\pi_i^2} \\ &= \frac{n}{\pi_i}; \text{ Since in a multinomial distribution, } \mathbb{E}(x_i) = n\pi_i \end{aligned}$$

The Jeffreys' prior is

$$\begin{aligned} p(\pi_i) &\propto |I(\pi)|^{-1/2} \\ &\propto \frac{n}{\pi_i} \end{aligned}$$

This case happens when the prior is distributed as a Dirichlet distribution with $\alpha_i = 1/2$. Therefore, we can say that for a multinomial likelihood model with K categories, the Jeffrey's prior is distributed as,

$$p(\pi_i) \sim \text{Dir}(1/2, \ldots, 1/2)$$

4.

Solution: *a.* The Galaxies dataset has been downloaded from the following website `https://r-data.pmagunia.com/dataset/r-dataset-package-mass-galaxies`. This dataset consists of velocities in km/sec of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. We desire to fit a normal location mixture model of the form

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{Normal}(y_i \mid \mu_k, \sigma^2)$$

The EM Algorithm has been implemented in Python, and the steps to update the parameters of the normal location mixture distribution has been carried out according to the results taken from the lecture slides. (Pg. 39/96 - SDS-383C-F2022-M6-L3). To facilitate convergence, we keep the means fixed at initial values for the first 20 or so iterations and only update π_k's and σ. After 20 or so iterations, we update all parameters. The code is explained in detail in the comments in the code file. We arrive at the results for the fitted model as shown below.

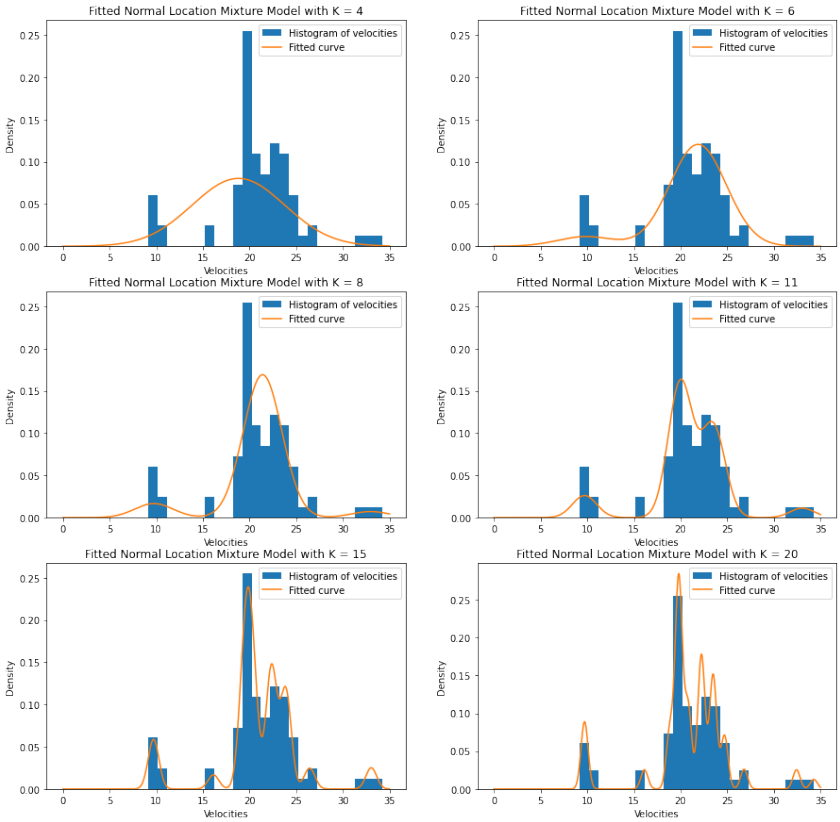

Figure 1: Normal Location Mixtures fitted for the Galaxies Dataset

b. To calculate the AIC and BIC values, we make use of the following relations. Here, $\mathcal{L}(\theta)$ is the maximum value of the likelihood function for the model and p is the number of parameters, n is the number of data points.

$$\text{AIC (Akaike Information Criteria)} = 2\{-\mathcal{L}(\theta) + p\}$$

$$\text{BIC (Bayesian Information Criteria)} = \{-2\mathcal{L}(\theta) + p \log n\}$$

Using these relations, we can calculate the values of AIC and BIC for the models. Among a finite set of models; models with lower AIC and BIC are generally preferred.

