Chapter 5: Normal Approximation and Laplace Approximation

ST4234: Bayesian Statistics

Semester 2, AY 2019/2020

Department of Statistics and Applied Probability

National University of Singapore

LI Cheng

stalic@nus.edu.sg

Introduction

- This chapter corresponds to Sections 5.1-5.7 in Jim Albert's book, and part of Chapters 2 and 3 in Martin A. Tanner's book.

- When the sampling density has a familiar functional form, such as a member of an exponential family, and a conjugate prior is chosen for the parameter, then the posterior distribution is often expressible in terms of standard probability distributions.

- However, for many multiparameter models, the joint posterior distribution is non-standard and difficult to sample from directly.

- In such cases, more advanced computational methods have to be employed. In this chapter, we discuss several methods based on normal approximation. We will discuss more methods in the next chapter.

Bayesian Computation: Computing Integrals

- A general problem in Bayesian computation is about computing posterior expectations.

- Suppose that the model is $p(\boldsymbol{y}|\theta)$, the prior is $p(\theta)$, and the posterior is $p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)p(\theta)$. Let $g(\theta)$ be a known function of θ. It is typically of interest to compute the posterior mean of $g(\theta)$:

$$\mathrm{E}[g(\theta)|\boldsymbol{y}] = \frac{\int g(\theta)p(\boldsymbol{y}|\theta)p(\theta)\mathrm{d}\theta}{\int p(\boldsymbol{y}|\theta)p(\theta)\mathrm{d}\theta}.$$

- Calculating this posterior expectation clearly involves evaluating two integrals over the parameter space.

- If $g(\theta) = \theta$, this is the posterior mean of θ. If $g(\theta) = \theta^2$, this is related to the posterior variance of θ. If $g(\theta) = \mathbb{1}(\theta \in A)$ for some set A, then this is the posterior probability of A, i.e. $\mathrm{P}(A|\boldsymbol{y})$.

Bayesian Computation: Computing Integrals

- Evaluate the integrals in such posterior expectations can be difficult, especially when θ has many dimensions.
- In this chapter, we will replace these integrals with approximations, such that the posterior expectations can be computed approximated. The methods we will introduce are normal approximation and Laplace approximation.
- In the next chapter, we will estimate these integrals using Monte Carlo methods, including rejection sampling and importance sampling.
- We will use two running examples to illustrate these computational methods. During this process, we will also introduce some Bayesian modeling techniques.

Outline

5.1 Example: Genetic Linkage Model

Suppose that 197 animals ($\boldsymbol{y}$) are distributed into four categories as follows (Rao 1997):

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

with cell probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4},\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta), \frac{\theta}{4}\right),$$

where $\theta \in [0, 1]$ is the parameter.

- We can treat $\boldsymbol{y}$ as a summary statistic of 197 numbers from the multinomial distribution with the cell probabilities given as above.
- The likelihood function is

$$\begin{aligned} p(\boldsymbol{y}|\theta) &\propto \left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1-\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4} \\ &\propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}. \end{aligned}$$

5.1 Example: Genetic Linkage Model

- Suppose that we impose a Uniform$(0, 1)$ prior on $\theta \in [0, 1]$, so $p(\theta) = 1$.
- The posterior density of θ is

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta)p(\theta) \\ &\propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}. \end{aligned}$$

- Note that this distribution is not a familiar distribution.
- Since $p(\theta|\boldsymbol{y})$ is a 1-dimensional distribution, we can numerically find the normalizing constant for $p(\theta|\boldsymbol{y})$. In R, we can use the function `integrate()` to find the normalizing constant.

5.1 Example: Genetic Linkage Model

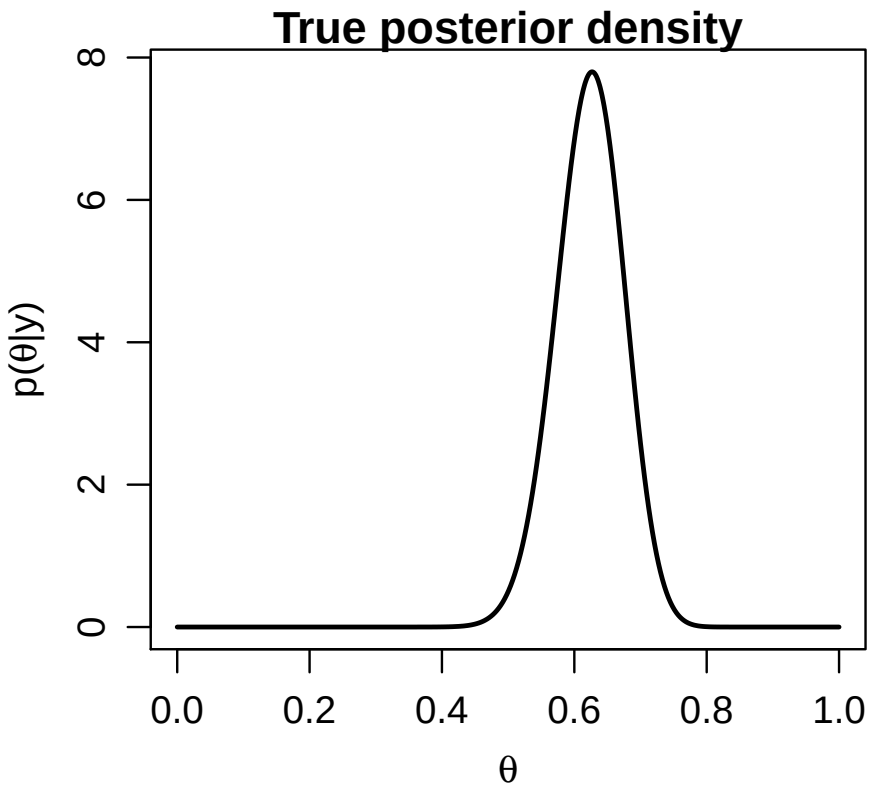
- The function `integrate(f, lower, upper)` has three main arguments.
- `f` is the name of the function to integrate. `lower` and `upper` are the lower and upper limits of the integration.
- We can first write a function called `post0` to represent the unnormalized posterior density. Then we find the normalizing constant by integrating `post0` from 0 to 1.

```
Y = c(125, 18, 20, 34)
post0 <- function(theta, Y) {
    (2+theta)^(Y[1])*(1-theta)^(Y[2]+Y[3])*theta^(Y[4])
}
> (Const <- integrate(post0, 0, 1, Y=Y)$value)
[1] 2.357695e+28
```

5.1 Example: Genetic Linkage Model

- With this normalizing constant, we can plot the true posterior density (which integrates to 1).

```
theta.grid <- seq(from=0, to=1, by=0.001)
plot(theta.grid, post0(theta.grid,Y=Y)/Const, type="l", lwd=2,
 ylab=expression(paste("p(",theta,"|y)")),
     xlab=expression(theta), main="True posterior density")
```



5.1 Example: Genetic Linkage Model

Now we consider the following questions:

(i) How to find the normalizing constant without using numerical integration (such as the `integrate()` function)?

(ii) How to approximate the posterior density $p(\theta|\boldsymbol{y})$?

(iii) How to estimate the posterior mean and posterior standard deviation of $p(\theta|\boldsymbol{y})$, based on a reasonable approximation?

We introduce normal approximation and Laplace approximation to address these questions.

