Exercise 1: (chapter 6.2) Let $\{Y_i\}_{i=1}^n$ be independent and identically distributed random variables that follow a Bernoulli distribution with parameter $0 \leq \theta \leq 1$. The probability mass function of the Bernoulli distribution is

$$p_Y(y) = \Pr(Y = y) = \theta^y (1-\theta)^{1-y} \qquad y \in \{0, 1\}\,.$$

Let the prior on θ be improper with density $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$.

1. Find the posterior $p(\theta \,|\, y)$ and the corresponding normal approximation at its mode.
2. Show that the improper prior on θ is equivalent to a uniform prior on the logit $\beta = \log\{\theta/(1-\theta)\}$.
3. Find the posterior $p(\beta \,|\, y)$ and the corresponding normal approximation at its mode.
4. Is it more sensible to derive a normal approximation on the probability or logit scale?

Solution: The posterior density is

$$p(\theta \,|\, y) \propto \left[\prod_{i=1}^n \theta^{y_i} (1-\theta)^{1-y_i}\right] \theta^{-1}(1-\theta)^{-1} = \theta^{n\bar{y}-1}(1-\theta)^{n-n\bar{y}-1}\,,$$

which is the kernel of the density of a $\mathrm{Beta}(n\bar{y}, n-n\bar{y})$ distribution. The mode of the posterior density is

$$\frac{\mathrm{d}\log p(\theta \,|\, y)}{\mathrm{d}\theta} = \frac{n\bar{y}-1}{\theta} - \frac{n-n\bar{y}-1}{1-\theta} \overset{!}{=} 0 \Rightarrow \hat{\theta}_{\mathrm{MAP}} = \frac{n\bar{y}-1}{n-2}\,,$$

which is readily known from the properties of the Beta distribution as $(\alpha-1)/(\alpha+\beta-2)$. The observed Fisher information is

$$\mathcal{J}(\theta) = -\frac{\mathrm{d}^2 \log p(\theta \,|\, y)}{\mathrm{d}\theta^2} = \frac{n\bar{y}-1}{\theta^2} + \frac{n-n\bar{y}-1}{(1-\theta)^2}\,,$$

At the mode, the observed Fisher information is

$$\mathcal{J}(\theta)|_{\theta=\hat{\theta}_{\mathrm{MAP}}} = \frac{(n-2)^2}{n\bar{y}-1} + \frac{(n-2)^2}{n-n\bar{y}-1}\,.$$

The normal approximation to $p(\theta \,|\, y)$ is therefore

$$\theta \,|\, y \overset{\text{approx.}}{\sim} \mathrm{Normal}\left(\theta \,\middle|\, \frac{n\bar{y}-1}{n-2}, \left[\frac{(n-2)^2}{n\bar{y}-1} + \frac{(n-2)^2}{n-n\bar{y}-1}\right]^{-1}\right).$$

The prior density on β is

$$p(\beta) = p(g(\beta))\left|\frac{\theta}{\beta}\right| \propto \left(\frac{e^\beta}{1+e^\beta}\right)^{-1}\left(1-\frac{e^\beta}{1+e^\beta}\right)^{-1}\frac{e^\beta}{(1+e^\beta)^2} = \frac{\left(1+e^\beta\right)^2}{e^\beta}\frac{e^\beta}{\left(1+e^\beta\right)^2} = 1\,,$$

which is an improper uniform prior density. The posterior density is then

$$p(\beta \mid y) \propto \prod_{i=1}^{n} \left(\frac{e^{\beta}}{1+e^{\beta}} \right)^{y_i} \left(1 - \frac{e^{\beta}}{1+e^{\beta}} \right)^{1-y_i} = \frac{e^{\beta n \bar{y}}}{(1+e^{\beta})^n} \,.$$

The mode of the posterior density is

$$\frac{\mathrm{d} \log p(\beta \mid y)}{\mathrm{d}\beta} = n\bar{y} - \frac{ne^{\beta}}{1+e^{\beta}} \overset{!}{=} 0 \Rightarrow \hat{\beta}_{\mathrm{MAP}} = \log\left\{ \frac{\bar{y}}{1-\bar{y}} \right\} .$$