Model Components	**AIC value**	**BIC value**
4	352.7817689704255	362.4086459594825
6	335.5590562953825	349.999371778968
8	380.22546857572144	399.4792225538355
11	323.0554458958644	349.5293576157712
15	357.3091456396654	393.40993434862924
20	314.15244074768043	362.2868256929655

From this table, we can choose the model with 11 components as the best fit model, as both the AIC and BIC values are comparatively low.

c. Now, we fit normal location scale mixture models with form as follows.

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

The EM Algorithm has been implemented in Python, and the steps to update the parameters of the normal location mixture distribution has been carried out according to the results taken from the lecture slides. (Pg. 36/96 - SDS-383C-F2022-M6-L3). To facilitate convergence, we keep the means fixed at initial values for the first 20 or so iterations and only update π_k's and σ_k's. After 20 or so iterations, we update all parameters. We arrive at the results for the fitted model as shown below.

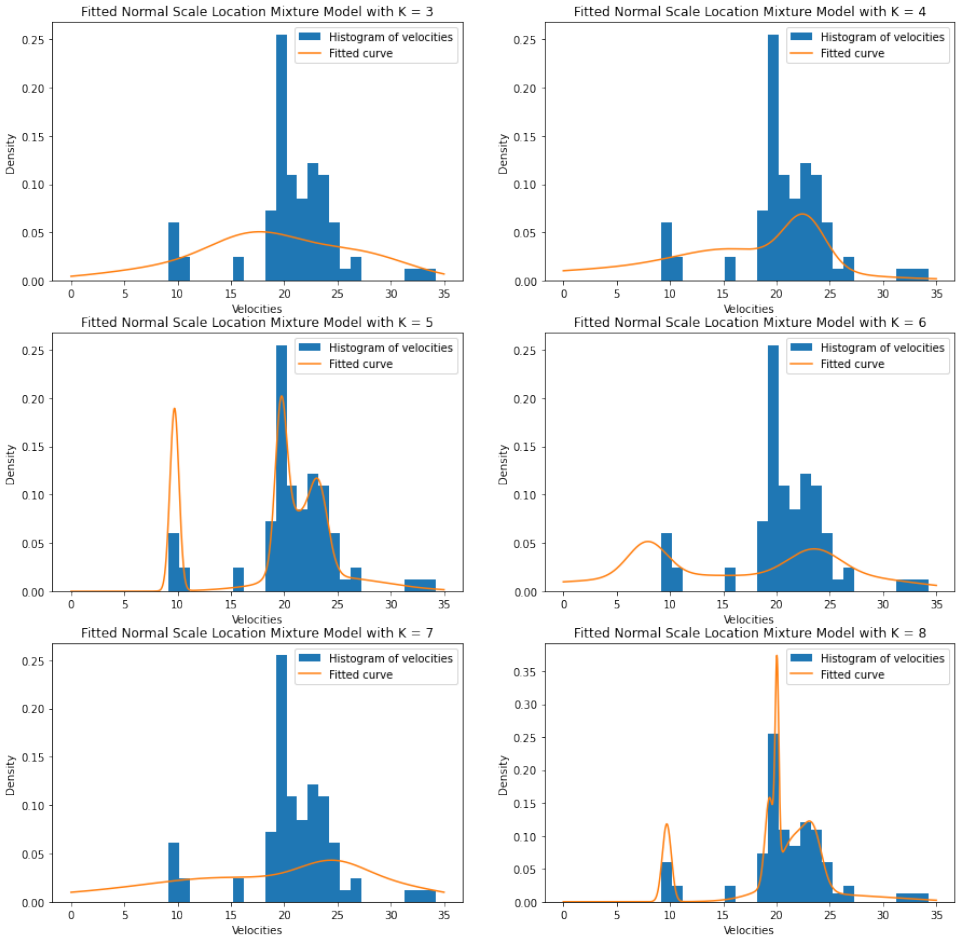

Figure 2: Normal Location-Scale Mixtures fitted for the Galaxies Dataset

*d.*Using the relations same as problem (b), we can calculate the values of AIC and BIC for the models. Among a finite set of models; models with lower AIC and BIC are generally preferred.

