Homework 4

Juliette Franqueville

November 9, 2022

(1) For a normal likelihood model with conjugate Normal-Inverse-Gamma prior on (μ, σ^2), show that a-priori and a-posteriori μ and σ^2 are dependent but uncorrelated.

They are evidently dependent because σ^2 is needed to calculate the density of μ. They are uncorrelated as shown below:

$$\begin{aligned}
cov(\mu, \sigma^2) &= E(\mu \cdot \sigma^2) - E(\mu)E(\sigma^2) \\
&= E_{\sigma^2} E_{\mu|\sigma^2}(\mu \cdot \sigma^2 | \sigma^2) - E(\mu)E(\sigma^2) \\
&= \mu_0 E_{\sigma^2} \sigma^2 - \mu_0 \nu_0 \sigma_0^2/2/(\nu_0/2 - 1) \\
&= \mu_0 \nu_0 \sigma_0^2/2/(\nu_0/2 - 1) - \mu_0 \nu_0 \sigma_0^2/2/(\nu_0/2 - 1) = 0
\end{aligned}$$

The posterior is also of NIG form, so the a-priori μ and σ^2 are also dependent but uncorrelated.

(2) Consider again a normal likelihood model but with NIP $p(\mu, \sigma^2) \propto \sigma^{-2}$.

(a) Show that a-posteriori μ and σ^2 are dependent but uncorrelated.

This is essentially the same question as earlier. They are dependent because σ^2 appears in the pdf of μ. For the posterior of σ^2, let $\alpha = n/2$ and $\beta = 1/2[(n-1)s^2 + n(\mu - \bar{y})^2]$ in $IG(\alpha, \beta)$

$$\begin{aligned}
cov(\mu, \sigma^2 | y) &= E(\mu \cdot \sigma^2 | y) - E(\mu|y)E(\sigma^2|y) \\
&= E_{\sigma^2} E_{\mu|\sigma^2}(\mu \cdot \sigma^2 | \sigma^2, y) - E(\mu|y)E(\sigma^2|y) \\
&= \bar{y} E_{\sigma^2}(\sigma^2|y) - \bar{y}\beta/(\alpha - 1) \\
&= \bar{y}\beta/(\alpha - 1) - \bar{y}\beta/(\alpha - 1) = 0
\end{aligned}$$

(b) Find out the marginal posteriors $p(\mu|y_{1:n})$ and $p(\sigma^2|y_{1:n})$.

We use the joint density and integrate w.r.t μ and σ^2.

$$
\begin{aligned}
P(\sigma^2|y) &\propto \int P(\sigma^2, \mu|y) d\mu \\
&\propto \int P(\sigma^2, \mu|y) d\mu \\
&\propto \int \sigma^{-(n/2+1)} e^{\frac{-1}{2\sigma^2}[(n-1)s^2 + n(\mu-\bar{y})^2]} d\mu \\
&\propto \sigma^{-(n/2+1)} e^{\frac{-1}{2\sigma^2}(n-1)s^2} \int e^{\frac{-1}{2\sigma^2}[n(\mu-\bar{y})^2]} d\mu \\
&\propto \sigma^{-(n/2+1)} e^{\frac{-1}{2\sigma^2}[(n-1)s^2} (\sigma^2)^{1/2} \\
&\propto \sigma^{-\left(\frac{n-1}{2}\right)-1} e^{\frac{-1}{2\sigma^2}(n-1)s^2} \\
&= IG\left(\frac{n-1}{2}, \frac{1}{2}(n-1)s^2\right)
\end{aligned}
$$

$$
\begin{aligned}
P(\mu|y) &\propto \int P(\sigma^2, \mu|y) d\sigma^2 \\
&\propto \int P(\sigma^2, \mu|y) d\sigma^2 \\
&\propto \int \sigma^{-(n/2+1)} e^{\frac{-1}{2\sigma^2}[(n-1)s^2 + n(\mu-\bar{y})^2]} \\
&\propto [(n-1)s^2 + n(\mu-\bar{y})^2]^{-n/2} \\
&\propto [1 + n(\mu-\bar{y})^2/(n-1)s^2]^{-n/2} \\
&= t_{n-1}(\bar{y}, s^2/n)
\end{aligned}
$$

(b) (c) Find out the predictive distribution $p(y_{new}|y_{1:n})$.

