1 Exercise 1

For a patient newly diagnosed with particular form of cancer, the survival time to death(in years), Y, has a Weibull(θ,r) distribution. The probability density function of $Y \sim Weibull(\theta, r)$ is

$$p(y|\theta, r) = r\theta y^{r-1} \exp\{-\theta y^r\} \;\; y > 0$$

1.1 a

A parametric family belongs to the exponential family of distributions if it can be written in the following form

$$p(y|\theta) = f(y)g(\theta)exp\{h(\theta)t(y)\}$$

In the case of this Weibull distribution r is known while θ is unknown. This means that r can be treated as a constant. In the expression, $f(y)$ is the normalizing constant. $t(y)$ is the sufficient statistic and $h(\theta)$ is the natural parameter. As concern the Weibull distribution it can be decoposed as follows:

- $f(y) = ry^{r-1}$
- $g(\theta) = \theta$
- $h(\theta) = -\theta$
- $t(y) = y^r$

Where θ is the canonical parameter and $t(y) = y^r$ is a sufficient statistic for the Weibull distribution. In order to compute a conjugate prior we have to use a result from the lecture notes. We can derive a conjugate prior distribution for the Weibull using the result of the exponential family. If a r.v. Y belongs to the exponential family of distributions a conjugate prior can be written in the following way

$$p(\theta) \propto g(\theta)^\nu \exp\{h(\theta)\delta\} \tag{1.1}$$

Where $g(\theta)$ and $h(\theta)$ have been already defined. The equation 1.1 can be clearly viewed as a likelihood based on ν independent observations and $\delta = \sum_{i=1}^{\nu} t(x_i)$.

With the Weibull we have that $g(\theta) = \theta$ and $h(\theta) = -\theta$. A conjugate prior for a Weibull distribution with r known is:

$$p(\theta) \propto \theta^\nu \exp\{-\theta\delta\} \;\text{ for } \theta > 0$$

This expression, function of θ, recall the kernel of a $gamma(\alpha, \beta)$ distribution, where $\nu = \alpha - 1$ and $\delta = \beta$. Finally, a conjugate prior for a weibull distribution with r known is a $Gamma(\alpha, \beta)$

1.2 b

From the Bayes' rule we know that the posterior distribution can be computed using the following formula:

$$\text{posterior distribution} \propto \text{prior distribution} \times \text{Likelihood}$$
$$p(\theta|\mathbf{x}) \propto p(\theta) \times p(\mathbf{x}|\theta)$$

We have already computed the conjugate prior distribution, we need to derive the expression for the likelihood.

Let's assume to have a random vector $Y = (Y_1, \ldots, Y_n)$ from which we have observed an i.i.d. sample $\mathbf{y} = (y_1, \ldots, y_n)$. We can write the likelihood function as the product of the marginals.

$$
\begin{aligned}
p(\mathbf{y}|\theta, r) = L(\theta|\mathbf{y}, r) &= \prod_{i=1}^{n} r\theta y_i^{r-1} \exp\{-\theta y_i^r\} \text{ for } \theta > 0 \\
&= r^n \theta^n \prod_{i=1}^{n} y_i^{r-1} \exp\{-\theta \sum_i y_i^r\} \;\; \theta > 0
\end{aligned}
$$

Where with $L(\theta|\mathbf{y}, r)$ we denoted the likelihood function. At this point using the conjugate prior and the likelihood, we can compute the posterior distribution.

$$
\begin{aligned}
&\text{posterior distribution} \propto \text{prior distribution} \times \text{Likelihood} \\
&\qquad p(\theta|\mathbf{y}) \propto p(\theta) \times p(\mathbf{y}|\theta, r) \\
&\qquad\qquad \propto \theta^{\alpha-1} \exp\{-\theta\beta\} \times r^n \theta^n \prod_{i=1}^{n} y_i^{r-1} \exp\{-\theta \sum_i y_i^r\} \text{ for } \theta > 0 \\
&\qquad\qquad \propto \theta^{\alpha-1} \exp\{-\theta\beta\} \times \theta^n \exp\{-\theta \sum_i y_i^r\} \; \theta > 0 \\
&\qquad\qquad \text{leaving out the elements that does not depend on } \theta \\
&\qquad\qquad = \theta^{n+\alpha-1} \exp\{-\theta(\beta + \sum_i y_i^r)\} \;\; \theta > 0
\end{aligned}
$$

Finally, the posterior distribution is $p(\theta|\mathbf{y}) \propto \theta^{n+\alpha-1} \exp\{-\theta(\beta + \sum_i y_i^r)\}$ which is the kernel of $Gamma(\alpha^* = n + \alpha, \beta^* = \beta + \sum_i y_i^r)$

1.3 c

Suppose that the two doctors, Andrea and Brian, agree to use the same $Gamma(5, 10)$ prior for θ

1.3.1 i

From a well known results in the lecture notes we know that a conjugate prior derived from the exponential family of distributions can be written as

$$
p(\theta) \propto g(\theta)^{\nu} e^{h(\theta)\delta}
$$

where $g(\theta)$, $h(\theta)$ have been already defined. This expression which is function of θ, can be interpreted as a likelihood of ν i.i.d. observations with $\delta = \sum_i t(y_i)$. In the previous part, we have proved that a conjugate prior for a Weibull distribution, with r known, is a Gamma.

