ST4234: Bayesian Statistics

Tutorial 6 Solution, AY 19/20

Solutions

1.

(a) For this problem, you may assume that $a_0 + y - 1 > 0$ and that $b_0 + n - y - 1 > 0$. Otherwise, the posterior mode is at either 0 or 1 (on the boundary). In such cases, normal approximation and Laplace approximation are problematic.

(i) Since the posterior distribution of θ is $\mathrm{Beta}(a_0 + y, b_0 + n - y)$, the log posterior and its derivatives are

$$
\begin{aligned}
p(\theta|y) &= \frac{1}{\mathrm{B}(a_0 + y, b_0 + n - y)} \theta^{a_0+y-1}(1-\theta)^{b_0+n-y-1}, \\
\ell(\theta) &= \log p(\theta|y) = (a_0 + y - 1)\log\theta + (b_0 + n - y - 1)\log(1-\theta), \\
\ell'(\theta) &= \frac{\mathrm{d}\log p(\theta|y)}{\mathrm{d}\theta} = \frac{a_0 + y - 1}{\theta} - \frac{b_0 + n - y - 1}{1-\theta}, \\
\ell'(\theta) &= 0 \implies \widehat{\theta} = \frac{a_0 + y - 1}{a_0 + b_0 + n - 2}, \\
\ell''(\theta) &= \frac{\mathrm{d}^2\log p(\theta|y)}{\mathrm{d}\theta^2} = -\frac{a_0 + y - 1}{\theta^2} - \frac{b_0 + n - y - 1}{(1-\theta)^2}, \\
\implies -\ell''(\widehat{\theta}) &= \frac{(a_0 + b_0 + n - 2)^2}{a_0 + y - 1} + \frac{(a_0 + b_0 + n - 2)^2}{b_0 + n - y - 1} \\
&= \frac{(a_0 + b_0 + n - 2)^3}{(a_0 + y - 1)(b_0 + n - y - 1)}.
\end{aligned}
$$

Therefore, the normal approximation to the posterior of θ is

$$
\theta|y \overset{\text{approx}}{\sim} \mathrm{N}\left(\frac{a_0 + y - 1}{a_0 + b_0 + n - 2}, \frac{(a_0 + y - 1)(b_0 + n - y - 1)}{(a_0 + b_0 + n - 2)^3}\right).
$$

(ii) For the Laplace approximation, we take $g(\theta) = \theta$ as in Chapter 5 page 27-28. Then for the approximation method 1, the Laplace approximation is

$$
\mathrm{E}(\theta|y) \approx g(\widehat{\theta}) = \widehat{\theta} = \frac{a_0 + y - 1}{a_0 + b_0 + n - 2}.
$$

For the approximation method 2, we define the function $h^*(\theta)$ (as in Chapter 5 page 28) as

$$
\begin{aligned}
h^*(\theta) &= -n^{-1}\left[\log g(\theta) + \ell(\theta)\right] \\
&= -n^{-1}\left[\log\theta + (a_0 + y - 1)\log\theta + (b_0 + n - y - 1)\log(1-\theta)\right] \\
&= -n^{-1}\left[(a_0 + y)\log\theta + (b_0 + n - y - 1)\log(1-\theta)\right],
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}h^*(\theta)}{\mathrm{d}\theta} &= -n^{-1}\left[\frac{a_0+y}{\theta} - \frac{b_0+n-y-1}{1-\theta}\right], \\
\frac{\mathrm{d}h^*(\theta)}{\mathrm{d}\theta} = 0 &\implies \theta^* = \frac{a_0+y}{a_0+b_0+n-1}, \\
\frac{\mathrm{d}^2h^*(\theta)}{\mathrm{d}\theta^2} &= -n^{-1}\left[\frac{a_0+y}{\theta^2} + \frac{b_0+n-y-1}{(1-\theta)^2}\right], \\
\widehat{\sigma}^{*2} &= \left[\frac{\mathrm{d}^2h^*(\theta)}{\mathrm{d}\theta^2}\right]^{-1}\Bigg|_{\theta=\theta^*} \\
&= n\left[\frac{a_0+y}{\theta^{*2}} + \frac{b_0+n-y-1}{(1-\theta^*)^2}\right]^{-1} \\
&= n\left[\frac{(a_0+b_0+n-1)^2}{a_0+y} + \frac{(a_0+b_0+n-1)^2}{b_0+n-y-1}\right]^{-1} \\
&= n\frac{(a_0+y)(b_0+n-y-1)}{(a_0+b_0+n-1)^3}.
\end{aligned}
$$