Model Components	**AIC value**	**BIC value**
3	536.0535205389705	543.2736782807632
4	550.7503355546576	560.3772125437146
5	379.0150026525681	391.0485988888894
6	646.6301254358095	661.0704409193951
7	706.5958783561925	723.4429130870424
8	433.5025155703568	452.75626954847087

From this table, we can choose the model with 5 components as the best fit model, as both the AIC and BIC values are comparatively low. *e.* In summary, we have fitted both

the normal location and location scale mixtures using the EM algorithm. The EM is an iterative algorithm, that has 2 main steps.

$$\text{E - Step: Compute } Q(\theta, \theta^{(m)}) = \mathbb{E}_{z \sim p(z|y,\theta^m)} \log\left(p(y, z|\theta)\right)$$
$$\text{M - Step: Compute } \theta^{(m+1)} = \arg\max_{\theta} Q(\theta, \theta^{(m)})$$

The convergence criteria has been specified to occur when the difference between the likelihood values between successive iterations becomes less than 0.0001. We see that in case of the normal location mixture model fitting, at larger values of K (=15,20) and in case of the normal location scale mixture model fitting, at larger values o K (=8), there are big spikes that accompany the fitted curve at certain points. This is due to overfitting, and is not ideal in choosing our best model.

5.

Solution: Now, we use the stochastic EM algorithm to fit location and location scale mixtures of normal to the galaxies dataset. The algorithm for updating the parameters using the stochastic EM algorithm is taken from the lecture slides (Pg. 96/96 - SDS-383C-F2022-M6-L3) We obtain the following graphs for the Location and Location Scale Mixture respectively. The code is explained in detail in the comments in the code file. To facilitate convergence, we keep the means fixed at initial values for the first 20 or so iterations and only update π_k's and σ and σ_k's. After 20 or so iterations, we update all parameters.

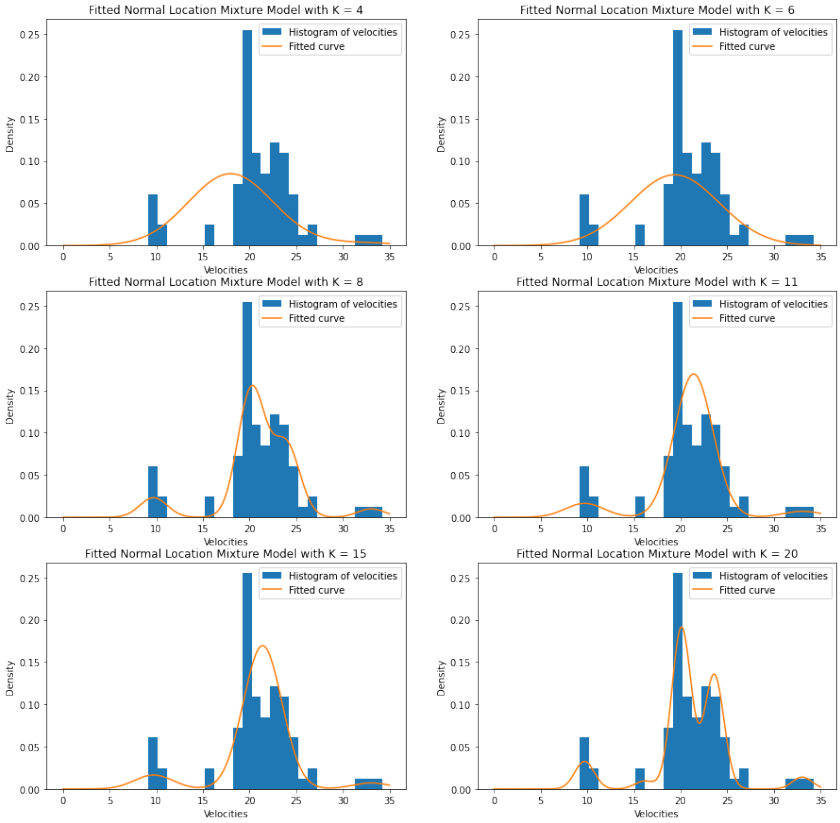

Figure 3: Normal Location Mixtures fitted for the Galaxies Dataset using the Stochastic EM Algorithm

The AIC and BIC values for the Normal Location mixture models fitted are tabulated as in correspondance to the relationships established in question 4b. The values are calculated as follows.