The observed Fisher information is

$$\mathcal{J}(\beta) = -\frac{\mathrm{d}^2 \log p(\beta \mid y)}{\mathrm{d}\beta^2} = \frac{ne^{\beta}}{(1+e^{\beta})^2} \,.$$

At the mode, the observed Fisher information is

$$\mathcal{J}(\beta)\big|_{\beta=\hat{\beta}_{\mathrm{MAP}}} = n\bar{y}(1-\bar{y}) \,.$$

The normal approximation to $p(\beta \mid y)$ is therefore

$$\beta \mid y \overset{\text{approx.}}{\sim} \mathrm{Normal}\left(\theta \,\middle|\, \log\left\{ \frac{\bar{y}}{1-\bar{y}} \right\}, \frac{1}{n\bar{y}(1-\bar{y})} \right) .$$

The logit β ranges from $-\infty$ to ∞. It could therefore be more sensible to approximate $p(\beta \mid y)$ by a normal approximation since the support and parameter space agree.

Exercise 2 (chapter 6.2): Let $\{Y_i\}_{i=1}^n$ be independent and identically distributed random variables that follow a Poisson distribution with rate parameter $\lambda > 0$. The probability mass function of the Poisson distribution is

$$p_Y(y) = \Pr(Y = y) = \frac{\lambda^y}{y!} \exp\{-\lambda\} \qquad y = 0, 1, \ldots$$

Assume that $\mathbb{E}[\lambda] = 2$ and $\Pr(\lambda > 3) = 0.01$.

1. Describe the prior on λ by a normal distribution and find the posterior $p(\lambda \mid y)$.
2. Derive a normal approximation to the posterior $p(\lambda \mid y)$ at its mode using 100 Poisson observations

y_i	0	1	2	3	4	5	≥ 6
#	18	32	27	15	6	2	0

   and compute the posterior probability $\Pr(\lambda > 2 \mid y)$.
3. Although $\lambda > 0$, the support of the normal prior on λ is unconstrained. Which reparameterization under the bijection $\theta = g(\lambda) \Leftrightarrow \lambda = h(\theta)$ would yield an unconstrained parameter? Describe the prior on θ by a normal distribution using $\mathbb{E}[\theta] = \log 2$ and $\Pr(\theta > \log 3) = 0.01$ and find the posterior $p(\theta \mid y)$.
4. Derive a normal approximation to the posterior $p(\theta \mid y)$ at its mode using same data as above and

compute the posterior probability $\Pr(\lambda > 2 \,|\, y)$ by translating back to the original parameter space (you may use R to find the mode and observed Fisher information).

Solution: It is know that $\mathbb{E}[\lambda] = 2$ and $\Pr(\lambda > 3) = 0.01$ so that

$$\Pr(\lambda > 3) = \Pr\left(\frac{\lambda - 2}{\sigma} > \frac{3-2}{\sigma}\right) = \Pr\left(Z > \frac{1}{\sigma}\right) = 0.01 \Rightarrow \frac{1}{\sigma} = \Phi^{-1}(0.99) = 2.33\,.$$

The parameters of the Normal prior on λ are thus $\mu = 2$ and $\sigma^2 = 0.18$. The posterior density is

$$p(\lambda \,|\, y) \propto \left[\prod_{i=1}^{n} \lambda^{y_i} \exp\{-\lambda\}\right] \exp\Big\{-2.63(\lambda-2)^2\Big\} = \lambda^{n\bar{y}} \exp\big\{-2.63\lambda^2 + (10.52 - n)\lambda\big\}$$

The mode of the posterior density is

$$\frac{\mathrm{d}\log p(\lambda \,|\, y)}{\mathrm{d}\lambda} = \frac{n\bar{y}}{\lambda} - 5.24\lambda + 10.52 - n \overset{!}{=} 0$$

$$\Rightarrow \hat{\lambda}_{\text{MAP}} = \frac{-(n-10.52) + \sqrt{(n-10.52)^2 + 4(5.24)(n\bar{y})}}{2(5.24)} \qquad \text{since } \lambda > 0$$