$$
\begin{aligned}
p(y_{new}|y_{1:n}) &= \int\int p(y_{new}|y_{1:n}, \sigma^2, \mu) P(\mu, \sigma^2|y) d\mu d\sigma^2 \\
&= \int\int p(y_{new}|y_{1:n}, \sigma^2, \mu) P(\mu|\sigma^2, y) P(\sigma^2|y) d\mu d\sigma^2
\end{aligned}
$$

Take the inner integral first

$$\int p(y_{new}|y_{1:n},\sigma^2,\mu)P(\mu|\sigma^2,y)d\mu$$
$$\int \propto (\sigma^2)^{-1/2}e^{-1/2\sigma^2(y_{new}-\mu)^2}(\sigma^2)^{-1/2}e^{-n/2\sigma^2(\mu-\bar{y})^2}d\mu$$
$$\int \propto (\sigma^2)^{-1}e^{-1/2\sigma^2[(y_{new}-\mu)^2+n(\mu-\bar{y})^2]}d\mu$$
$$\int \propto (\sigma^2)^{-1}e^{-1/2\sigma^2[y_{new}^2-2y_{new}\mu+\mu^2+n\mu^2-2n\mu\bar{y}+n\bar{y}^2]}d\mu$$
$$\int \propto (\sigma^2)^{-1}e^{-1/2\sigma^2[\mu^2(n+1)-2\mu(y_{new}+n\bar{y})+y_{new}^2]}d\mu$$
$$\int \propto (\sigma^2)^{-1}e^{-(n+1)/2\sigma^2[\mu^2-2\mu(y_{new}+n\bar{y})/(n+1)+y_{new}^2/(n+1)]}d\mu$$
$$\int \propto (\sigma^2)^{-1}e^{-(n+1)/2\sigma^2[\{\mu-(y_{new+n\bar{y}}/(n+1)\}^2-\{(y_{new+n\bar{y}}/(n+1)\}^2+y_{new}^2/(n+1)]}d\mu$$
$$\propto (\sigma^2)^{-1}(\sigma^2)^{1/2}e^{-(n+1)/2\sigma^2[-\{(y_{new+n\bar{y}}/(n+1)\}^2+y_{new}^2/(n+1)]}$$
$$\propto (\sigma^2)^{-1/2}e^{-(n+1)(n+1)^{-2}/2\sigma^2[-y_{new}^2-2y_{new}n\bar{y}-n^2\bar{y}^2+y_{new}^2(n+1)]}$$
$$\propto (\sigma^2)^{-1/2}e^{-n(n+1)^{-1}/2\sigma^2[y_{new}^2-2y_{new}\bar{y}]}$$
$$\propto (\sigma^2)^{-1/2}e^{-n(n+1)^{-1}/2\sigma^2[y_{new}^2-\bar{y}]^2}$$
$$\propto \mathcal{N}(\bar{y},(n+1)\sigma^2/n)$$

Now the full integral

$$p(y_{new}|y_{1:n}) \propto \int P(\sigma^2|y)[(\sigma^2)^{-1/2}e^{-n(n+1)^{-1}/2\sigma^2[y_{new}^2-\bar{y}]^2}d\sigma^2$$
$$\int \propto \sigma^{-\left(\frac{n-1}{2}\right)-1}e^{\frac{-1}{2\sigma^2}(n-1)s^2}e^{-n(n+1)^{-1}/2\sigma^2[y_{new}^2-\bar{y}]^2}d\sigma^2$$
$$\int \propto \sigma^{-n/2-1}e^{\frac{-1}{2\sigma^2}[(n-1)s^2+n/(n+1)(y_{new}^2-\bar{y})^2]}d\sigma^2$$
$$\propto [(n-1)s^2+n/(n+1)(y_{new}^2-\bar{y})^2]^{-n/2}\int IG(n/2,1/2[(n-1)s^2+n/(n+1)(y_{new}^2-\bar{y})^2])d\sigma^2$$
$$\propto [(n-1)s^2+n/(n+1)(y_{new}^2-\bar{y})^2]^{-n/2}$$
$$\propto [1+n/(n+1)(y_{new}^2-\bar{y})^2/(n-1)s^2]^{-n/2}$$
$$= t_{n-1}(\bar{y},(n+1)s^2/n)$$

f

(3) For a multinomial likelihood model with K categories, show that the Jeffreys' prior for the category probabilities is $Dir(1/2,\ldots,1/2)$.