Similarly, we can derive $\widehat{\sigma}$ from $h(\theta)$

$$
\begin{aligned}
h(\theta) &= -n^{-1}\ell(\theta) \\
&= -n^{-1}\left[(a_0+y-1)\log\theta + (b_0+n-y-1)\log(1-\theta)\right], \\
\frac{\mathrm{d}h(\theta)}{\mathrm{d}\theta} &= -n^{-1}\left[\frac{a_0+y-1}{\theta} - \frac{b_0+n-y-1}{1-\theta}\right], \\
\frac{\mathrm{d}h(\theta)}{\mathrm{d}\theta} = 0 &\implies \widehat{\theta} = \frac{a_0+y-1}{a_0+b_0+n-2}, \\
\frac{\mathrm{d}^2h(\theta)}{\mathrm{d}\theta^2} &= n^{-1}\left[\frac{a_0+y-1}{\theta^2} + \frac{b_0+n-y-1}{(1-\theta)^2}\right], \\
\widehat{\sigma}^{2} &= \left[\frac{\mathrm{d}^2h(\theta)}{\mathrm{d}\theta^2}\right]^{-1}\Bigg|_{\theta=\widehat{\theta}} \\
&= n\left[\frac{a_0+y-1}{\widehat{\theta}^2} + \frac{b_0+n-y-1}{(1-\widehat{\theta})^2}\right]^{-1} \\
&= n\left[\frac{(a_0+b_0+n-2)^2}{a_0+y-1} + \frac{(a_0+b_0+n-2)^2}{b_0+n-y-1}\right]^{-1} \\
&= n\frac{(a_0+y-1)(b_0+n-y-1)}{(a_0+b_0+n-2)^3}.
\end{aligned}
$$

Therefore (after some tedious calculations), the approximation method 2 estimates $\mathrm{E}(\theta|y)$ by

$$
\begin{aligned}
\mathrm{E}(\theta|y) &\approx \frac{\widehat{\sigma}^*\theta^* p(y|\theta^*)p(\theta^*)}{\widehat{\sigma}p(y|\widehat{\theta})p(\widehat{\theta})} \\
&= \frac{\sqrt{n\frac{(a_0+y)(b_0+n-y-1)}{(a_0+b_0+n-1)^3}}\cdot\theta^*\cdot(\theta^*)^{a_0+y-1}(1-\theta^*)^{b_0+n-y-1}}{\sqrt{n\frac{(a_0+y-1)(b_0+n-y-1)}{(a_0+b_0+n-2)^3}}\cdot(\widehat{\theta})^{a_0+y-1}(1-\widehat{\theta})^{b_0+n-y-1}} \\
&= \frac{(a_0+y)^{a_0+y+1/2}(a_0+b_0+n-2)^{a_0+b_0+n-1/2}}{(a_0+y-1)^{a_0+y-1/2}(a_0+b_0+n-1)^{a_0+b_0+n+1/2}}.
\end{aligned}
$$

(b) For this problem, you may assume that $a_0+t-1>0$ where $t=\sum_{i=1}^n y_i$. Otherwise, the posterior mode is at zero (on the boundary). In that case, normal approximation and Laplace approximation are problematic.