In part c), Andrea and Brian agree to use the following prior distribution, a Gamma(5,10). Which can be written

$$
\begin{aligned}
p(\theta) &= \frac{\theta^{5-1} e^{-\theta 10} 10^5}{\Gamma(5)} \text{ for } \theta > 0 \\
&\propto \theta^{5-1} e^{-\theta 10} \text{ for } \theta > 0
\end{aligned}
$$

Which in terms of sample data can be interpreted as a likelihood function of a Weibull distribution based on 4 independent and identically distributed observations and $\sum_{i=1}^{\nu} y_i^r = 10$.

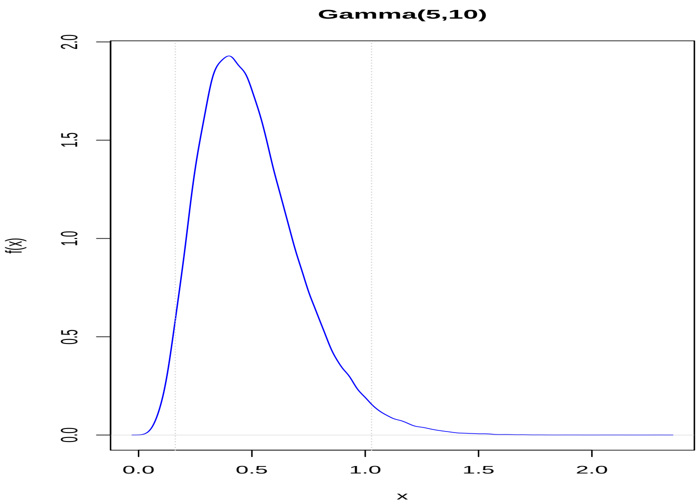

Figure 1.1: Density of a Gamma(5,10). With the grey-ish dashed line is represented the 95% interval

The specification of this prior means that the two doctors, according to their prior beliefs, expect θ to be non negative and distributed around an expected value of 0.5. In fact, the specification of a Gamma(5,10) lead to a distribution with an expectation $E[\theta] = \alpha/\beta = 5/10 = 0.5$ and with a variance $Var(\theta) = \alpha/\beta^2 = 5/100 = 0.05$. Furthermore, with a probability of 0.95 they believe that θ is in the range $[0.16, 1.02]$.

In other words, for Andrea and Brian is very unlikely to observe a value for θ greater than 1.02 and a value that is lower that 0.16, considering their prior belief. The choice of this prior intuitively means that the two doctors have reasonably enough prior information to not specify a vague prior. A final consideration can be done computing the Maximum likelihood estimator for θ, which can be proved to be $\hat{\theta} = \frac{n}{\sum_i Y^r}$. Interpreting the Gamma(5,10) as a Weibull likelihood we obtain a ML estimate $\hat{\theta} = 4/10 = 0.4$, which correspond to the mode of a Gamma(5,10): This support the idea that the choice of the prior is reasonably justified by some prior data that the two doctors could have collected or observed.

1.3.2 ii

Suppose of having observed two different sets of patients. Andrea's patients $\mathbf{a} = (a_1, \dots, a_n)$ and Brian's patients are $\mathbf{b} = (b_1, \dots, b_n)$. The Andrea's posterior can be written as:

$$
\begin{aligned}
\text{posterior distribution} &\propto \text{prior distribution} \times \text{Likelihood} \\
p(\theta|\mathbf{a}) &\propto p(\theta) \times p(\mathbf{a}|\theta) \\
&\quad \text{using the expression for the Likelihood derived before} \\
&\propto \theta^{\alpha-1} \exp\{-\theta\beta\} \times \prod_{i=1}^{n} r\theta a_i^{r-1} \exp\{-\theta a_i^r\} \;\; \theta > 0 \\
&= \theta^{\alpha-1} \exp\{-\theta\beta\} r^n \theta^n \prod_{i=1}^{n} a_i^{r-1} \exp\{-\theta \sum_i a_i^r\} \;\; \theta > 0 \\
&\propto \theta^{n+\alpha-1} \exp\{-\theta(\beta + \sum_i a_i^r)\} \;\; \theta > 0
\end{aligned}
$$

Finally the posterior distribution of Andrea is

$$p(\theta|\mathbf{a}) \propto \theta^{n+5-1} \exp\{-\theta(10 + \sum_i a_i^r)\} \text{ for } \theta > 0$$

We can observe that the posterior is proportional to the kernel of a $Gamma(n+\alpha, \beta + \sum_i a_i^r)$, thus, this is a proper posterior distribution.