Model Components	AIC value	BIC value
4	370.06879972830274	383.63869188214187
6	350.0827245349783	364.52304001856385
8	321.1848947063752	335.1423880122533
11	296.1514964613857	322.62540818129247
15	304.1514964613857	340.2522851703495
20	322.3315294430886	361.62052957897447

From this table, we can see that the best AIC and BIC values is given my the model with 11 components, as lower AIC and BIC values suggest better model perfomance.

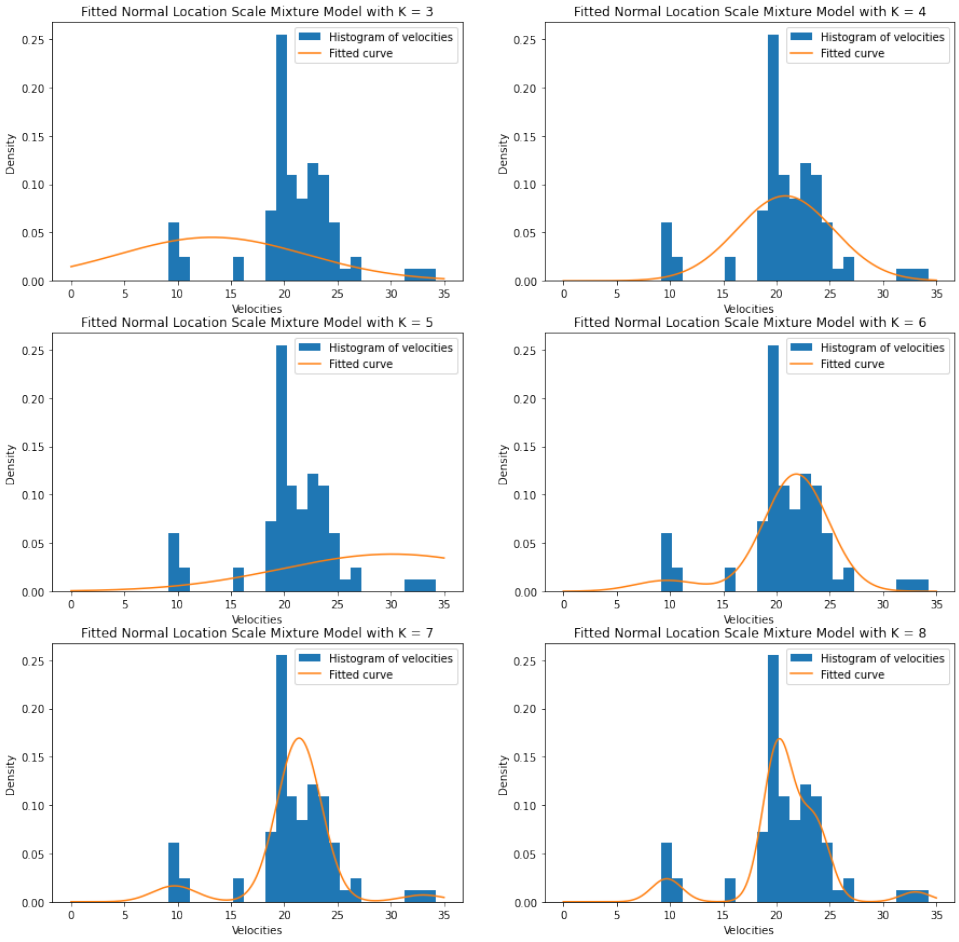

Figure 4: Normal Location - Scale Mixtures fitted for the Galaxies Dataset using the Stochastic EM Algorithm

Model Components	AIC value	BIC value
3	445.592073517817	452.8122312596098
4	337.969862945876	347.596739934933
5	475.3588669056884	487.39246314200966
6	322.34396464344843	336.784280127034
7	288.1514964613857	312.5526698778936
8	312.3150580081916	346.4431622586945

From this table, we can see that the best AIC and BIC values is given my the model with 7 components, as lower AIC and BIC values suggest better model perfomance. We also observe that in the stochastic setting, the model is prone to quite a lot of error in the starting conditions. Also, the model fits the data well on a whole, which is expected from any stochastic algorithm. We also observe that the runtime and number of iterations taken to achieve convergence of the stochastic algorithm in both cases have been reduced significantly.