(i) Since the posterior distribution of θ is $\text{Gamma}(a_0+t, b_0+n)$, the log posterior and its derivatives are

$$
\begin{aligned}
p(\theta|\boldsymbol{y}) &\propto \theta^{a_0+t-1}e^{-(b_0+n)\theta},\\
\ell(\theta) &= \log p(\theta|\boldsymbol{y}) = (a_0+t-1)\log\theta-(b_0+n)\theta,\\
\ell'(\theta) &= \frac{\mathrm{d}\log p(\theta|\boldsymbol{y})}{\mathrm{d}\theta} = \frac{a_0+t-1}{\theta}-(b_0+n),\\
\ell'(\theta) &= 0 \implies \widehat{\theta} = \frac{a_0+t-1}{b_0+n},\\
\ell''(\theta) &= \frac{\mathrm{d}^2\log p(\theta|\boldsymbol{y})}{\mathrm{d}\theta^2} = -\frac{a_0+t-1}{\theta^2},\\
\implies -\ell''(\widehat{\theta}) &= \frac{(b_0+n)^2}{a_0+t-1}.
\end{aligned}
$$

Therefore, the normal approximation to the posterior of θ is

$$
\theta|\boldsymbol{y} \overset{\text{approx}}{\sim} \mathrm{N}\left(\frac{a_0+t-1}{b_0+n}, \frac{a_0+t-1}{(b_0+n)^2}\right).
$$

(ii) For the Laplace approximation, we take $g(\theta)=\theta$ as in Chapter 5 page 27-28. Then for the approximation method 1, the Laplace approximation is

$$
\mathrm{E}(\theta|\boldsymbol{y}) \approx g(\widehat{\theta}) = \widehat{\theta} = \frac{a_0+t-1}{b_0+n}.
$$

For the approximation method 2, we define the function $h^*(\theta)$ (as in Chapter 5 page 28) as

$$
\begin{aligned}
h^*(\theta) &= -n^{-1}\left[\log g(\theta)+\ell(\theta)\right]\\
&= -n^{-1}\left[\log\theta+(a_0+t-1)\log\theta-(b_0+n)\theta\right]\\
&= -n^{-1}\left[(a_0+t)\log\theta-(b_0+n)\theta\right],\\
\frac{\mathrm{d}h^*(\theta)}{\mathrm{d}\theta} &= -n^{-1}\left[\frac{a_0+t}{\theta}-(b_0+n)\right],\\
\frac{\mathrm{d}h^*(\theta)}{\mathrm{d}\theta} &= 0 \implies \theta^* = \frac{a_0+t}{b_0+n},\\
\frac{\mathrm{d}^2h^*(\theta)}{\mathrm{d}\theta^2} &= n^{-1}\frac{a_0+t}{\theta^2},\\
\widehat{\sigma}^{*2} &= \left[\frac{\mathrm{d}^2h^*(\theta)}{\mathrm{d}\theta^2}\right]^{-1}\Bigg|_{\theta=\theta^*}\\
&= \frac{n\theta^{*2}}{a_0+t}\\
&= \frac{n(a_0+t)}{(b_0+n)^2}.
\end{aligned}
$$

Similarly, we can derive $\widehat{\sigma}$ from $h(\theta)$

$$\begin{aligned}
h(\theta) &= -n^{-1}\ell(\theta) \\
&= -n^{-1}\left[(a_0+t-1)\log\theta - (b_0+n)\theta\right], \\
\frac{\mathrm{d}h(\theta)}{\mathrm{d}\theta} &= -n^{-1}\left[\frac{a_0+t-1}{\theta} - (b_0+n)\right], \\
\frac{\mathrm{d}h(\theta)}{\mathrm{d}\theta} = 0 &\implies \widehat{\theta} = \frac{a_0+t-1}{b_0+n}, \\
\frac{\mathrm{d}^2h(\theta)}{\mathrm{d}\theta^2} &= n^{-1}\frac{a_0+t-1}{\theta^2}, \\
\widehat{\sigma}^2 &= \left[\frac{\mathrm{d}^2h(\theta)}{\mathrm{d}\theta^2}\right]^{-1}\Big|_{\theta=\widehat{\theta}} \\
&= \frac{n\widehat{\theta}^2}{a_0+t-1} \\
&= \frac{n(a_0+t-1)}{(b_0+n)^2}.
\end{aligned}$$