The Brian's posterior is

$$\begin{aligned}
\text{posterior distribution} &\propto \text{prior distribution} \times \text{Likelihood} \\
p(\theta|\mathbf{b}) &\propto p(\theta) \times p(\mathbf{b}|\theta) \\
&\propto \theta^{\alpha-1} \exp\{-\theta\beta\} \times \prod_{i=1}^{n} r\theta b_i^{r-1} \exp\{-\theta b_i^r\} \;\; \theta > 0 \\
&= \theta^{\alpha-1} \exp\{-\theta\beta\} r^n \theta^n \prod_{i=1}^{n} b_i^{r-1} \exp\{-\theta \sum_i b_i^r\} \;\; \theta > 0 \\
&\propto \theta^{n+\alpha-1} \exp\{-\theta(\beta + \sum_i b_i^r)\} \;\; \theta > 0
\end{aligned}$$

Finally the posterior distribution of Brian is

$$p(\theta|\mathbf{b}) \propto \theta^{n+5-1} \exp\{-\theta(10 + \sum_i b_i^r)\} \text{ for } \theta > 0$$

Also in this case, we can recognize the kernel of a Gamma distribution, precisely a $Gamma(n+\alpha, \beta + \sum_i b_i^r)$, thus, this is a proper posterior distribution. The posterior distribution that we have obtained is not very different from the Andre's ones. The kernel are practically identical, both a Gamma, the only difference lies in the rate parameter which are different.

1.3.3 iii

Andrea now tells Brian the survival times $\mathbf{a}$ of his patients, and, in exchange, Brian tells Andrea the survival times of his patients, $\mathbf{b}$,. Both can now use this information to update their belief about θ.

In the prevoius point, we have derived the posterior for Biran $p(\theta|\mathbf{b})$, considering now the new information available, this posterior is now Brian's new prior. This means that the new information available can be used to update Brian's belief. To explain this, we can consider the posterior distribution of Brian after having observed Andrea's patient.

$$p(\theta|\mathbf{b}, \mathbf{a}) \propto p(\mathbf{b}, \mathbf{a}|\theta) \times p(\theta)$$

As $\mathbf{b}$ and $\mathbf{a}$ are two independent sample we can express the right hand side as

$$p(\mathbf{b}, \mathbf{a}|\theta) \times p(\theta) = p(\mathbf{a}|\theta) \underbrace{p(\mathbf{b}|\theta)p(\theta)}_{p(\theta|\mathbf{b})} \tag{1.2}$$

We can notice from equation 1.2 that $p(\mathbf{b}|\theta)p(\theta)$ in the Brian's posterior. Meaning that we can compute the new posterior distribution by simply multiplying the Brian's posterior with the likelihood function of a Weibull distribution after having observed the sample $\mathbf{a}$. We have then

$$
\begin{aligned}
p(\theta|\mathbf{b},\mathbf{a}) &\propto p(\theta|\mathbf{b}) \times p(\mathbf{a}|\theta) \\
&\propto \theta^{n+\alpha-1} exp\{-\theta(\beta + \sum_i b_i^r\} \times \theta^n \exp\{-\theta \sum_i a_i^r\} \\
&= \theta^{2n+\alpha-1} \exp\{-\theta(\beta + \sum_i b_i^r + \sum_i a_i^r)\}
\end{aligned}
$$

The Brian's posterior distribution is

$$
p(\theta|\mathbf{b},\mathbf{a}) \propto \theta^{2n+\alpha-1} \exp\{-\theta(\beta + \sum_i b_i^r + \sum_i a_i^r)\} \;\; \theta > 0
$$

From which we recognize a $Gamma(2n + \alpha, \beta + \sum_i b_i^r + \sum_i a_i^r)$.

Using the same reasoning we can derive the posterior of Andrea

$$
\begin{aligned}
p(\theta|\mathbf{a},\mathbf{b}) &\propto p(\theta|\mathbf{a}) \times p(\mathbf{b}|\theta) \\
&\propto \theta^{n+\alpha-1} exp\{-\theta(\beta + \sum_i a_i^r\} \times \theta^n \exp\{-\theta \sum_i b_i^r\} \\
&= \theta^{2n+\alpha-1} \exp\{-\theta(\beta + \sum_i b_i^r + \sum_i a_i^r)\}
\end{aligned}
$$

The Andrea's posterior is

$$
p(\theta|\mathbf{a},\mathbf{b}) \propto \theta^{2n+\alpha-1} \exp\{-\theta(\beta + \sum_i b_i^r + \sum_i a_i^r)\} \;\; \text{for } \theta > 0
$$

From which we recognize a $Gamma(2n + \alpha, \beta + \sum_i b_i^r + \sum_i a_i^r)$, the exactly same posterior that we found for Brian.

1.3.4 iv

The two posterior are identical and will lead to the exactly same inference about theta. Theoretically, to update the belief of the two doctors we are using the probability theory which is a coherent method. In fact, in whichever way we do the conditioning the result does not change. In other words, the two doctors at this stage have the same amount of information about the phenomenon. This element and the fact that they both used the same prior, lead to the same posterior distribution in part iii).