6.

a. Question 6 asks us to model a multivariate distribution of data. The dataset to be modelled is the "Old Faithful Geyser" dataset. This dataset was downloaded from the website `https://r-data.pmagunia.com/dataset/r-dataset-package-datasets-faithful`. The dataset consists of the waiting times between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We desire to fit a Multivariate Normal Mixture Distribution with K componenets of the form

$$\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

The EM Algorithm is used to fit these models. We start at an initial guess for the model parameters and iteratively update the values until convergence occurs. The convergence criteria is attained, when the log likelihood values between successive iterations has a tolerance of 0.0001. The equations to update the parameters of this model are implemented in Python, and the code has been explained in the comments. To speed up our calculations, we also have scaled down the values of each observation using the MinMaxScaler available in Python. We observe the following results. First, we plot the contour plots of the fitted densities superimposed over a scatterplot of the data points.

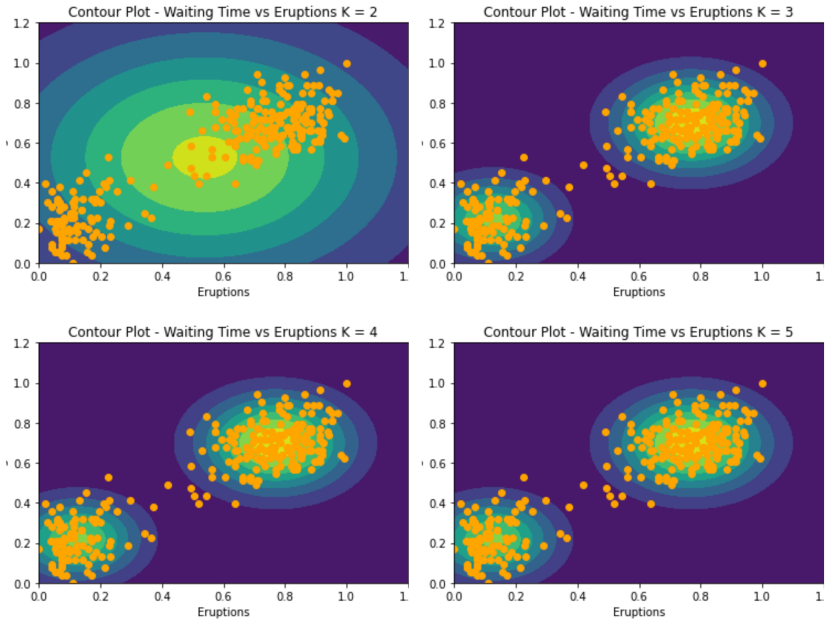

Figure 5: Contours of the Multivariate Normal Mixtures fitted for the Old Faithful Geyser Dataset using the EM Algorithm

Further, we also plot the densities of these fitted mixture models in 3 dimensions to best visualize how the multivariate normal mixture functions.