For the multinomial distribution:

$$\mathcal{L}(y_1,\ldots y_k|n,p_1\ldots p_k) = \log\frac{n}{x_1!\ldots x_k!}\prod p_i^{x_i}$$
$$= \log\frac{n}{x_1!\ldots x_k!}+\sum \log p_i^{x_i}$$

$$\frac{\partial}{\partial p_i}\mathcal{L}(y_1,\dots y_k|n,p_1\dots p_k) = \frac{x_i}{p_i}$$

$$\frac{\partial^2}{\partial p_i \partial p_j}\mathcal{L}(y_1,\dots y_k|n,p_1\dots p_k) = \begin{cases} -\frac{x_i}{p_i^2}, & i=j \\ 0, & \text{otherwise} \end{cases}$$

In order words, this is a k by k diagonal matrix with diagonal elements $-x_i/p_i^2$. We also have $-E(-x_i/p_i^2) = E(x_i)/p_i^2 = np_i/p_i^2 = n/p_i$, so the information matrix is

$$I(p_1,\dots,p_k) = \begin{bmatrix} n/p_1 & & \\ & \ddots & \\ & & n/p_k \end{bmatrix}$$

and Jeffreys prior is:

$$\begin{aligned} p_1,\dots,p_n &\propto \det \begin{bmatrix} n/p_1 & & \\ & \ddots & \\ & & n/p_k \end{bmatrix}^{1/2} \\ &\propto \prod p_k^{-1/2} \\ &\propto \prod p_k^{1/2-1} \\ &\propto Dir(1/2,\dots 1/2) \end{aligned}$$

(4) (a) Using the EM algorithm, fit location mixtures of normals to the Galaxy data.

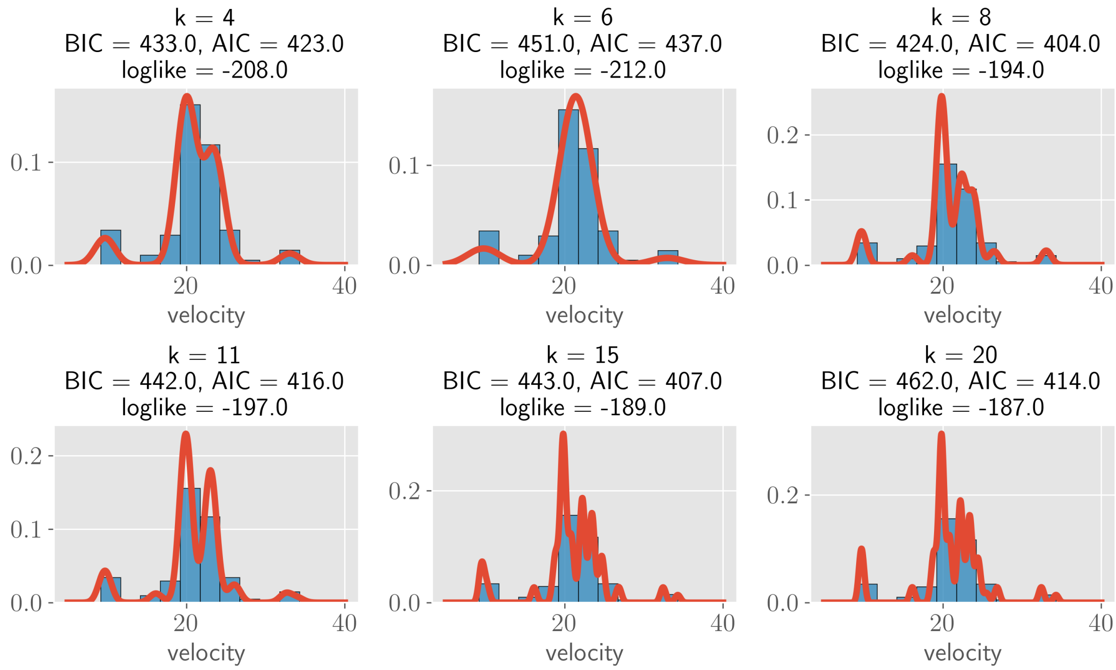

Figure 1: Results when the variances and pooled.