For the above data, the mode is $\hat{\lambda}_{\text{MAP}} = 1.67$. The observed Fisher information is

$$\mathcal{J}(\lambda) = -\frac{\mathrm{d}^2\log p(\lambda \,|\, y)}{\mathrm{d}\lambda^2} = \frac{n\bar{y}}{\lambda^2} + 5.24\,.$$

At the mode, the observed Fisher information is $\mathcal{J}(\lambda)|_{\lambda=\hat{\lambda}_{\text{MAP}}} = 63.78$. The normal approximation to $p(\lambda \,|\, y)$ is therefore

$$\lambda \,|\, y \overset{\text{approx.}}{\sim} \text{Normal}\big(\lambda \,|\, 1.67, 63.78^{-1}\big)$$

with $\Pr(\lambda > 2) = 0.0042$. A possibly sensible reparameterization is $\theta = \log\lambda$. It is know that $\mathbb{E}[\theta] = \log 2$ and $\Pr(\theta > \log 3) = 0.01$ so that

$$\Pr(\theta > \log 3) = \Pr\left(\frac{\theta - \log 2}{\psi} > \frac{\log 3/2}{\psi}\right) = \Pr\left(Z > \frac{\log 3/2}{\psi}\right) = 0.01 \Rightarrow \frac{1}{\psi} = \frac{\Phi^{-1}(0.99)}{\log 3/2} = 5.74\,.$$

The parameters of the Normal prior on θ are thus $\omega = 0.69$ and $\psi^2 = 0.03$. The posterior density is

$$p(\lambda \,|\, y) \propto \left[\prod_{i=1}^{n} e^{\theta y_i} \exp\{-e^{\theta}\}\right] \exp\Big\{-16.67(\theta - \log 2)^2\Big\} = \exp\big\{-16.67\theta^2 + (n\bar{y} + 23.11)\theta - ne^{\theta}\big\}$$

Using the above data and R, the mode of the posterior density is $\hat{\theta}_{\text{MAP}} = 0.53$ and observed Fisher information at the mode is $\mathcal{J}(\theta)|_{\theta=\hat{\theta}_{\text{MAP}}} = 203.71$. The normal approximation to $p(\lambda \,|\, y)$ after transforming back to the original parameter space is therefore

$$\lambda \,|\, y \overset{\text{approx.}}{\sim} \text{Normal}\big(\log\lambda \,|\, 0.53, 203.71^{-1}\big)\lambda^{-1}$$

with $\Pr(\lambda > 2 \,|\, y) = 0.0055$.

Exercise 3 (chapter 6.2): Let $\{Y_i\}_{i=1}^{n}$ be independent and identically distributed random variables that follow an Exponential distribution with rate parameter $\lambda > 0$. The density of the Exponential distribution is

$$f_Y(y) = \lambda \exp\{-\lambda y\} \qquad y > 0\,.$$

Assume that the prior on λ can be described by the following density

$$p(\lambda) \propto \exp\Big\{-20(\lambda - 0.25)^2\Big\} \qquad \lambda > 0\,.$$

1. Find the posterior $p(\lambda \,|\, y)$ and an expression for the normalizing constant.
2. Derive a normal approximation to the posterior at its mode using $n = 10$ and $\bar{y} = 0.5$. Plot the normal approximation together with the true posterior density.

Solution: The posterior density is

$$p(\lambda \,|\, y) \propto \left[\prod_{i=1}^{n} \lambda \exp\{-\lambda y_i\}\right] \exp\Big\{-20(\lambda - 0.25)^2\Big\} = \lambda^n \exp\big\{-20\lambda^2 + (10 - n\bar{y})\lambda\big\}\,.$$

The normalizing constant of the posterior density is

$$c^{-1} = \int_0^{\infty} \lambda^n \exp\big\{-20\lambda^2 + (10 - n\bar{y})\lambda\big\}\,\mathrm{d}\lambda\,.$$