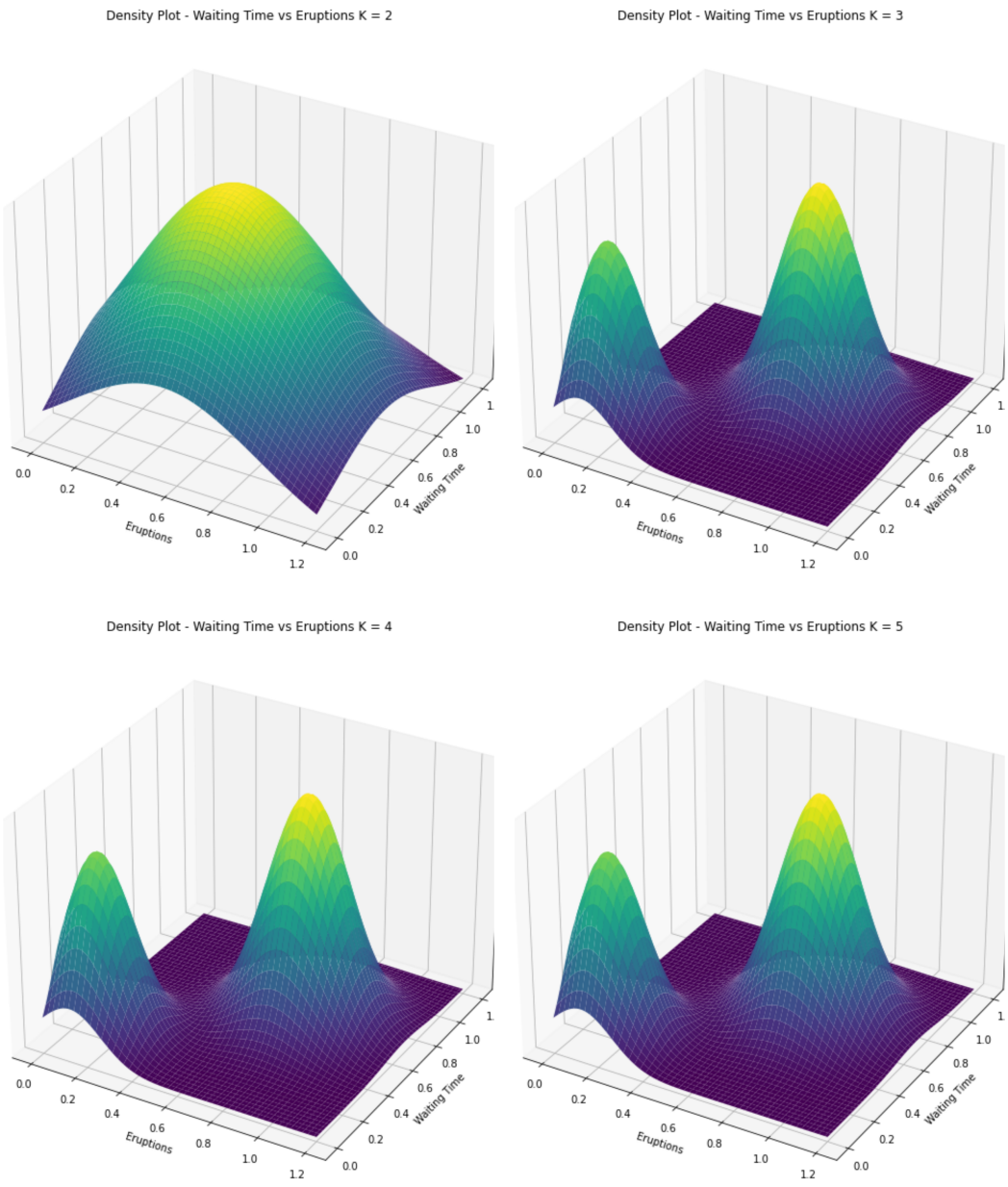

Figure 6: Densities of the Multivariate Normal Mixtures fitted for the Old Faithful Geyser Dataset using the EM Algorithm

b. To identify the best model among the following mixtures, we make use of the AIC and BIC values. These values are calculated similar to question 4. That is,

$$\text{AIC (Akaike Information Criteria)} = 2\{-\mathcal{L}(\theta) + p\}$$

$$\text{BIC (Bayesian Information Criteria)} = \{-2\mathcal{L}(\theta) + p \log n\}$$

where p is the number of parameters, n is the number of data points, and $\mathcal{L}(\theta)$ is the

log-likelihood value. Using these relations, we arrive at the following values for the AIC and BIC parameters.

Model Components	**AIC value**	**BIC value**
2	-180.90964109228753	146.7566404232498
3	136.87965618032885	-173.69807017059864
4	-180.9112823836408	145.39070277577534
5	-180.932910192116	-173.70826874375487

The lower the value for AIC and BIC, the better the fit of the model. The absolute value of the AIC or BIC value is not important. It can be positive or negative. From this table, we can conclude that therefore, the model with 5 components offers a better fit for our distribution.