(b) Tabulate AIC and BIC values for each case and report the 'best' model(s)

The AIC and BIC values are shown on the plots. $k = 8$ has both the lowest BIC and AIC.

(c) Next, fit location-scale mixtures of normals [where the variances are not pooled]

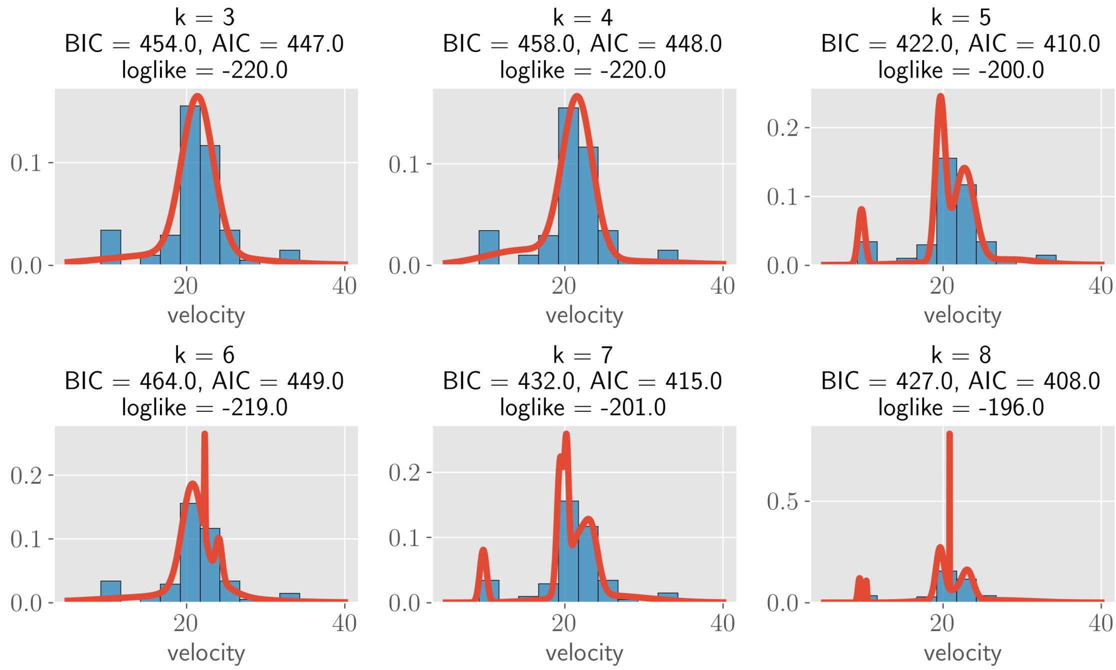

Figure 2: Results when the variances and not pooled (k variances).

(d) Tabulate AIC and BIC values for each case and report the 'best' model(s)

The AIC and BIC values are shown on the plots. $k = 5$ had the lowest BIC but $k = 8$ has the lowest AIC.

(e) Summarize your general findings

As expected, increasing the number of k generally increases the likelihood of the data, but can cause overfitting. The AIC and BIC measures take this into account by penalizing model complexity. Having a different variance for each normal allows for more flexibility in the model. The results seemed very dependent on starting conditions.

(5) Repeat everything you did in Problem No 3 above but this time using the stochastic EM algorithm