1.4 d)

To express the posterior probability asked, we need to derive the posterior predictive distribution.

$$
p(\tilde{y}|\mathbf{y}) = \int_{\theta\in\Theta} p(\tilde{y}|\theta)p(\theta|\mathbf{y})d\theta
$$

Where with Θ we defined the parametric space of θ and with $\tilde{y}$ the new information.

$$p(\tilde{y}|\mathbf{y}) = \int_0^{+\infty} \underbrace{r\theta\tilde{y}^{r-1}e^{-\theta\tilde{y}^r}}_{p(\tilde{y}|\theta)} \underbrace{\frac{\theta^{2n+\alpha-1}e^{-\theta(\beta+\sum_i b_i^r+\sum_i a_i^r)}(\beta+\sum_i b_i^r+\sum_i a_i^r)^{(2n+\alpha)}}{\Gamma(2n+\alpha)}}_{p(\theta|\mathbf{y})} d\theta$$

$$= \frac{(\beta+\sum_i b_i^r+\sum_i a_i^r)^{(2n+\alpha)}}{\Gamma(2n+\alpha)} r\tilde{y}^{r-1} \int_0^{+\infty} \theta^{(2n+\alpha+1)-1}e^{-\theta(\beta+\sum_i b_i^r+\sum_i a_i^r+\tilde{y}^r)}d\theta$$

inside the integral we can recognize the kernel of a gamma distribution

Let $\frac{(\beta+\sum_i b_i^r+\sum_i a_i^r)^{(2n+\alpha)}}{\Gamma(2n+\alpha)}r = C$. We can now multiply and divide by $\Gamma(2n+\alpha+1)$ and $(\beta + \sum_i b_i^r + \sum_i a_i^r + \tilde{y}^r)^{(2n+\alpha+1)}$ to complete the integral.

$$p(\tilde{y}|\mathbf{y}) = C\frac{\tilde{y}^{r-1}\Gamma(2n+\alpha+1)}{(\beta+\sum_i b_i^r+\sum_i a_i^r+\tilde{y}^r)^(2n+\alpha+1)}\times$$
$$\times \underbrace{\int_0^{+\infty} \frac{\theta^{(2n-\alpha+1)-1}e^{-\theta(\beta+\sum_i b_i^r+\sum_i a_i^r+\tilde{y}^r)}(\beta+\sum_i b_i^r+\sum_i a_i^r+\tilde{y}^r)^{(2n+\alpha+1)}}{\Gamma(2n+\alpha+1)}d\theta}_{Gamma(2n+\alpha+1,\beta+\sum_i b_i^r+\sum_i a_i^r+\tilde{y}^r)=1}$$

As the terms under the integral sign integrate to 1, we obtain the following expression for the posterior predictive distribution:

$$p(\tilde{y}|\mathbf{y}) = \frac{(\beta+\sum_i b_i^r+\sum_i a_i^r)^{(2n+\alpha)}\Gamma(2n+\alpha+1)r\tilde{y}^{r-1}}{(\beta+\sum_i b_i^r+\sum_i a_i^r+\tilde{y}^r)^{(2n+\alpha+1)}\Gamma(2n+\alpha)}$$

simplifying the factorials we obtain

$$p(\tilde{y}|\mathbf{y}) = \frac{(\beta+\sum_i b_i^r+\sum_i a_i^r)^{(2n+\alpha)}(2n+\alpha)r\tilde{y}^{r-1}}{(\beta+\sum_i b_i^r+\sum_i a_i^r+\tilde{y}^r)^{(2n+\alpha+1)}} \quad \tilde{y} > 0$$

This expression in the predictive posterior distribution which is function of $\tilde{y}$. We can now express the probability that a newly diagnosed patient will die within one year. Which can be rephrased in, compute the probability that the survival time of a new patient Y is less than 1.

$$Pr(\tilde{y} < 1|\mathbf{y}) = A\int_0^1 \frac{\tilde{y}^{r-1}}{(B+\tilde{y}^r)^C}d\tilde{y}$$

where:

- $A = (\beta + \sum_i b_i^r + \sum_i a_i^r)^{(2n+\alpha)}(2n+\alpha)r$
- $B = \beta + \sum_i b_i^r + \sum_i a_i^r$
- $C = 2n + \alpha + 1$

2 Exercise 2

Let Y denote the income of a random high income individual living in the country. And assume that Y can be described by a Pareto distribution:

$$p_Y(y|\theta, k) = \theta k^\theta y^{-(\theta+1)} \quad y > k$$

2.1 a

We can show that $X = \log(Y/k)$ has an exponential distribution. Using the usual formula for transformation of random variables

$$f_X(x) = f_Y(y)(g^{-1}(x)) \left| \frac{dg^{-1}(x)}{dx} \right|$$

We have then $e^X = \frac{Y}{k} \to Y = ke^X$, we can then easily compute the jacobian $\frac{dy}{dx} = ke^X$. With $x \in (0, +\infty)$

$$\begin{aligned} f_X(x|\theta, k) &= f_Y(y|\theta, r) \left| \frac{dy}{dx} \right| \\ &= \theta k^\theta (ke^x)^{-(\theta+1)} ke^x \\ &= \cancel{k^{\theta+1}} \cancel{k^{-(\theta+1)}} \theta e^{-x(\theta+1)+x} \\ &= \theta e^{-x\theta} \end{aligned}$$

Then we obtain

$$f_X(x|\theta, k) = \theta e^{-x\theta} \quad x > 0$$

Which is clearly the pdf of an $Exp(\theta)$.