Outline

5.2 Normal approximation of the posterior

- One method of summarizing a multivariate posterior distribution is based on the behavior of the density about its **mode**.
- Let θ be a vector-valued parameter with prior density $p(\theta)$. Suppose we observe data $\boldsymbol{y}$ with likelihood $p(\boldsymbol{y}|\theta)$.
- Let

$$\ell(\theta) = \log p(\theta|\boldsymbol{y}) + C \tag{1}$$

  where C is an additive constant not depending on θ.
- Then $\ell(\theta)$ and $p(\theta|\boldsymbol{y})$ share the same modes. The first and second derivatives of the log posterior, $\log p(\theta|\boldsymbol{y})$, are equal to $\ell'(\theta)$ and $\ell''(\theta)$ respectively.

5.2 Normal approximation of the posterior

- Let $\hat{\theta}$ denote the posterior mode of θ.
- If we expand $\ell(\theta)$ in a second-order Taylor series about $\hat{\theta}$, we have

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^\top \ell''(\hat{\theta})(\theta - \hat{\theta}) \tag{2}$$

  where $\ell''(\hat{\theta})$ is the Hessian of $\ell(\theta)$ evaluated at the mode. Here, $\top$ is the matrix transpose.

- If θ is d-dimensional, then $\ell'(\theta)$ is a $d \times 1$ vector of functions, and $\ell''(\theta)$ is a $d \times d$ symmetric matrix of functions.
- Note that in Equation (2), there is no first order term, because $\ell'(\hat{\theta}) = 0$ when $\hat{\theta}$ is a mode of $\ell(\theta)$ (why?).

5.2 Normal approximation of the posterior

- From (1) and (2), we can derive that

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \exp\{\ell(\theta) - C\} \\ &\approx \exp\left\{\ell(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^\top \ell''(\hat{\theta})(\theta-\hat{\theta}) - C\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\theta-\hat{\theta})^\top[-\ell''(\hat{\theta})](\theta-\hat{\theta})\right\}, \end{aligned}$$

  which is proportional to the density of a multivariate normal distribution with mean $\hat{\theta}$ and covariance matrix $-[\ell''(\hat{\theta})]^{-1}$.

- A random vector $\mathbf{X} = (X_1, \ldots, X_p)^\top$ is said to follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, denoted $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if the pdf of $\mathbf{X}$ is

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}\{\det(\boldsymbol{\Sigma})\}^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

5.2 Normal approximation of the posterior

- Thus we obtain a normal approximation to the posterior density:

$$\theta|\boldsymbol{y} \sim \mathrm{N}(\hat{\theta}, -[\ell''(\hat{\theta})]^{-1}). \tag{3}$$

- The normal approximation is often reasonable due to the asymptotic normality of posterior distributions:

Theorem (Bayesian Central Limit Theorem, or Bernstein-von Mises Theorem)

Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p(x|\theta)$ *and* $p(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$. *Suppose the prior* $p(\theta)$ *and the likelihood* $p(\boldsymbol{x}|\theta)$ *are positive and twice differentiable near* $\hat{\theta}$, *the posterior mode of* θ. *Then under suitable regularity conditions,* **as** $n \to \infty$, *the posterior distribution* $p(\theta|\boldsymbol{x})$ *can be approximated by a normal distribution with mean equal to the posterior mode, and covariance matrix equal to the negative of the inverse Hessian of the log posterior evaluated at the mode.*

5.2 Normal approximation of the posterior

- To apply this approximation, we need to find the posterior mode. That is, we need to find θ such that $\ell'(\theta) = 0$.
- One optimization algorithm for finding this mode is Newton's method.
  - Suppose θ_0 is an initial value.
  - If θ_{t-1} is the estimate of the mode at the $t-1$ iteration, then the next iteration is given by

$$\theta_t = \theta_{t-1} - [\ell''(\theta_{t-1})]^{-1}\ell'(\theta_{t-1}),$$

    where $\ell'(\theta_{t-1})$ and $\ell''(\theta_{t-1})$ are the gradient and Hessian of $\ell(\theta)$ evaluated at the current value θ_{t-1}.
  - One continues these iterations until convergence.

5.2 Normal approximation of the posterior

Genetic Linkage Model

- Let us find a normal approximation for the genetic linkage model. From the plot, we have seen that the posterior $p(\theta|\boldsymbol{y})$ is unimodal.
- The log posterior and its derivatives are given by

$$\ell(\theta) = \log p(\theta|\boldsymbol{y}) = y_1 \log(2+\theta) + (y_2 + y_3)\log(1-\theta) + y_4 \log\theta,$$
$$\ell'(\theta) = \frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta},$$
$$\ell''(\theta) = -\frac{y_1}{(2+\theta)^2} - \frac{y_2+y_3}{(1-\theta)^2} - \frac{y_4}{\theta^2}.$$

- Therefore, once we find the mode $\hat{\theta}$, the variance of normal approximation will be

$$-[\ell''(\hat{\theta})]^{-1} = 1\Big/\left[\frac{y_1}{(2+\hat{\theta})^2} + \frac{y_2+y_3}{(1-\hat{\theta})^2} + \frac{y_4}{\hat{\theta}^2}\right].$$

5.2 Normal approximation of the posterior

Genetic Linkage Model

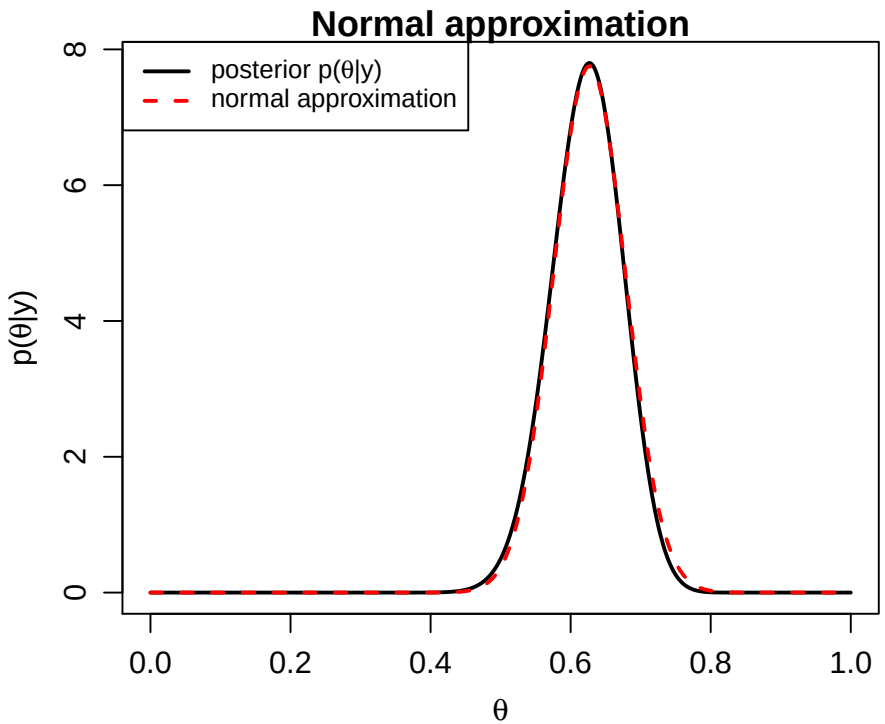
- We can use the R function `optimize()` to directly find the mode of $p(\theta|\boldsymbol{y})$.
- Note that to search for the mode, we only need to optimize the unnormalized posterior density `post0()`.
- The Hessian can be computed using the analytical formula in the last page.

```
> (out <- optimize(post0, interval=c(0,1), Y=Y, maximum=TRUE))
$maximum
[1] 0.6268101
$objective
[1] 1.838839e+29
> theta.hat <- out$maximum
> (var.theta <- 1/(Y[1]/(2+theta.hat)^2+
          (Y[2]+Y[3])/(1-theta.hat)^2+Y[4]/theta.hat^2))
[1] 0.002648982
```