(a) Using the EM algorithm, fit location mixtures of normals to the Galaxy data

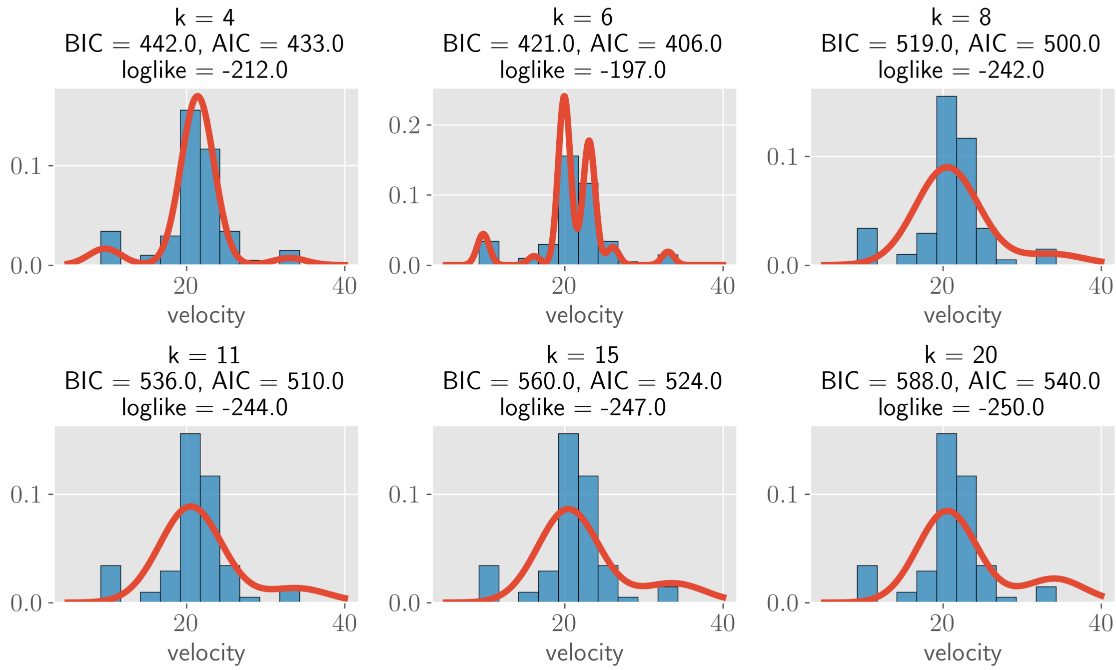

Figure 3: Results when the variances and pooled, stochastic.

(b) Tabulate AIC and BIC values for each case and report the 'best' model(s)

The AIC and BIC values are shown on the plots. $k = 6$ has the lowest BIC and AIC.

(c) Next, fit location-scale mixtures of normals [where the variances are not pooled]

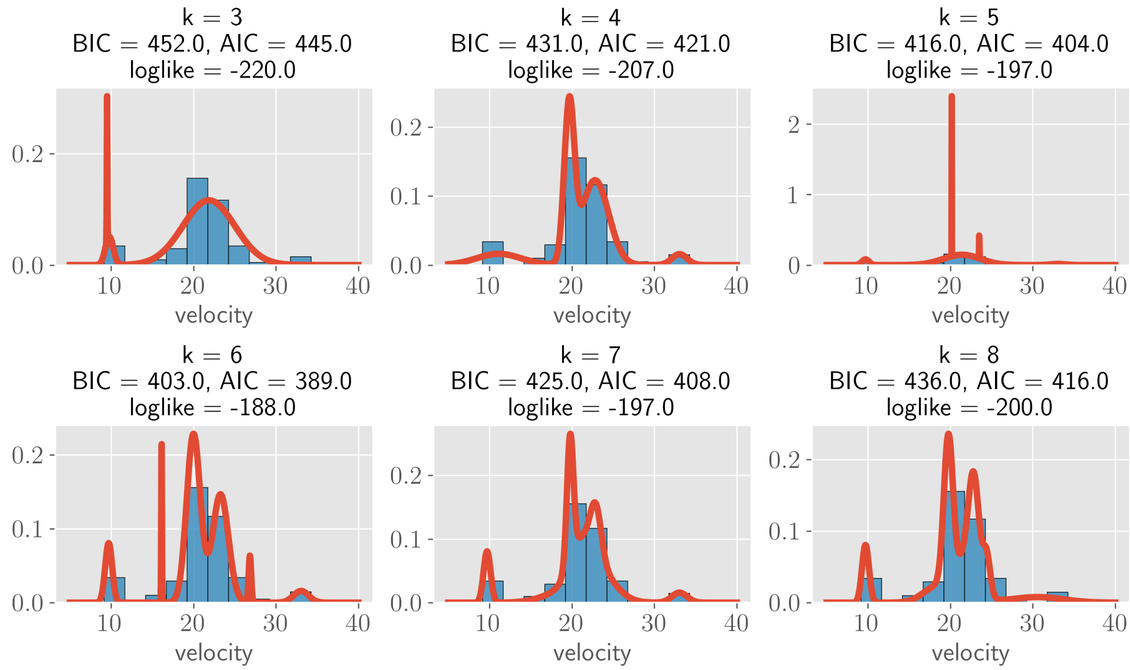

Figure 4: Results when the variances and not pooled (k variances), stochastic.