Therefore (after some tedious calculations), the approximation method 2 estimates $\mathrm{E}(\theta|\boldsymbol{y})$ by

$$\begin{aligned}
\mathrm{E}(\theta|\boldsymbol{y}) &\approx \frac{\widehat{\sigma}^*\theta^* p(\boldsymbol{y}|\theta^*)p(\theta^*)}{\widehat{\sigma}p(\boldsymbol{y}|\widehat{\theta})p(\widehat{\theta})} \\
&= \frac{\sqrt{\frac{n(a_0+t)}{(b_0+n)^2}}\cdot\theta^*\cdot(\theta^*)^{a_0+t-1}e^{-(b_0+n)\theta^*}}{\sqrt{\frac{n(a_0+t-1)}{(b_0+n)^2}}\cdot\widehat{\theta}^{a_0+t-1}e^{-(b_0+n)\widehat{\theta}}} \\
&= \frac{(a_0+t)^{a_0+t+1/2}}{(a_0+t-1)^{a_0+t-1/2}(b_0+n)e},
\end{aligned}$$

where $e = 2.71828...$

2. $\theta = \log \frac{p}{1-p} \Rightarrow p = \frac{e^\theta}{1+e^\theta}$ and $1-p = \frac{1}{1+e^\theta}$. The prior for θ is $\theta \sim N(\mu, \sigma^2)$. The posterior density of θ is

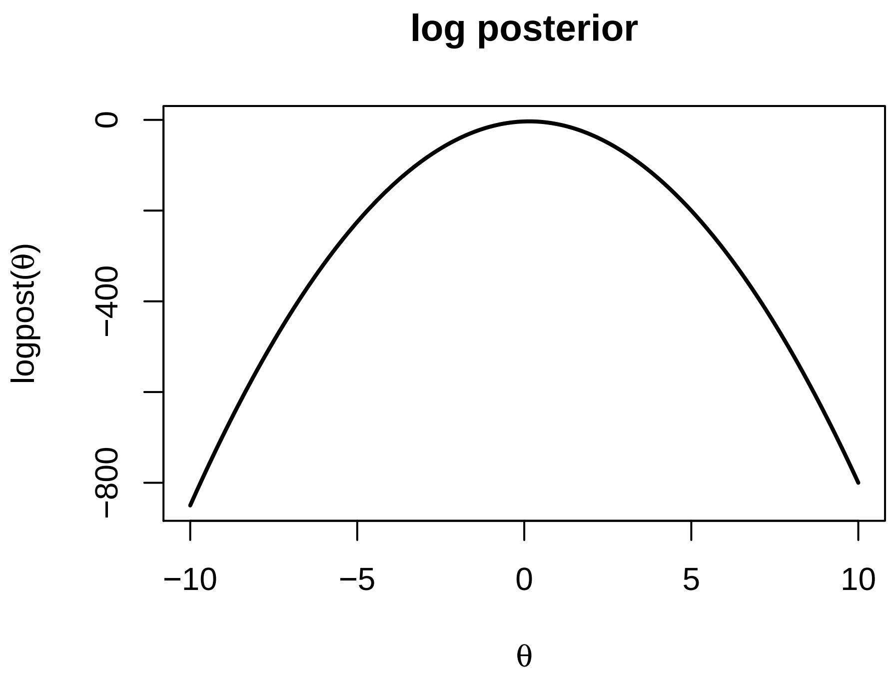
$$\begin{aligned}
p(\theta|y) &\propto p(y|\theta)p(\theta) \\
&\propto \binom{n}{y} p^y (1-p)^{n-y} \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\} \\
&\propto \left(\frac{e^\theta}{1+e^\theta}\right)^y \left(\frac{1}{1+e^\theta}\right)^{n-y} \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\} \\
&\propto \frac{\exp(y\theta)}{[1+\exp(\theta)]^n} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\}.
\end{aligned}$$