5.2 Normal approximation of the posterior

Genetic Linkage Model

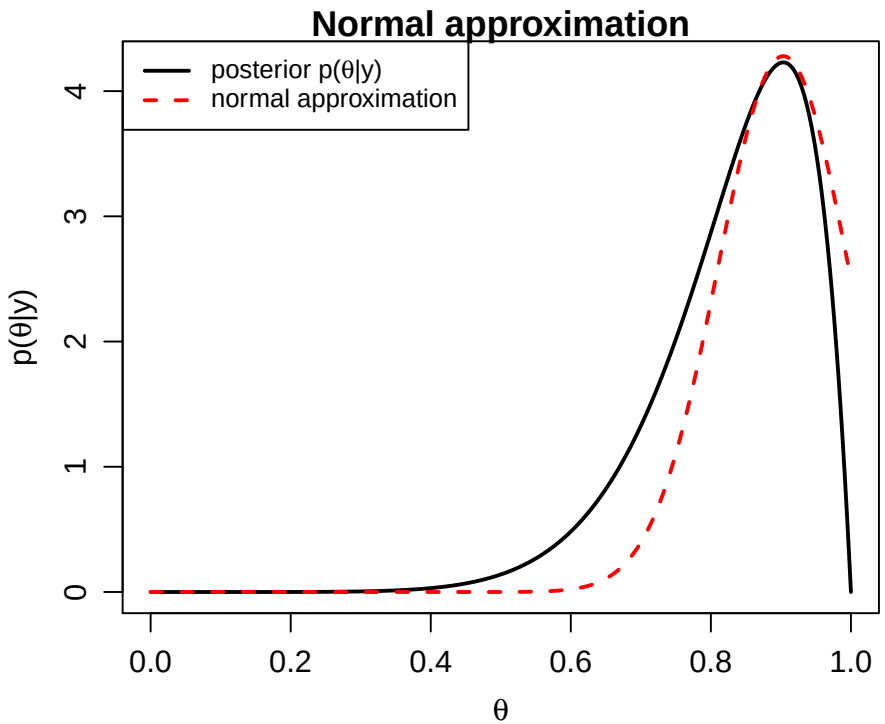
- Therefore, we can approximate the true posterior $p(\theta|\boldsymbol{y})$ with a normal distribution $\mathrm{N}(0.6268, 0.002649)$.
- We can plot the overlayed normal density with the true posterior density.



5.2 Normal approximation of the posterior

Genetic Linkage Model

- Now we modify the original data a little bit: we let $Y = (14, 0, 1, 5)$. We repeat the whole procedure again.
- The normal approximation gives $\mathrm{N}(0.9034, 0.008691)$.
- The density plot now becomes



5.2 Normal approximation of the posterior

Genetic Linkage Model

- We can see that the normal approximation becomes less accurate for $Y = (14, 0, 1, 5)$, because the posterior is skewed while the normal density is always symmetric.
- This prompts us to find some better alternatives to a normal density.
- Since we already have the unnormalized posterior density, a natural idea is to only estimate the normalizing constant, and then divide the unnormalized posterior density by this estimated normalizing constant.
- This can address the shape problem of a normal density. The shape will be only determined by the already known unnormalized posterior density.

Outline

5.3 Laplace approximation

- Recall that in the posterior $p(\theta|\boldsymbol{y})$, the normalizing constant is defined to be $\int p(\boldsymbol{y}|\theta)p(\theta)\mathrm{d}\theta$.
- We present a general method for estimating the integral

$$\mathcal{I} = \int g(\theta)p(\boldsymbol{y}|\theta)p(\theta)\mathrm{d}\theta.$$

- If we choose $g(\theta) = 1$, then $\mathcal{I}$ is the normalizing constant in $p(\theta|\boldsymbol{y})$.
- If we choose $g(\theta) = \theta$, then $\mathcal{I}$ becomes the numerator in $\mathrm{E}(\theta|\boldsymbol{y})$, since

$$\mathrm{E}[g(\theta)|\boldsymbol{y}] = \frac{\int g(\theta)p(\boldsymbol{y}|\theta)p(\theta)\mathrm{d}\theta}{\int p(\boldsymbol{y}|\theta)p(\theta)\mathrm{d}\theta}.$$

5.3 Laplace approximation

Univariate θ

- Let θ be a scalar. The Laplace method (Tierney and Kadane 1986) works for integrals of the following format:

$$\mathcal{I} = \int f(\theta) \cdot \exp\{-nh(\theta)\} \mathrm{d}\theta. \tag{4}$$

- The Laplace method approximates $\mathcal{I}$ with

$$\widehat{\mathcal{I}} = f(\hat{\theta}) \sqrt{\frac{2\pi}{n}} \hat{\sigma} \exp\left\{-nh(\hat{\theta})\right\}.$$

- $\hat{\theta}$ is the minimizer of $h(\theta)$, and $\hat{\sigma} = \left[h''(\theta)\Big|_{\theta=\hat{\theta}}\right]^{-1/2}$.

5.3 Laplace approximation

Univariate θ

- The derivation of this approximation is similar to normal approximation. We can use a Taylor-series expansion around the mode $\theta = \hat{\theta}$:

$$\begin{aligned}
\mathcal{I} &\approx \int f(\hat{\theta}) \exp\left\{-n\left[h(\hat{\theta}) + (\theta - \hat{\theta})h'(\hat{\theta}) + \frac{h''(\hat{\theta})}{2}(\theta - \hat{\theta})^2\right]\right\} \mathrm{d}\theta \\
&= f(\hat{\theta}) \exp\left\{-nh(\hat{\theta})\right\} \int \exp\left\{-\frac{n(\theta - \hat{\theta})^2}{2\hat{\sigma}^2}\right\} \mathrm{d}\theta \\
&= f(\hat{\theta}) \exp\left\{-nh(\hat{\theta})\right\} \cdot \sqrt{\frac{2\pi}{n}}\hat{\sigma} = \widehat{\mathcal{I}},
\end{aligned}$$

  where $\hat{\sigma} = \left[h''(\hat{\theta})\right]^{-1/2}$.

- Again we use the fact $h'(\hat{\theta}) = 0$.