2.2 b

Assume that a sample of n high income is observed and their income y_i $i = 1, \ldots, n$ are recorded and assuming that we have observed an i.i.d. sample, we can write the likelihood function as the product of the marginals.

$$\begin{aligned} p(\mathbf{x}|\theta) = L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \end{aligned}$$

Using a Gamma(a,b) prior and the standard result from the Bayes' theorem we can compute the posterior distribution.

$$\begin{aligned} \text{posterior distribution} &\propto \text{prior distribution} \times \text{Likelihood} \\ p(\theta|\mathbf{x}) &\propto p(\theta) \times p(\mathbf{x}|\theta) \\ &\propto \theta^{a-1} e^{-b\theta} \times \prod_{i=1}^{n} \theta e^{-\theta x_i} \\ &= \theta^{a-1} e^{-b\theta} \times \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \\ &= \theta^{a+n-1} e^{-\theta(b + \sum_{i=1}^{n} x_i)} \quad \theta > 0 \end{aligned}$$

We can recognize inthe posterior distribution the kernel of a $Gamma(\alpha^* = \alpha + n, \beta^* = b + \sum_{i=1}^{n} x_i)$.
Where $A = \alpha + n$ and $B = b + \sum_{i=1}^{n} x_i$.

$$
\begin{aligned}
B &= b + \sum_{i=1}^{n} x_i \\
&\text{knwing that } \; x_i = \log{(y_i/k)} \\
&= b + \sum_{i=1}^{n} \log \frac{y_i}{k} \\
&\text{using the property of the log} \\
&= b + \sum_{i=1}^{n} \log y_i - n \log k \\
&= b + n \left[\frac{1}{n} \sum_{i=1}^{n} \log y_i - \log k \right] \\
&\text{we can apply } \; \exp \text{ and } \log \\
&= b + n \left\{ \log \left[\exp{(n^{-1} \sum_{i=1}^{n} \log y_i - \log k)} \right] \right\} \\
&\text{where } \; -\log{(k)} = \log k^{-1} = \log 1/k \\
&= b + n \left\{ \log \left[\exp{(n^{-1} \sum_{i=1}^{n} \log y_i)} \exp{(\log{(\frac{1}{k})})} \right] \right\} \\
&= b + n \left\{ \log \left[\frac{\exp{\{n^{-1} \sum_{i=1}^{n} \log y_i\}}}{k} \right] \right\}
\end{aligned}
$$

Where we have expressed $B = b + n \log \frac{G}{k}$, where $G = \exp{\{n^{-1} \sum_{i=1}^{n} \log y_i\}}$

2.3 c

A government is considering a tax policy. Let T_1 and T_2 the amount of tax under each regime, considering only on the population whose income is greater that k, we can define T_1 and T_2 as:

$$
T_1 = \begin{cases} t & \text{if } y > k \\ 0 & \text{otherwise} \end{cases}
$$

Under the first regime, if a person have an income greater than k then he or she pays a fixed amount t. While if he or she has an income $< k$ the individual does not pay this pool tax.

Under the second regime we have:

$$
T_2 = \begin{cases} s & \text{if } y > m \\ 0 & \text{otherwise} \end{cases}
$$

if a person has an income greater than m he or she pays a fixed amount of taxes. Now for each policy we want to compute the expected tax paid by a randomly chosen individual conditional on θ

Under the first regime we have that $E[T_1|\theta] = tPr(y > k|\theta) = t \times 1$, which is simply t as we are considering in our analysis only individuals who have income greater than k. As concern the second regime we have that $E[T_2|\theta] = sPr(y > m|\theta)$. This probability can be computed using the income distribution. Then we have

$$
\begin{aligned}
Pr(y > m|\theta) &= \int_m^{+\infty} \theta k^\theta y^{-(\theta+1)} dy \\
&= \theta k^\theta \int_m^{+\infty} y^{-(\theta+1)} dy \\
&= \theta k^\theta \left| \frac{y^{-(\theta+1)+1}}{-(\theta+1)+1} \right|_m^{+\infty} \\
&= \theta k^\theta \left| \frac{y^{-\theta}}{-\theta} \right|_m^{+\infty} \\
&= -k^\theta \left[0 - m^{-\theta} \right] = \left(\frac{k}{m} \right)^\theta
\end{aligned}
$$

Finally, $E[T_2|\theta] = s\left(\frac{k}{m}\right)^\theta$ where $\frac{k}{m} < 1$ this since $k < m$

2.4 d

Under the first tax regime we can define a loss function as

$$Loss_1 = -E[T_1|\theta] = -tPr(y > k|\theta)$$

Under the second tax regime, the loss function can be defined as

$$Loss_2 = -E[T_2|\theta] = -sPr(y > m|\theta)$$

Taking the expectation respect to the posterior distribution er can then compute the expected posterior loss. We can write

$$
\begin{aligned}
E_{\theta|\mathbf{x}}[-E[T_i|\theta]] &= -\delta_i \int_{\theta \in \Theta} Pr(y > \nu_i|\theta) p(\theta|\mathbf{y}) d\theta \\
&\text{we can write in an explicit way that probability} \\
&- \delta_i \int_{\theta \in \Theta} \int_{\nu_i}^{+\infty} p(y|\theta) p(\theta|\mathbf{y}) dy d\theta \\
&\text{using the Fubini theorem we can write} \\
&- \delta_i \int_{\nu_i}^{+\infty} \left[\int_{\theta \in \Theta} p(y|\theta) p(\theta|y) d\theta \right] dy \\
&\text{we can recognize the posterior predictive distribution inside the integral} \\
&- -\delta_i \int_{\nu_i}^{+\infty} p(\tilde{y}|\mathbf{y}) dy \\
&= -\delta_i Pr(\tilde{y} > \nu_i|\mathbf{y})
\end{aligned}
$$