We have

$$\log p(\theta|y) = \underbrace{\theta y - n\log(1+e^\theta) - \frac{(\theta-\mu)^2}{2\sigma^2}}_{\ell(\theta)} + C,$$

where C is an additive constant not depending on θ. A normal approximation to $p(\theta|y)$ is given by $N(\hat{\theta}, -[\ell''(\hat{\theta})]^{-1})$ where $\hat{\theta}$ is the posterior mode. The figure below shows a plot of $\ell(\theta)$ and we can see that the mode is close to 0.

```r
mu <- 0
sigma2 <- 0.25^2
n <- 5
y <- 5  # 5 heads

# log posterior function
logpost <- function(theta,y,n,mu,sigma2){
        theta*y - n*log(1+exp(theta)) - 0.5*(theta-mu)^2/sigma2
}

# plot the log posterior
theta.grid <- seq(from=-10,to=10,by=0.1)
plot(theta.grid, logpost(theta.grid,y,n,mu,sigma2),type="l",lwd=2,
     ylab=expression(paste("logpost(",theta,")")),xlab=expression(theta),
         main="log posterior")
```



Note that since the log posterior is a one-dimensional function, we can use `optimize()`, but `optimize()` does not return the Hessian and we need to calculate it analytically. This is not difficult as the derivatives of $\ell(\theta)$ are

$$\begin{aligned}
\ell'(\theta) &= y - n\frac{\mathrm{e}^\theta}{1+\mathrm{e}^\theta} - \frac{(\theta-\mu)}{\sigma^2}, \\
\ell''(\theta) &= -n\frac{\mathrm{e}^\theta(1+\mathrm{e}^\theta) - \mathrm{e}^\theta \mathrm{e}^\theta}{(1+\mathrm{e}^\theta)^2} - \frac{1}{\sigma^2} \\
&= -n\frac{\mathrm{e}^\theta}{(1+\mathrm{e}^\theta)^2} - \frac{1}{\sigma^2}.
\end{aligned}$$

```
(out <- optimize(logpost, interval=c(-10,10), maximum=TRUE,
                 y=y,n=n,mu=mu,sigma2=sigma2))

## $maximum
## [1] 0.1449459
##
## $objective
## [1] -3.284565

(post.mode <- out$maximum)

## [1] 0.1449459

(post.var <- 1/(n*exp(post.mode)/(1+exp(post.mode))^2 + 1/sigma2))

## [1] 0.05799301
```

An alternative way is to use `optim()` to find the posterior mode, but with the method option `method="Brent"`. This method requires the specification of `lower` and `upper` for the interval on which θ is optimized over.

```
(out <- optim(par=0, fn=logpost, hessian=TRUE, control=list(fnscale=-1),
        method="Brent", lower=-10, upper=10,y=y,n=n,mu=mu,sigma2=sigma2))

## $par
## [1] 0.1449459
##
## $value
## [1] -3.284565
##
## $counts
## function gradient
##       NA       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
##
## $hessian
##            [,1]
## [1,] -17.24346

(post.mode <- out$par)

## [1] 0.1449459

(post.var <- -1/out$hessian)

##             [,1]
## [1,] 0.05799301
```

The results from two optimization methods agree with each other. We find that the posterior mode $\hat{\theta} = 0.145$ and $-[h''(\hat{\theta})]^{-1} = 0.058$. The normal approximation to $p(\theta|y)$ is N(0.145, 0.058). From this normal approximation the $P(\theta > 0|y) \approx 0.726$.

```r
1 - pnorm(0, mean=post.mode, sd=sqrt(post.var))
```

```
## [1] 0.7263768
```