5.3 Laplace approximation

Univariate θ

- It can be shown that in general, $\widehat{\mathcal{I}} = \mathcal{I}\left[1 + \mathcal{O}\left(\frac{1}{n}\right)\right]$.
- To compute the posterior related integrals, we can set

$$h(\theta) = -\frac{1}{n}\left[\log p(\boldsymbol{y}|\theta) + \log p(\theta)\right].$$

- Suppose that the unnormalized posterior $\widetilde{p}(\theta|\boldsymbol{y}) = p(\boldsymbol{y}|\theta)p(\theta)$ is known. Then the Laplace method approximates the posterior density by

$$\widehat{p}(\theta|\boldsymbol{y}) = \frac{\widetilde{p}(\theta|\boldsymbol{y})}{\widehat{\mathcal{I}}} = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{\sqrt{2\pi}\hat{s}p(\boldsymbol{y}|\hat{\theta})p(\hat{\theta})},$$

$$\text{where } \hat{s} = \left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log\widetilde{p}(\theta|\boldsymbol{y})\right]^{-1/2}\Bigg|_{\theta=\hat{\theta}}.$$

5.3 Laplace approximation

Multivariate θ

- If θ is d-dimensional for $d \geq 1$, then the Laplace method approximates the posterior density by

$$\widehat{p}(\theta|\boldsymbol{y}) = \frac{\widetilde{p}(\theta|\boldsymbol{y})}{\widehat{\mathcal{I}}} = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{(2\pi)^{d/2}\hat{s}p(\boldsymbol{y}|\hat{\theta})p(\hat{\theta})},$$

$$\text{where } \hat{s} = \left[\det\left\{-\frac{\partial^2}{\partial\theta^2}\log\widetilde{p}(\theta|\boldsymbol{y})\right\}\right]^{-1/2}\Bigg|_{\theta=\hat{\theta}}.$$

- The number $(2\pi)^{d/2}\hat{s}$ actually comes from the normalizing constant of a multivariate normal distribution in the previous normal approximation.

5.3 Laplace approximation

Approximation of posterior mean

- Now we can use the Laplace method to estimate $\mathrm{E}[g(\theta)|\boldsymbol{y}]$.
- Again, first assume that θ is a scalar. Since

$$\mathrm{E}[g(\theta)|\boldsymbol{y}] = \frac{\int g(\theta)p(\boldsymbol{y}|\theta)p(\theta)\mathrm{d}\theta}{\int p(\boldsymbol{y}|\theta)p(\theta)\mathrm{d}\theta},$$

we can directly apply the Laplace method for Equation (4) to the numerator with $f(\theta) = g(\theta)$ and the denominator with $f(\theta) = 1$, and derive the following estimate of $\mathrm{E}[g(\theta)|\boldsymbol{y}]$:

Approximation 1

$$\begin{aligned}\mathrm{E}[g(\theta)|\boldsymbol{y}] &\approx \frac{g(\hat{\theta})\sqrt{\frac{2\pi}{n}}\hat{\sigma}\exp\left\{-nh(\hat{\theta})\right\}}{1\cdot\sqrt{\frac{2\pi}{n}}\hat{\sigma}\exp\left\{-nh(\hat{\theta})\right\}} \\ &= g(\hat{\theta}).\end{aligned}$$

- Approximation 1 is simple but the order of error is large. It can be shown that $\mathrm{E}[g(\theta)|\boldsymbol{y}] = g(\hat{\theta})\left[1 + \mathcal{O}\left(\frac{1}{n}\right)\right]$.
- A better way is to approximate the numerator by choosing $f(\theta) = 1$ and $h^*(\theta) = -n^{-1}[\log g(\theta) + \log p(\boldsymbol{y}|\theta) + \log p(\theta)]$ in Equation (4).
- We define

$$\begin{aligned} h^*(\theta) &= -n^{-1}[\log g(\theta) + \log p(\boldsymbol{y}|\theta) + \log p(\theta)] \\ &= -n^{-1}\log g(\theta) + h(\theta), \\ \theta^* &= \arg\min_{\theta\in\Theta} h^*(\theta), \\ \hat{\sigma}^* &= \left[h^{*''}(\theta^*)\right]^{-1/2}. \end{aligned}$$

Approximation 2

$$\begin{aligned}\mathrm{E}[g(\theta)|\boldsymbol{y}] &\approx \frac{\sqrt{\frac{2\pi}{n}}\hat{\sigma}^* \exp\left\{-nh^*(\theta^*)\right\}}{1 \cdot \sqrt{\frac{2\pi}{n}}\hat{\sigma} \exp\left\{-nh(\hat{\theta})\right\}} \\ &= \frac{\hat{\sigma}^* g(\theta^*) p(\boldsymbol{y}|\theta^*) p(\theta^*)}{\hat{\sigma} p(\boldsymbol{y}|\hat{\theta}) p(\hat{\theta})}.\end{aligned}$$

- It can be show that

$$\mathrm{E}[g(\theta)|\boldsymbol{y}] = \frac{\hat{\sigma}^* g(\theta^*) p(\boldsymbol{y}|\theta^*) p(\theta^*)}{\hat{\sigma} p(\boldsymbol{y}|\hat{\theta}) p(\hat{\theta})} \left[1 + \mathcal{O}\left(\frac{1}{n^2}\right)\right].$$

- The last formula also holds for multivariate θ, by replacing $\hat{\sigma}$ and $\hat{\sigma}^*$ with the following formulas:

$$\hat{\sigma} = \left[\det \left\{ \frac{\partial^2}{\partial \theta^2} h(\theta) \right\} \right]^{-1/2} \Big|_{\theta=\hat{\theta}},$$

$$\hat{\sigma}^* = \left[\det \left\{ \frac{\partial^2}{\partial \theta^2} h^*(\theta) \right\} \right]^{-1/2} \Big|_{\theta=\theta^*},$$

where $\hat{\theta}$ and θ^* are the minimizers of $h(\theta)$ and $h^*(\theta)$, respectively.

5.3 Laplace approximation

Genetic Linkage Model

- We now revisit the genetic linkage model. First we approximate the normalizing constant in the posterior density using the Laplace method.
- Let $\hat{\theta}$ be the posterior mode. We use the formula on page 25 and page 16,

$$\begin{aligned}\hat{s} &= \left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log \widetilde{p}(\theta|\boldsymbol{y})\right]^{-1/2}\Bigg|_{\theta=\hat{\theta}} \\ &= \left[\frac{y_1}{(2+\hat{\theta})^2}+\frac{y_2+y_3}{(1-\hat{\theta})^2}+\frac{y_4}{\hat{\theta}^2}\right]^{-1/2}.\end{aligned}$$

- Then the Laplace method approximates the posterior density by

$$\widehat{p}(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{\sqrt{2\pi}\hat{s}p(\boldsymbol{y}|\hat{\theta})p(\hat{\theta})}.$$

5.3 Laplace approximation

Genetic Linkage Model

- We consider both $Y = (125, 18, 20, 34)$ and $Y = (14, 0, 1, 5)$.
- We only need to compute $\hat{\theta}$ and $\hat{s}$.

```
Y = c(125, 18, 20, 34)
Const <- integrate(post0, 0, 1, Y=Y)$value
out <- optimize(log.post, interval=c(0,1), Y=Y, maximum=TRUE)
theta.hat <- out$maximum
s.hat <- 1/sqrt(Y[1]/(2+theta.hat)^2+
        (Y[2]+Y[3])/(1-theta.hat)^2+Y[4]/theta.hat^2)
```