Where $\delta_i = t, s$, $\nu_i = k, m$ and $E_{\theta|\mathbf{x}}[.]$ denotes the expectation taken respect to the posterior distribution.

We can recognize $Pr(\tilde{y} > \nu_i|\mathbf{y})$ be a probability computed under the posterior predictive distribution. We can conclude that in order to compute the expected posterior loss we need to find an expression for the predictive posterior distribution. This is reasonable as, under the Bayesian domain, we use the posterior predictive distribution to predict the outcome tanking into account the uncertainty about θ. To compute the posterior predictive we can use the result derived in part a), namely $X = \log(Y/k)$.

Denoting with $\tilde{x}$ a new observation of a randomly selected individual, we can express the posterior predictive as:

$$p(\tilde{x}|\mathbf{x}) = \int_{\theta\in\Theta} p(\tilde{x}|\theta)p(\theta|\mathbf{x})d\theta \quad \tilde{x} > 0$$

Where Θ denote the parametric space, $p(\tilde{x}|\theta)$ is the exponential density evaluated at $\tilde{x}$ and $p(\theta|\mathbf{x})$ is the posterior distribution derived in part b)

$$= \int_0^{+\infty} \theta e^{-\tilde{x}\theta}.\frac{\theta^{(a+n)-1}(b+\sum_i x_i)^{(a+n)}e^{-\theta(b+\sum_i x_i)}}{\Gamma(a+n)}d\theta$$

putting out all the terms that does not depend on θ

$$= \frac{(b+\sum_i x_i)^{(a+n)}}{\Gamma(a+n)}\int_0^{+\infty} \theta^{(a+n+1)-1}e^{-\theta(b+\sum_i x_i+\tilde{x})}d\theta$$

Under the integral sign we can recognize the kernel of a gamma distribution, multiplying and dividing by $\Gamma(\alpha+n+1)$ and by $(b+\sum_i x_i+\tilde{x})^{(a+n+1)}$ we can complete the integral as follows:

$$= \frac{(b+\sum_i x_i)^{(a+n)}\Gamma(\alpha+n+1)}{(b+\sum_i x_i+\tilde{x})^{(a+n+1)}\Gamma(a+n)}\underbrace{\int_0^{+\infty}\frac{\theta^{(a+n+1)-1}(b+\sum_i x_i+\tilde{x})^{(a+n+1)}e^{-\theta(b+\sum_i x_i+\tilde{x})}}{\Gamma(a+n+1)}d\theta}_{=1}$$

Finally, the posterior predictive distribution is

$$p(\tilde{x}|x) = \frac{(b+\sum_i x_i)^{(a+n)}(a+n)}{(b+\sum_i x_i+\tilde{x})^{(a+n+1)}} \quad \tilde{x} > 0$$

Where we have simplified the ratio between the two gamma functions.

We can now compute the expected posterior loss under each policy.

Under the first policy

$$E_{\theta|\mathbf{x}}[Loss_1] = -tPr(\tilde{y} > k|\mathbf{y}) = -tPr(\tilde{x} > 0|\mathbf{x}) = t \times 1$$

As we are considering only the individuals that have income greater that k than $p(\tilde{y} > k|\mathbf{y}) = p(\tilde{x} > 0|\mathbf{x}) = 1$

Under the second regime we have

$$E_{\theta|\mathbf{x}}[Loss_2] = -s\ Pr(\tilde{y} > m|\mathbf{y}) = -s\ Pr(\tilde{x} > \log(m/k)|\mathbf{x})$$

Where the last expression comes for the fact that $X = \log(\frac{Y}{k})$, then if $Y > m$ then $X > \log\frac{m}{k}$ and the inequity does not change since the natural logarithm is a monotone increasing function.

$$
\begin{aligned}
Pr(\tilde{x} > \log\left(\frac{m}{k}\right)|\mathbf{x}) &= \int_{\log\frac{m}{k}}^{+\infty} (b+\sum_i x_i)^{(a+n)}(a+n)(b+\sum_i x_i+\tilde{x})^{-(a+n+1)}d\tilde{x} \\
&= (a+n)(b+\sum_i x_i)^{(a+n)} \left| \frac{(b+\sum_i x+\tilde{x})^{-(a+n+1)-1}}{-(a+n)} \right|_{log\frac{m}{k}}^{+\infty} \\
&= (a+n)(b+\sum_i x_i)^{(a+n)} \left(0+\frac{(b+\sum_i x_i+\log\frac{m}{k})^{-(a+n)}}{(a+n)}\right) \\
&= \frac{\cancel{(a+n)}(b+\sum_i x_i)^{(a+n)}(b+\sum_i x_i+\log\frac{m}{k})^{-(a+n)}}{\cancel{(a+n)}} \\
&= (b+\sum_i x_i)^{(a+n)}(b+\sum_i x_i+\log\frac{m}{k})^{-(a+n)}
\end{aligned}
$$