The mode of the posterior density is

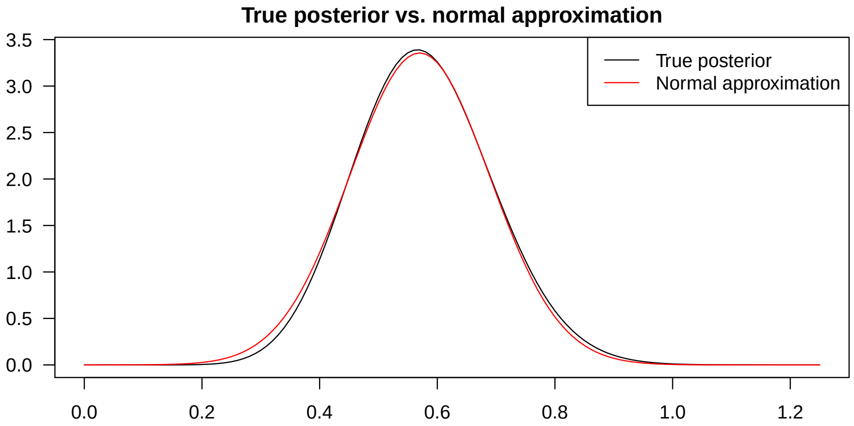
$$\begin{aligned} \frac{\mathrm{d}\log p(\lambda \,|\, y)}{\mathrm{d}\lambda} &= \frac{n}{\lambda} - 40\lambda + 10 - n\bar{y} \stackrel{!}{=} 0 \\ \Rightarrow \hat{\lambda}_{\mathrm{MAP}} &= \frac{-(n\bar{y} - 10) + \sqrt{(n\bar{y} - 10)^2 + 4(40)(n)}}{2(40)} \qquad \text{since } \lambda > 0 \end{aligned}$$

For the above data, the mode is $\hat{\lambda}_{\mathrm{MAP}} = 0.57$. The observed Fisher information is

$$\mathcal{J}(\lambda) = -\frac{\mathrm{d}^2 \log p(\lambda \,|\, y)}{\mathrm{d}\lambda^2} = \frac{n}{\lambda^2} + 40\,.$$

At the mode, the observed Fisher information is $\mathcal{J}(\lambda)|_{\lambda = \hat{\lambda}_{\mathrm{MAP}}} = 70.78$. The normal approximation to $p(\lambda \,|\, y)$ is therefore

$$\lambda \,|\, y \stackrel{\text{approx.}}{\sim} \mathrm{Normal}\big(\lambda \,|\, 0.57, 70.78^{-1}\big)$$



Exercise 4: Let $\{Y_i\}_{i=1}^n$ be independent and identically distributed random variables that follow an Normal distribution with location μ and precision parameter $\tau > 0$. The density of the Normal distribution with precision parameter τ is

$$f_Y(y) = \sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{\tau}{2}(y-\mu)^2\right\}.$$

Assume that $\mu \,|\, \tau \sim \text{Normal}\big(0, \tau^{-1}\big)$ and $\tau \sim \text{Gamma}(1, 1)$.

1. Derive the variational densities $q^\star(\mu \,|\, y) = \exp\{\mathbb{E}_\tau[\ln p(\mu, \tau, y)] - \ln c_\mu\}$ and $q^\star(\tau \,|\, y)$ under the mean–field assumption.
2. Implement a variational algorithm that refines the parameters of the variational distribution until convergence occurs.
3. Compare the variational algorithm to Gibbs sampling with respect to bias and speed using the following simulated data: `set.seed( 50 ) ; y <- rnorm( 100 )`

Solution: The variational density of μ is

$$\begin{aligned}
\ln q^\star(\mu \,|\, y) &= \mathbb{E}_\tau[\ln p(\mu, \tau, y)] + \text{const.} \\
&= \mathbb{E}_\tau\left[\sum_{i=1}^n \ln p(y_i \,|\, \mu, \tau) + \ln p(\mu \,|\, \tau)\right] + \text{const.} \\
&= -\frac{1}{2}\left\{\mathbb{E}[\tau]\sum_{i=1}^n (y_i - \mu)^2 + \mathbb{E}[\tau]\mu^2\right\} + \text{const.} \\
&= -\frac{1}{2}\left\{\mu^2(n\mathbb{E}[\tau] + \mathbb{E}[\tau]) - 2\mu\mathbb{E}[\tau]\sum_{i=1}^n y_i\right\} + \text{const.}
\end{aligned}$$