5.3 Laplace approximation

Genetic Linkage Model

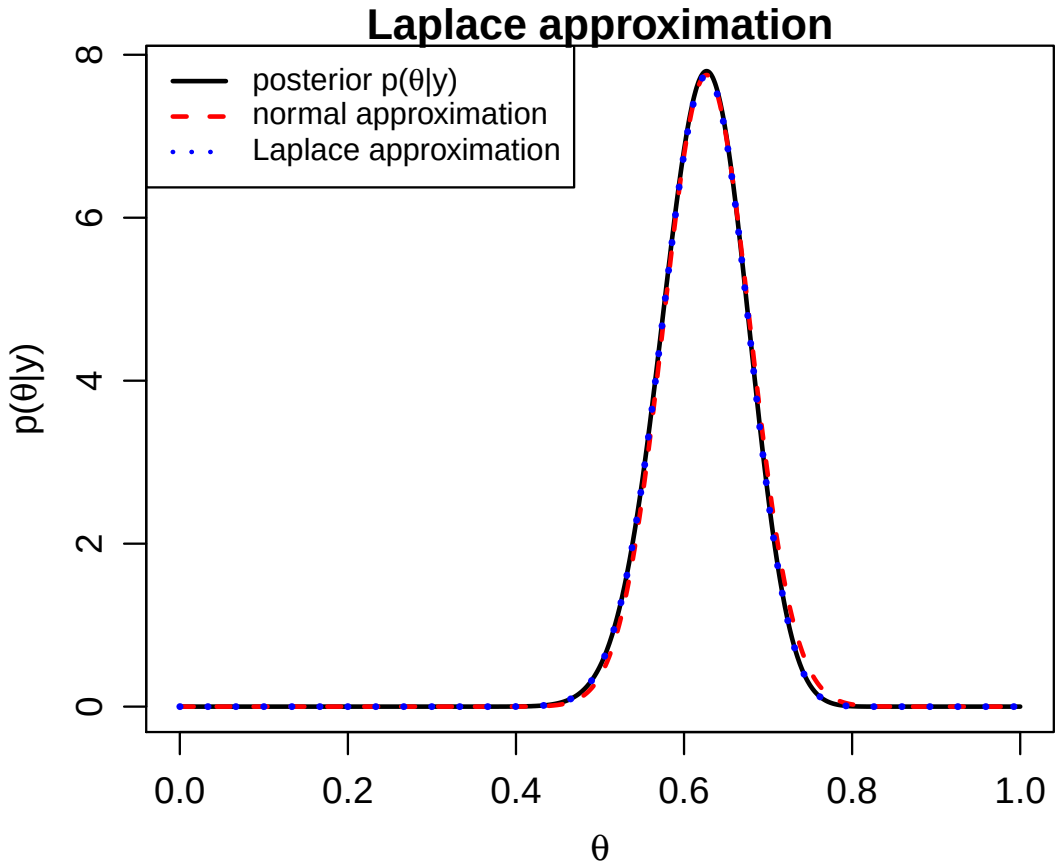
- Then we plot the true posterior density, the density from normal approximation, and the Laplace approximation together.

```
par(mar=c(3.5,3.5,1,1))
par(mgp=c(2.1,0.8,0))
theta.grid <- seq(from=0, to=1, by=0.001)
plot(theta.grid, post0(theta.grid,Y=Y)/Const, type="l", lwd=2,
     ylab=expression(paste("p(",theta,"|y)")),
     xlab=expression(theta),
     main="Laplace approximation")
points(theta.grid, dnorm(theta.grid,mean=theta.hat,sd=s.hat),
type="l", lwd=2, lty=2, col="red")
points(theta.grid,
       post0(theta.grid,Y)/(sqrt(2*pi)*s.hat*post0(theta.hat,Y)),
       type="l", lwd=2, lty=3, col="blue")
legend("topleft",legend =
       c(expression(paste("posterior p(", theta, "|y)")),
       "normal approximation", "Laplace approximation"),
       col=c("black","red","blue"), lty=c(1,2,3), lwd=2, cex=0.8)
```

5.3 Laplace approximation

Genetic Linkage Model

- For $Y = (125, 18, 20, 34)$,

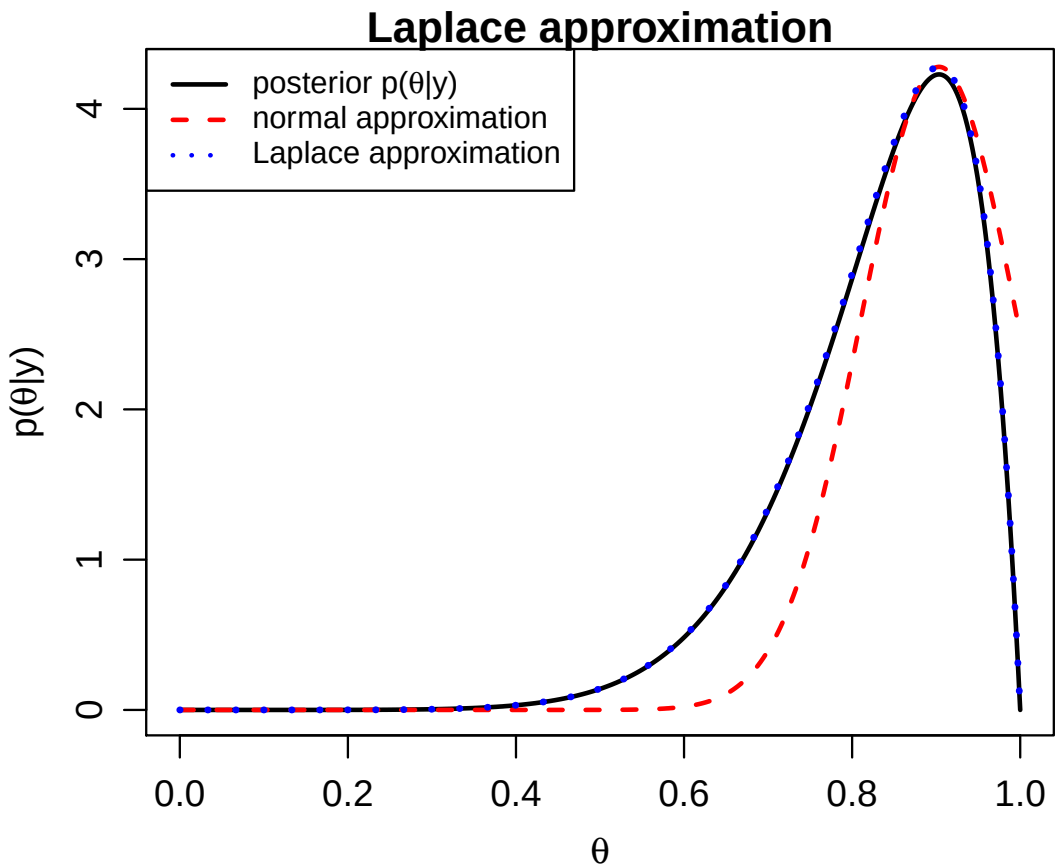


- Both normal approximation and Laplace approximation are accurate.

5.3 Laplace approximation

Genetic Linkage Model

- For $Y = (14, 0, 1, 5)$,



- Normal approximation cannot fit the skewness of the posterior well. Laplace approximation gives a very accurate fit.

5.3 Laplace approximation

Genetic Linkage Model

- We now consider the estimation of posterior mean.
- We use the approximation formulas on page 27 and 29.
- For posterior mean, we let $g(\theta) = \theta$.
- The $h^*(\theta)$ function and its derivatives are

$$h^*(\theta) = -\frac{1}{n}[\log\theta + \log\widetilde{p}(\theta|\boldsymbol{y})]$$

$$h^{*'}(\theta) = -\frac{1}{n}\left[\frac{1}{\theta} + \frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta}\right]$$

$$h^{*''}(\theta) = \frac{1}{n}\left[\frac{1}{\theta^2} + \frac{y_1}{(2+\theta)^2} + \frac{y_2+y_3}{(1-\theta)^2} + \frac{y_4}{\theta^2}\right].$$

- Then we can calculate $\hat{\sigma}^* = [h^{*''}(\theta^*)]^{-1/2}$.