Finally we have that the expected posterior loss under the first regime is

$$E_{\theta|\mathbf{x}}[Loss_1] = -t$$

Under the second regime is

$$E_{\theta|\mathbf{x}}[Loss_2] = -s\left[\frac{(b+\sum_i x_i)}{(b+\sum_i x_i+\log\left(\frac{m}{k}\right))}\right]^{(a+n)}$$

The second regime would be more profitable if the loss under the second regime is lower that the loss under the status quo. In other words, if this inequity holds.

$$E_{\theta|\mathbf{x}}[Loss_2] - E_{\theta|\mathbf{x}}[Loss_1] < 0$$

Let $A = (a+n)$ and $B = (b+\sum_i x_i)$ and substituting the expressions just derived for the posterior expected loss. We can prove that new policy is worthwhile only if:

$$
\begin{aligned}
&-s\left[\frac{(b+\sum_i x_i)}{(b+\sum_i x_i+\log\left(\frac{m}{k}\right))}\right]^{(a+n)} - (-t) < 0 \\
&-s\left[\frac{B}{B+\log\left(\frac{m}{k}\right)}\right]^{A} + t < 0 \\
&\quad \text{changing the sign} \\
&\quad s\left[\frac{B}{B+\log\left(\frac{m}{k}\right)}\right]^{A} > t \\
&\quad s > t\left[\frac{B}{B+\log\left(\frac{m}{k}\right)}\right]^{-A} \\
&\quad s > t\left[1+\frac{\log\left(\frac{m}{k}\right)}{B}\right]^{A}
\end{aligned}
$$

In fact, we recall that m, k and t are fixed quantities and $\sum_i x_i = \sum_i \log(y_i/k)$ comes from the observed sample of high income individuals. Thus, the only undefined quantity is s. We can note that in the second line the inequity does not change as $a, b > 0$ (they are the parameters of the Gamma function and are supposed to be strictly positive), $n > 0$ and $\sum_i x_i$ as they are a transformation of the income who is supposed to be a non negative quantity. This means that the term in square brackets is positive.

3 Exercise 3

3.1 a

When a patient with Multiple Sclerosis are given a dose of the drug Campath their lymphocyte count falls temporarily. It then recovers slowly to its original level. Four measurements of the lymphocyte count are taken to patient i, at time $t_{i1}, t_{i2}, t_{i3}, t_{i4}$ after administration of the drug. A model for the lymphocyte count is as follows.

For $i = 1, \dots, n$ and $j = 1, 2, 3, 4$

$$
\begin{aligned}
&Y_{ij}|\alpha_i\beta_i, t_{ij} \sim N(\mu_{ij}, \tau^{-1})\\
&\mu_{ij} = \alpha_i - \frac{\beta_i}{t_{ij}+1}\\
&\alpha_i|\mu_\alpha, \tau_\alpha \sim N(\mu_\alpha, \tau_\alpha^{-1})\\
&\beta_i|\mu_\beta, \tau_\beta \sim N(\mu_\beta, \tau_\beta^{-1})
\end{aligned}
$$

Where Y_{ij} is the lymphocyte count of the i^{th} patient at time t_{ij} The following independent and vague prior distribution are assumed for the hyperparameters of the model.

$$
\begin{aligned}
&\mu_\alpha \sim N(0, 10^6)\\
&\mu_\beta \sim N(0, 10^6)\\
&\tau_\alpha \sim Gamma(0.001, 0.001)\\
&\tau_\beta \sim Gamma(0.001, 0.001)\\
&\tau \sim Gamma(0.001, 0.001)
\end{aligned}
$$

The model can we expressed in the following DAG:

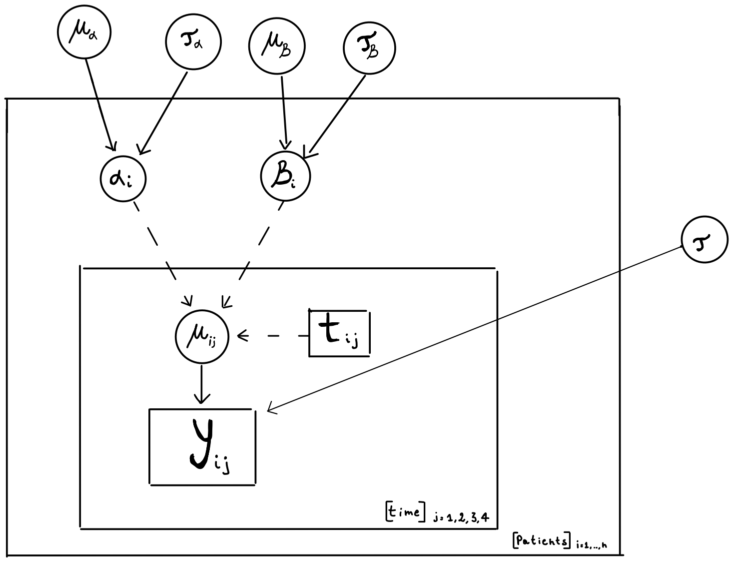

Figure 3.1: DAG, the dashed arrows denote deterministic dependencies

In this DAG is represented the model specified in the exercise 3. With circular nodes we have represented unknown r.vs, while with square nodes are represented known r.vs. This DAG has two repetitive structures, one for the patients and one for the time at which the measurements are taken. Finally, with solid lines we denoted stochastic dependencies, while with dashed arrows we denoted deterministic dependencies.