3. (a) When we change the parameters from (β, μ) to (θ_1, θ_2), we use the change-of-variable formula and multiply the original density by the determinant of Jacobian matrix. Using the transformations $\theta_1 = \log \beta$, $\theta_2 = \log(t_1 - \mu)$, we have

$$\beta = \mathrm{e}^{\theta_1}, \ \frac{\mathrm{d}\beta}{\mathrm{d}\theta_1} = \mathrm{e}_1^{\theta}, \quad \mu = t_1 - \mathrm{e}^{\theta_2}, \ \frac{\mathrm{d}\mu}{\mathrm{d}\theta_2} = -\mathrm{e}_2^{\theta}$$
$$\Rightarrow J = \begin{bmatrix} \mathrm{e}_1^{\theta} & 0 \\ 0 & -\mathrm{e}_2^{\theta} \end{bmatrix}, \quad |\det(J)| = \mathrm{e}^{\theta_1+\theta_2}.$$

Hence the posterior is

$$p(\theta_1, \theta_2|\text{data}) \propto \mathrm{e}^{-s\theta_1} \exp\left\{-\frac{t - n(t_1 - \mathrm{e}^{\theta_2})}{\mathrm{e}^{\theta_1}}\right\} \cdot \mathrm{e}^{\theta_1+\theta_2}$$
$$\propto \exp\left\{(1-s)\theta_1 + \theta_2 - \mathrm{e}^{-\theta_1}(t - nt_1 + n\mathrm{e}^{\theta_2})\right\}.$$

We have $n = 15$, $s = 8$, $t = 15962989$, and $t_1 = 237217$ from the question.

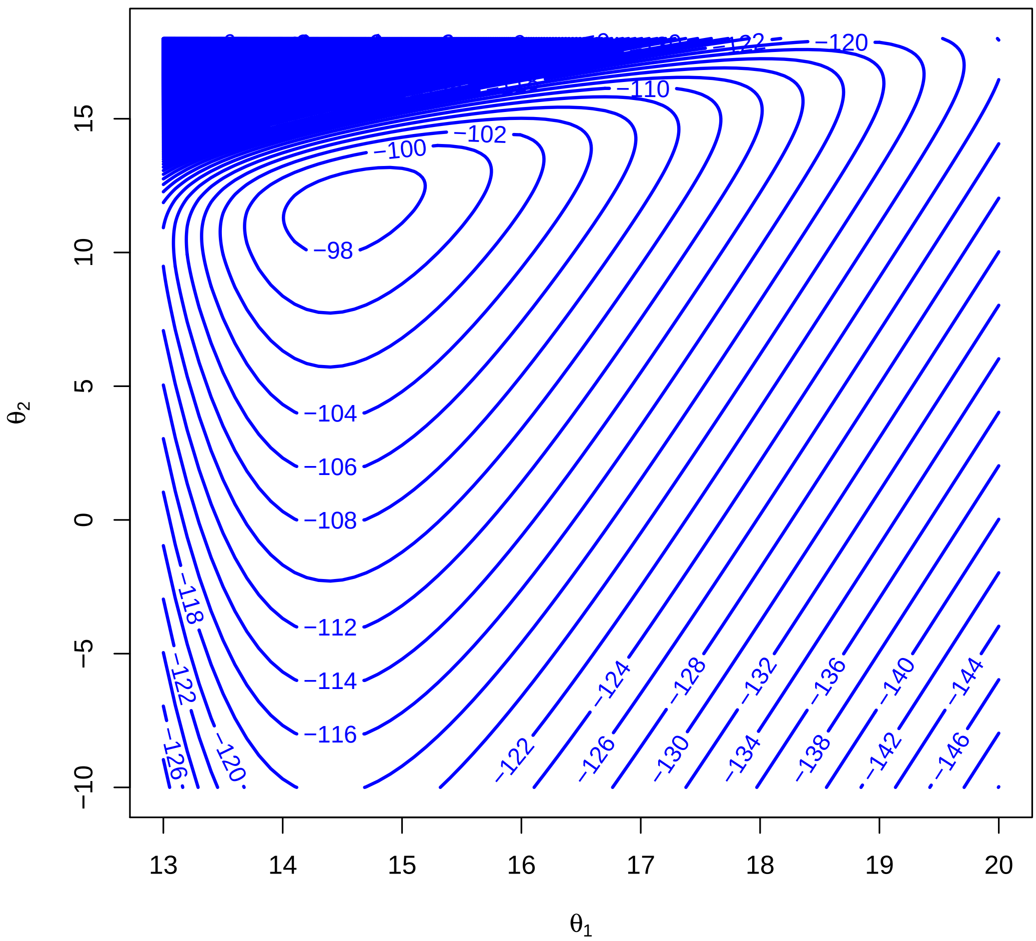
(b) The R function to compute the log posterior is as follows.

```r
s <- 8
n <- 15
t <- 15962989
t1 <- 237217
logpost <- function(theta, s, n, t, t1){
        theta1 <- theta[1]
        theta2 <- theta[2]
        return((1-s)*theta1 + theta2 -
                 (t-n*t1)*exp(-theta1) - n*exp(theta2-theta1))
}
```