5.3 Laplace approximation

Genetic Linkage Model

For $Y = (125, 18, 20, 34)$,

```
> numerator <- function(theta, Y) {
+ theta * (2+theta)^(Y[1])*(1-theta)^(Y[2]+Y[3])*theta^(Y[4])
+ }
> # true posterior mean
> (integrate(numerator, 0, 1, Y=Y)$value)/Const
[1] 0.6228061
> (theta.hat)      # Approximation 1
[1] 0.6268101
>
> out.star <- optimize(numerator, interval=c(0,1), Y=Y, maximum=TRUE)
> theta.star <- out.star$maximum
> s.star <- 1/sqrt(1/theta.star^2 + Y[1]/(2+theta.star)^2+
+           (Y[2]+Y[3])/(1-theta.star)^2 + Y[4]/theta.star^2)
> ((s.star*theta.star*post0(theta.star,Y))/
+  (s.hat*post0(theta.hat,Y)))      # Approximation 2
[1] 0.6227114
```

The true posterior mean is 0.6228. Approximation 2 gives a better estimate (0.6227) than Approximation 1 (0.6268).

5.3 Laplace approximation

Genetic Linkage Model

For $Y = (14, 0, 1, 5)$,

```
>  # true posterior mean
> (integrate(numerator, 0, 1, Y=Y)$value)/Const
[1] 0.831124
> (theta.hat)      # Approximation 1
[1] 0.9034481
>
> out.star <- optimize(numerator, interval=c(0,1), Y=Y, maximum=TRUE)
> theta.star <- out.star$maximum
> s.star <- 1/sqrt(1/theta.star^2 + Y[1]/(2+theta.star)^2+
+           (Y[2]+Y[3])/(1-theta.star)^2 + Y[4]/theta.star^2)
> ((s.star*theta.star*post0(theta.star,Y))/
+  (s.hat*post0(theta.hat,Y)))      # Approximation 2
[1] 0.8275301
```

The true posterior mean is 0.8311. Approximation 2 gives a much better estimate (0.8275) than Approximation 1 (0.9034).

- We can also estimate other posterior quantities, such as posterior standard deviation.
- We can set $g(\theta) = \theta^2$ and $g(\theta) = \theta$, and obtain the estimates of $\mathrm{E}(\theta^2|\boldsymbol{y})$ and $\mathrm{E}(\theta|\boldsymbol{y})$. Then we can plug in these estimates into the formula $\mathrm{Var}(\theta|\boldsymbol{y}) = \mathrm{E}(\theta^2|\boldsymbol{y}) - [\mathrm{E}(\theta|\boldsymbol{y})]^2$ and $\mathrm{sd}(\theta|\boldsymbol{y}) = \sqrt{\mathrm{Var}(\theta|\boldsymbol{y})}$.

5.3 Laplace approximation

Summary

- A brief summary of normal approximation and Laplace approximation:
  - Normal approximation can be used to approximate the full posterior.
  - Laplace approximation can be used to estimate the normalizing constant. Therefore, it can be used to approximate the full posterior, if the unnormalized posterior density is known.
  - Laplace approximation can also be used to estimate posterior expectations, by approximating the two integrals in the numerator and the denominator, respectively.

Outline

5.4 Example: Beta-binomial Model

We look at another example with multidimensional parameters.

- Tsutakawa et al. (1985) describe the problem of simultaneously estimating the rates of death from stomach cancer for males at risk in the age bracket 45–64 for the largest cities in Missouri.

- The data below shows the mortality rates for $n = 20$ cities, where each ordered pair contains the number of cancer deaths y_j and the number n_j at risk for a given city. Let $\boldsymbol{y} = (y_1, \ldots, y_{20})$.

(0, 1083)	(0, 855)	(2, 3461)	(0, 657)	(1, 1208)	(1, 1025)
(0, 527)	(2, 1668)	(1, 583)	(3, 582)	(0, 917)	(1, 857)
(1, 680)	(1, 917)	(54, 53637)	(0, 874)	(0, 395)	(1, 581)
(3, 588)	(0, 383)				

5.4 Example: Beta-binomial Model

- A first modeling attempt might assume that

$$y_j \overset{\text{indep}}{\sim} \text{Binomial}(n_j, p) \text{ for } j = 1, \ldots, n$$

where the probability of death is common across all cities.

- However, it can be shown that these data are overdispersed; the counts $\{y_j\}$ display more variation than would be predicted under a binomial model with a constant probability p.

- A better-fitting model assumes y_j is distributed from a beta-binomial model with parameters $0 < \eta < 1$ and $K > 0$:

$$p(y_j|\eta, K) = \binom{n_j}{y_j} \frac{\text{B}(K\eta + y_j, K(1-\eta) + n_j - y_j)}{\text{B}(K\eta, K(1-\eta))}, \quad y_j = 0, 1, \ldots, n_j.$$

where $\text{B}(\cdot, \cdot)$ denotes the beta function.

5.4 Example: Beta-binomial Model

- The beta-binomial distribution is the binomial distribution in which the probability of success at each trial is not fixed but random and follows the beta distribution. We can think of

$$y_j|p_j \overset{\text{indep}}{\sim} \text{Binomial}(n_j, p_j), \quad j = 1, \ldots, n$$

where p_j is a random variable and

$$p_j|K, \eta \overset{\text{indep}}{\sim} \text{Beta}(K\eta, K(1-\eta)).$$

- Then the likelihood $p(y_j|\eta, K)$ given earlier can be derived using

$$p(y_j|\eta, K) = \int p(y_j|p_j)p(p_j|K, \eta)\, \mathrm{d}p_j.$$

- This is the first hierarchical model we have seen so far.

5.4 Example: Beta-binomial Model

- Suppose we assign the parameters a "vague" prior for $\eta \in (0,1)$ and $K \in (0,\infty)$:

$$p(\eta, K) \propto \frac{1}{\eta(1-\eta)} \frac{1}{(1+K)^2}.$$

Note that this prior is improper (why?). Then the posterior density of (η, K) is

$$p(\eta, K|\boldsymbol{y}) \propto p(\eta, K) \prod_{j=1}^{20} p(y_j|\eta, K)$$

$$\propto \frac{1}{\eta(1-\eta)} \frac{1}{(1+K)^2} \prod_{j=1}^{20} \frac{\mathrm{B}(K\eta + y_j, K(1-\eta) + n_j - y_j)}{\mathrm{B}(K\eta, K(1-\eta))}.$$

This posterior is of a non-standard form and is known only up to a normalizing constant.

5.4 Example: Beta-binomial Model

- Let us first construct a contour plot of the parameters η and K.
- Taking logarithm of the posterior density,

$$\begin{aligned}\log p(\eta, K|\boldsymbol{y}) = &-\log\eta - \log(1-\eta) - 2\log(1+K) \\ &+ \sum_{j=1}^{20} \Big\{ \log B(K\eta + y_j, K(1-\eta) + n_j - y_j) \\ &- \log \mathrm{B}(K\eta, K(1-\eta)) \Big\} + C,\end{aligned}$$

where C is a constant not depending on η, K.