3.2 b

As concern the parameter interpretation:

α_i $i = 1, \ldots, n$ can be interpreted as the original level of lymphocyte in the i^{th} patient. With this hierarchical structure, we are assuming that each patient has his own original level of lymphocyte which is not effected by the drug.

β_i $i = 1, \ldots, n$ can be interpreted as the marginal expected change in the level or lymphocytes in the i^{th} patient when the time t_{ij} changes. From the description of the model, we can notice that the measurements of lymphocyte, for each patient, are taken in 4 different moments. The doctor expect that the falls in the level of lymphocyte is temporary. Thus, as the time passes, each patient is expected to recover more and more lymphocyte. In fact, when t_{ij} increases, $\frac{\beta_i}{t_{ij}+1} \to 0$.

μ_α is the location parameter of the distribution of α_i, distribution which is assumed equal across all the n patients. Meaning that we are assuming that the mean level of lymphocyte, not effected by the drug, across all the n patients is μ_α. Defining a common prior distribution for all the α_i's, we are not only assuming similarity in the original level of lymphocytes in each i^{th} patient, but we are also assuming conditional independence of the α_i given $(\mu_\alpha, \tau_\alpha^{-1})$.

3.3 c

3.3.1 i

The exchangeability can be justified given that, it is reasonable to assume that the baseline level of lymphocytes α_i is similar across all the patients. This seems reasonable as we don't have any further information about the patients, so we are not able to distinguish them. For the same reason, it is reasonable to think that the degree of lymphocytes recovery after taking the drug is similar in all patients.

In this model we are assuming some sort of hierarchical structure, we have in fact n specific parameters for each patient (α_i, β_i) and some overall hyperparameters $(\mu_\alpha, \tau_\alpha, \mu_\beta, \tau_\beta)$. The full joint distribution can be expressed as:

$$p(\mu_\alpha, \tau_\alpha, \mu_\beta, \tau_\beta, \tau, \alpha_i, \beta_i, y_{ij}) \propto p(\mu_\alpha)p(\tau_\alpha)p(\mu_\beta)p(\tau_\beta) \prod_{i=1}^{n} p(\alpha_i|\mu_\alpha, \tau_\alpha) \prod_{i=1}^{n} p(\beta_i|\mu_\beta, \tau_\beta) \times \prod_{i=1}^{n} \prod_{j=1}^{4} p(y_{ij}|\alpha_i, \beta_i, t_{ij})$$

We are assuming that given the hyperparameters the patients specific parameters are independent.

As concern the estimation, adopting this hierarchical structure, we are implying some sort of correlation among the patients specific parameters. This means that we are not estimating α_i and β_i just from the data about the i^{st} patient, we are estimating $(\alpha_i\ , \beta_i)$ directly from the data of the i^{st} patient and indirectly from the other source of evidence.

3.3.2 ii

There are two main reasons why these priors have been chosen:

1) Firstly, the priors for $\mu_\alpha, \tau_\alpha, \tau_\beta$ and τ were chosen because we are assuming a normal distribution for $\alpha_i, \beta_i, \varepsilon_{ij}$

Starting from the precision parameters τ_α, τ_β and τ, we have the constraint that the precision for a Gaussian distribution have to be greater than zero, $\sigma^2 = \tau^{-1} > 0$. An easy way to impose this constraint is to specify a prior with a positive support. A convenient choice is to adopt a Gamma distribution as a prior for τ_α, τ_β and τ .

For similar reasons μ_α has been specified. We can note that μ_α is the location parameter of the distribution of α_i. It's known that a valid range of values for the location parameter for a Normal distribution is the entire real line. As the normal distribution as a support $S \in (-\infty, +\infty)$ it make sense to specify a normal prior for μ_α.

Finally, under this specification, the data have a normal distribution $Y_{ij}|\alpha_i, \beta_i, t_{ij} \sim N(\alpha_i - \frac{\beta_i}{t_{ij}+1}, \tau^{-1})$, this means that the independent vague priors distributions specified for the location and the dispersion parameters are conjugate priors. This means that the posterior distributions for μ_α is expected to be Normal, while for τ_α, τ_β, τ is expected to be Gamma.

2)Secondly, the priors for $\mu_\alpha, \tau_\alpha, \tau_\beta$ and τ are chosen as we are assuming exchangeability. In other words, we are assuming that the patients are similar, not equal, to each other. This is reasonable if we consider that these data comes from the same study or we imagine we are considering patients from the same hospital.