(c) We can first try to find the global maximum point of (θ_1, θ_2) by using `optim`. Note that it is not always guaranteed that the solution is a global maximum. Sometimes, it may end up with a local maximum. So we can use trial and error: we first plot for a large area of the parameters in a contour plot, and then zoom in to a proper region.

```r
# some trial-and-error / exploration
(out <- optim(par=c(0,0), fn=logpost, hessian=TRUE,
              control=list(fnscale=-1), s=s, n=n, t=t, t1=t1))

## $par
## [1]  14.31215 -16.79486
##
## $value
```

```r
## [1] -124.5291
##
## $counts
## function gradient
##       39       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
##
## $hessian
##           [,1] [,2]
## [1,] -7.549201    0
## [2,]  0.000000    0

# this ends up in a local extremum value, not the global maximum

# contour plot: several attempts
# theta1.grid <- seq(from=-100, to=100, by=1)
# theta2.grid <- seq(from=-100, to=100, by=1)
# range too large, change the range

theta1.grid <- seq(from=13, to=20, by=.1)
theta2.grid <- seq(from=-10, to=18, by=.1)
theta.grid <- expand.grid(theta1.grid, theta2.grid)
grid.logpost <- apply(theta.grid, 1, logpost, s=s, n=n, t=t, t1=t1)
contour(theta1.grid, theta2.grid,
        matrix(grid.logpost,nrow=length(theta1.grid),
               ncol=length(theta2.grid)),
        col="blue", nlevels=800, lwd=2, labcex=0.9,
        xlab=expression(theta[1]), ylab=expression(theta[2]))
```



From the contour plot, we see that the mode is located at approximately (14.5,11) and we use this as a starting point for the optimization.

```
(out <- optim(par=c(14.5,11), fn=logpost, hessian=TRUE,
              control=list(fnscale=-1), s=s, n=n, t=t, t1=t1))

## $par
## [1] 14.54155 11.83311
##
## $value
## [1] -96.95903
##
## $counts
## function gradient
##       45       NA
##
## $convergence
```

```
## [1] 0
##
## $message
## NULL
##
## $hessian
##            [,1]       [,2]
## [1,] -7.0013119  0.9996118
## [2,]  0.9996118 -0.9996118
```

```
(post.mode <- out$par)
```

```
## [1] 14.54155 11.83311
```

```
(post.var <- -solve(out$hessian))
```

```
##           [,1]      [,2]
## [1,] 0.1666195 0.1666195
## [2,] 0.1666195 1.1670078
```

Hence a normal approximation of the posterior density is

$$(\theta_1, \theta_2) \mid \text{data} \overset{\text{approx}}{\sim} \mathrm{N}\left(\begin{bmatrix}14.54\\11.83\end{bmatrix}, \begin{bmatrix}0.1666 & 0.1666\\0.1666 & 1.1670\end{bmatrix}\right).$$