5.4 Example: Beta-binomial Model

R Implementation

- First, install the R package `LearnBayes`.
- We load the dataset `cancermortality` from the `LearnBayes` package and create vectors `y` and `n` for storing the values of $\{y_j\}$ and $\{n_j\}$ respectively.
- Then we write a function `logpost0` which computes the log posterior (up to an additive constant not depending on η, K). Note that we compute $\log \mathrm{B}(\cdot,\cdot)$ using `lbeta` which is more stable.

```
require(LearnBayes)          # load R package "LearnBayes"
data(cancermortality)        # load dataset "cancermortality"
y <- cancermortality$y       # vector containing y_j values
n <- cancermortality$n       # vector containing n_j values
logpost0 <- function(eta,K,y,n){
 (sum(lbeta(K*eta+y,K*(1-eta)+n-y) - lbeta(K*eta,K*(1-eta)))
  - log(eta) - log(1-eta) -2*log(1+K))
}
```

5.4 Example: Beta-binomial Model

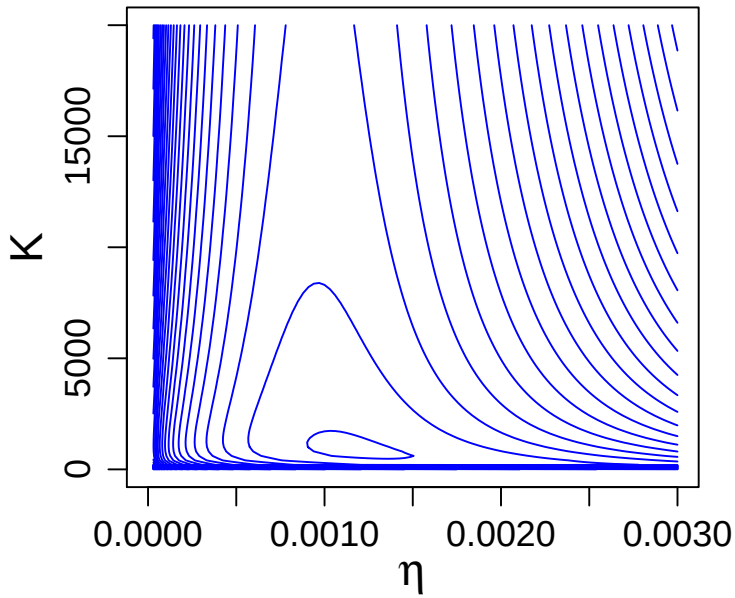
The contour plot can be obtained using the R-code below.

```
N <- 100
K <- seq(from=1, to=20000, length.out=N)
eta <- seq(from=0, to=0.003, length.out=N)
z <- matrix(0,N,N)
for (i in 1:N){
 for (j in 1:N){
   z[i,j] <- logpost0(eta[i],K[j],y,n)
}}
contour(x=eta,y=K,z,col="blue",nlevels=40,drawlabels=FALSE)
```

Alternatively, you may also use the `expand.grid` function to compute the function on 2d grids. See my previous R codes in Chapter 3.

5.4 Example: Beta-binomial Model

- The figure below shows a contour plot of the posterior density of (η, K).



- The contour plot of (η, K) shows strong skewness in the posterior density, especially towards large values of K. To reduce the skewness, we transform each parameter to the real line:

$$\theta_1 = \text{logit}(\eta) = \log \frac{\eta}{1-\eta}, \qquad \theta_2 = \log K.$$

5.4 Example: Beta-binomial Model

- Let $\theta = (\theta_1, \theta_2)$. As $\eta = \dfrac{e^{\theta_1}}{1 + e^{\theta_1}}$ and $K = e^{\theta_2}$,

$$\frac{d\eta}{d\theta_1} = \frac{e^{\theta_1}}{(1 + e^{\theta_1})^2} = \eta(1 - \eta), \quad \frac{dK}{d\theta_2} = e^{\theta_2} = K \quad \Rightarrow |J| = \eta(1 - \eta)K.$$

Here, J is the Jacobian matrix.

The posterior density $p(\theta|\boldsymbol{y})$ is then given by

$$p(\theta|\boldsymbol{y}) \propto p(\eta, K|\boldsymbol{y})|J|$$

$$\propto \frac{1}{\cancel{\eta(1 - \eta)}} \frac{1}{(1 + K)^2} \prod_{j=1}^{20} \frac{B(K\eta + y_j, K(1 - \eta) + n_j - y_j)}{B(K\eta, K(1 - \eta))} \cdot \cancel{\eta(1 - \eta)}K$$

$$\propto \frac{K}{(1 + K)^2} \prod_{j=1}^{20} \frac{B(K\eta + y_j, K(1 - \eta) + n_j - y_j)}{B(K\eta, K(1 - \eta))}.$$

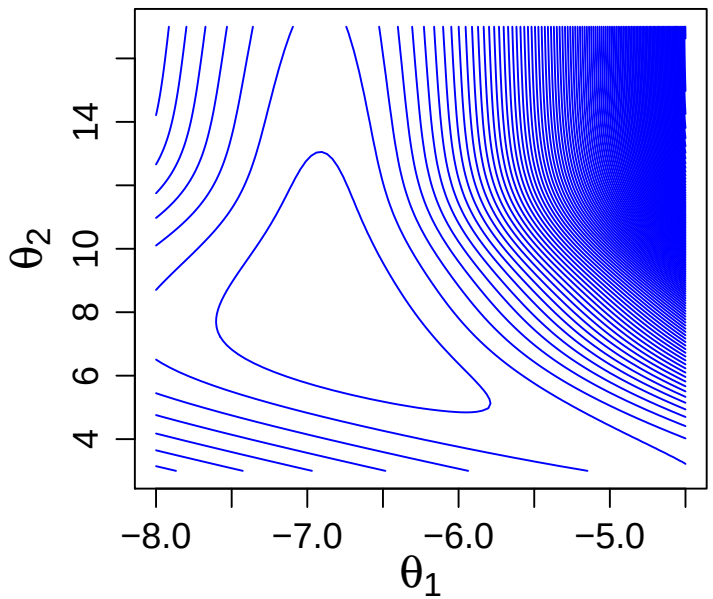
5.4 Example: Beta-binomial Model

- The log posterior density of the transformed parameters ($\log p(\theta|\boldsymbol{y})$) is programmed in the log-likelihood function below (up to an additive constant not depending on θ).

```
logpost <- function(theta,y,n){
  eta <- 1/(1+exp(-theta[1]))
  K <- exp(theta[2])
  L <- (sum(lbeta(K*eta+y,K*(1-eta)+n-y)
        - lbeta(K*eta,K*(1-eta))) + theta[2] - 2*log(1+K))
  return(L)
}
```

5.4 Example: Beta-binomial Model

- The contour plot of the log posterior of θ can be obtained similarly and is shown in the figure below.



- Although the posterior density has an unusual shape, the strong skewness has been reduced and the distribution is more amenable to the computational methods described in this chapter.

5.4 Example: Beta-binomial Model

- There are many alternative algorithms available for finding the posterior mode.
- In the following, we will use the Nelder-Mead algorithm, which is the default method in the R function `optim`.
- This algorithm uses only function values and is robust but relatively slow. It will work reasonably well for non-differentiable functions.
- The Nelder-Mead algorithm is sometimes preferable to Newton's method as it is less sensitive to the choice of starting value.

5.4 Example: Beta-binomial Model

- The usage of the function `optim` is as follows:

```
optim(par, fn, gr=NULL, ..., method=c("Nelder-Mead","BFGS",
    "CG","L-BFGS-B","SANN","Brent"), lower=-Inf, upper=Inf,
    control list(), hessian=FALSE)
```

par	: initial values for the parameters to be optimized
fn	: name of function to be minimized (or maximized)
gr	: a function to return the gradient (optional)
method	: name of method to be used (default is "Nelder-Mead")
lower, upper	: bounds on the variables
control	: a list of control parameters
hessian	: return a numerically differentiated Hessian matrix?