(d) Tabulate AIC and BIC values for each case and report the 'best' model(s).

The AIC and BIC values are shown on the plots. $k = 6$ had the lowest BIC and $k = 6$ had the lowest AIC.

(e) Summarize your general findings

The conclusions are similar as before; increasing the number of k generally increases the likelihood of the data, but can cause overfitting. I found it more difficult to get the stochastic EM algorithms to converge. In fact, I am not sure that the algorithm converged for all k values in the case where the variances were all the same. As before, the results seemed very dependent on starting conditions. With the stochastic method, I was not sure how to deal with the fact that there was no guarantee that all k would be sampled in $z_i \sim Mult(1, \pi_i)$. I made sure it was the case by sampling again if a k had not been sampled, but this made the algorithm slower.

(6) The 'faithful' dataset from package 'datasets' in R gives eruption and waiting times of the old faithful geyser in Yellowstone national park (export this dataset from R if you are using a different programming language)

(a) Using the EM algorithm, fit location-scale mixtures of bivariate normals

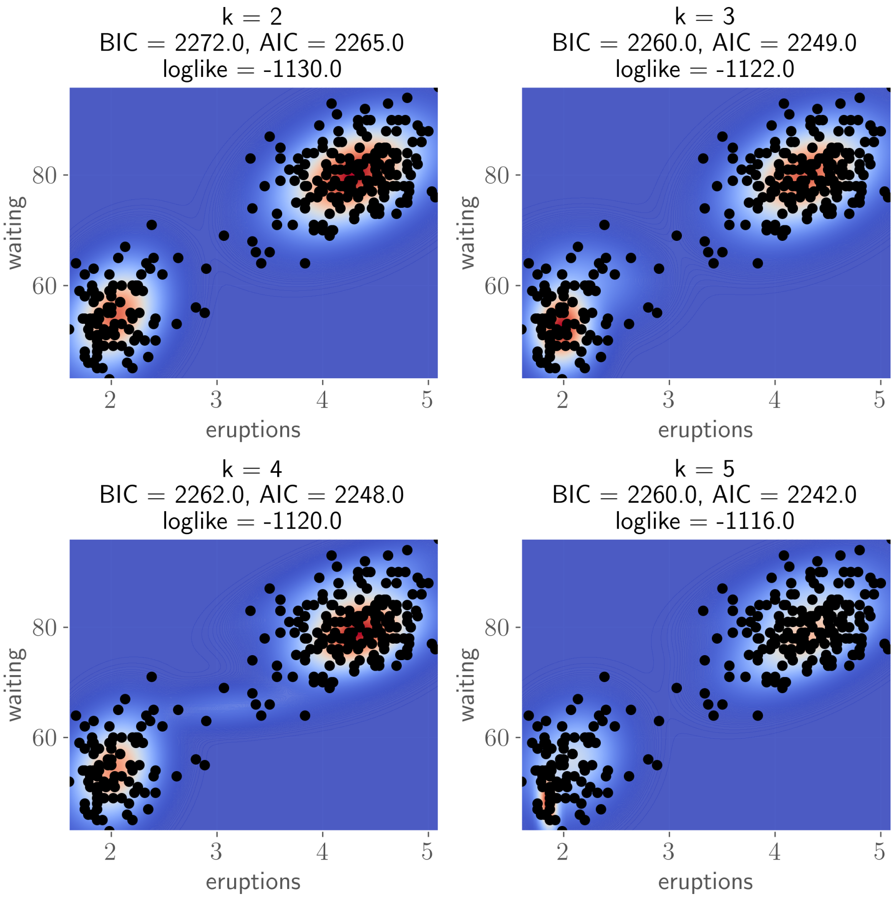

Figure 5: Results for k bivariates. Note that some of the bivariates have means close to each other, so it is difficult to distinguish them from the plots.

(b) Tabulate AIC and BIC values for each case and report the 'best' model(s).

The values are shown in the plot. $k = 5$ had the lowest BIC and AIC values, though they were very close to those for $k = 3, 4$. As before, I had trouble getting the algorithm to converge and it took many attempts with different starting values to obtain reasonable results.