Exponentiating $\ln q^\star(\mu \,|\, y)$ indicates that the optimal variational density of μ is a normal densities, that is,

$$q^\star(\mu \,|\, y) = \text{Normal}\big(\mu \,|\, \omega, \psi^{-1}\big)$$

with precision

$$\psi = \mathbb{E}[\tau](n+1) = \frac{\alpha(n+1)}{\beta}$$

and location parameter

$$\omega = \frac{n\bar{y}}{n+1}\,.$$

The variational density of τ is

$$\begin{aligned}
\ln q^{\star}(\tau \mid y) &= \mathbb{E}_{\mu}[\ln p(\mu, \tau, y)] + \text{const.} \\
&= \mathbb{E}_{\mu}\left[\sum_{i=1}^{n} \ln p(y_i \mid \mu, \tau) + \ln p(\mu \mid \tau) + \ln p(\tau)\right] + \text{const.} \\
&= \mathbb{E}_{\mu}\left[\frac{n}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{n}(y_i - \mu)^2 + \frac{1}{2}\log\tau - \frac{\tau}{2}\mu^2 - \tau\right] + \text{const.} \\
&= \left(\frac{n+1}{2} + 1 - 1\right)\log\tau - \left\{1 + \frac{1}{2}\left(\sum_{i=1}^{n}\mathbb{E}\left[(y_i-\mu)^2\right] + \mathbb{E}\left[\mu^2\right]\right)\right\}\tau + \text{const.}
\end{aligned}$$

Exponentiating $\ln q^{\star}(\tau \mid y)$ indicates that the optimal variational density of τ is a Gamma densities, that is,

$$q^{\star}(\tau \mid y) = \text{Gamma}(\tau \mid \alpha, \beta)$$

with shape

$$\alpha = 1 + \frac{n+1}{2}$$

and rate parameter

$$\begin{aligned}
\beta &= 1 + \frac{1}{2}\left(\sum_{i=1}^{n}\mathbb{E}\left[(y_i-\mu)^2\right] + \mathbb{E}\left[\mu^2\right]\right) \\
&= 1 + \frac{1}{2}\left(\sum_{i=1}^{n} y_i^2 - 2\mathbb{E}[\mu]\sum_{i=1}^{n} y_i + \mathbb{E}\left[\mu^2\right][n+1]\right) \\
&= 1 + \frac{1}{2}\left(\sum_{i=1}^{n} y_i^2 - 2\omega\sum_{i=1}^{n} y_i + \left[\psi^{-1} + \omega^2\right][n+1]\right)
\end{aligned}$$

An implementation of a variational algorithm in R is

```
sumY <- sum( y ) ; sumYSquare <- sum( y ^ 2 )
omega <- sumY / ( n + 1 )
alpha <- 1 + 0.5 * ( n + 1 )
psi <- beta <- 1
tolerance <- 10 ^ -6
repeat {
   psiOld <- psi ; betaOld <- beta
   psi <- alpha * ( n + 1 ) / beta
   beta <- 1 + 0.5 * ( ( n + 1 ) * ( 1 / psi + omega ^ 2 ) -
      2 * omega * sumY + sumYSquare
   )
   if( all( abs( c( psiOld - psi, betaOld - beta ) ) < tolerance ) ) {
      break
```

```
    }
}
```

An implementation of Gibbs sampling in R is

```
sumY <- sum( y )
omega <-  sumY / ( n + 1 ) ; alpha <- 1 + 0.5 * ( n + 1 )
nSamples <- 10000 ;
x <- matrix( 0, nSamples, 2 ) ; x[ 1, ] <- 1
for( ii in seq( 2, nSamples ) ) {
   x[ ii, 2 ] <- rgamma(
      1,
      alpha,
      1 + 0.5 * sum( ( y - x[ ii - 1, 1 ] ) ^ 2 ) + 0.5 * ( x[ ii - 1, 1 ] ) ^ 2
   )
   x[ ii, 1 ] <- rnorm( 1, omega, sqrt( 1 / x[ ii, 2 ] / ( n + 1 ) ) )
}
```

Both methods give equivalent results, but the variational algorithm is significantly faster than Gibbs sampling for 10 000 draws.