5.4 Example: Beta-binomial Model

- Let us find the normal approximation of $p(\theta|\boldsymbol{y})$ for the beta-binomial model. We have written a function `h` which is equal to $\log p(\theta|\boldsymbol{y})$ plus an additive constant. The mode of `h` and the hessian of `h` at the mode can be found as follows:

```
out <- optim(par=c(-7,7.5), fn=logpost, hessian=TRUE,
        control=list(fnscale=-1), y=y, n=n)
```

  - We set initial values for θ as $(-7, 7.5)$ based on the contour plot.
  - The parameter to be optimized, `theta`, must be the first argument of `logpost`. As `logpost` has other arguments `y` and `n`, we have to specify their values as well, inside the `optim` function directly.
  - Set `hessian=TRUE` so that the Hessian matrix is returned.
  - `optim` performs minimization by default. For maximization, add "`control=list(fnscale=-1)`".

5.4 Example: Beta-binomial Model

The output is :

```
$par
[1] -6.818978  7.573641
$value
[1] -571.3762
$counts
function gradient
      41       NA
$convergence
[1] 0
$hessian
           [,1]       [,2]
[1,] -15.974996 -1.7654723
[2,]  -1.765472 -0.9373622
```

- The mode $\hat{\theta}$ is given by `out$par`, and $\ell''(\hat{\theta})$ is given by `out$hessian`.
- If `out$convergence` is 0, then the algorithm is completed successfully.

5.4 Example: Beta-binomial Model

- The mean of the normal approximation is given by the posterior mode and the covariance matrix is given by negative of the inverse Hessian.

```
> (post.mode <- out$par)
[1] -6.818978  7.573641
> (post.cov <- -solve(out$hessian))
             [,1]       [,2]
[1,]  0.07905249 -0.1488912
[2,] -0.14889120  1.3472521
```

- Hence

$$\theta|y \sim \mathrm{N}\left(\begin{bmatrix}-6.82\\ 7.57\end{bmatrix}, \begin{bmatrix}0.08 & -0.15\\ -0.15 & 1.35\end{bmatrix}\right) \text{ approximately.}$$

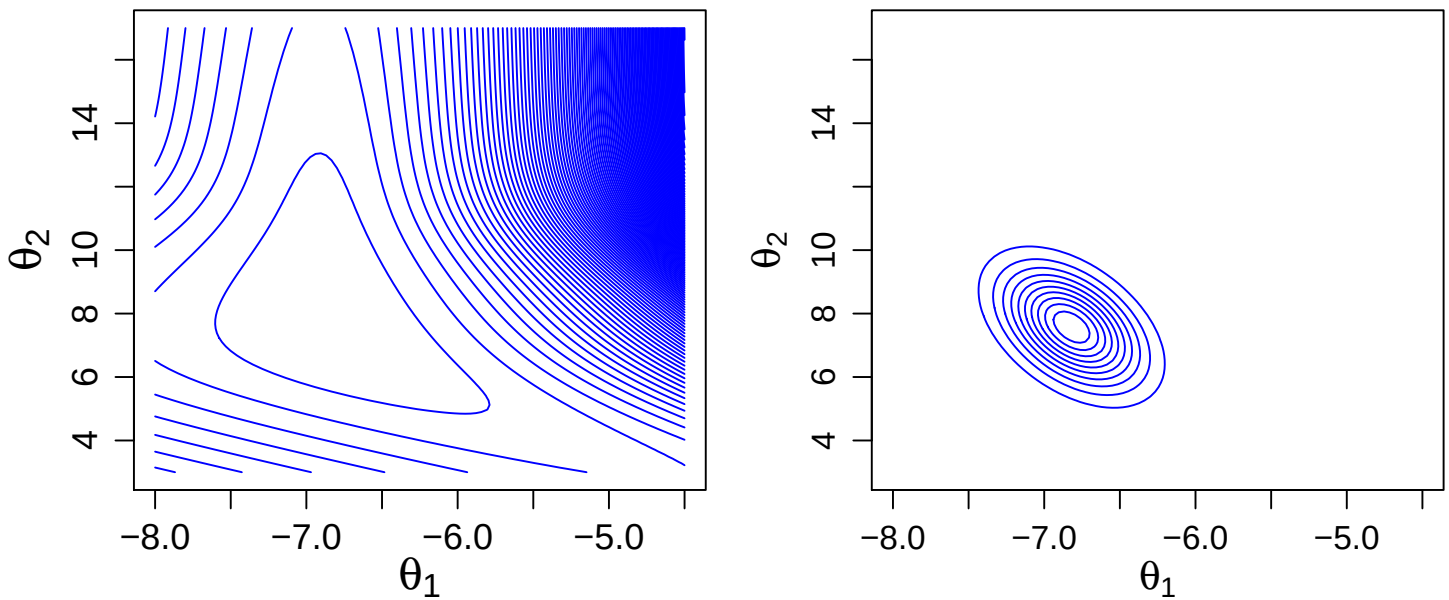
5.4 Example: Beta-binomial Model

- We can use the R-code below to plot the contours of the normal approximation. The R package `mvtnorm` provide functions `dmvnorm` for the density function and `rmvnorm` for random generation from the multivariate normal.

```
require(mvtnorm)
N <- 100
theta1 <- seq(from=-8, to=-4.5, length.out=N)
theta2 <- seq(from=3, to=17, length.out=N)
z <- matrix(0,N,N)
for (i in 1:N){
for (j in 1:N){
   theta <- c(theta1[i], theta2[j])
   z[i,j] <- dmvnorm(theta,mean=post.mode,sigma=post.cov)
}}
contour(x=theta1,y=theta2,z,drawlabels=FALSE)
```

5.4 Example: Beta-binomial Model

- The figure below shows the contour plot of the normal approximation to the posterior. We note that there are significant differences between the contours of the exact posterior and the approximate normal posterior. This is mainly because that in this example, the sample size $n = 20$ is small.



5.4 Example: Beta-binomial Model

- One advantage of this multivariate normal approximation is that we can obtain quick summaries of the parameters, using the property of normal distribution.
- For example, we can use the diagonal elements of the covariance matrix to construct approximate probability intervals for θ_1 and θ_2. The R-code below constructs 90% confidence intervals for the parameters:

```
> post.sd <- sqrt(diag(cov.mat))
> qnorm(c(0.05,0.95),mean=post.mode[1],sd=post.sd[1])
[1] -7.281449 -6.356506
> qnorm(c(0.05,0.95),mean=post.mode[2],sd=post.sd[2])
[1] 5.664440 9.482842
```

  A 90% CI for θ_1 is (-7.28, -6.36), and a 90% CI for θ_2 is (5.66, 9.48).
- However, if the approximation quality is poor (as we have seen in this example), then the accuracy of these CIs are questionable (compared to the true CIs from the true posterior of (θ_1, θ_2)).

Summary

A summary of normal approximation to a posterior distribution:

Step 1 : Suppose $p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)p(\theta) \propto f(\theta)$.

Step 2 : Let $h(\theta) = \log f(\theta)$.

Step 3 : Find the mode $\hat{\theta}$ by setting $h'(\theta) = 0$.

Step 4 : Find the inverse of the negative Hessian at the mode: $-[h''(\hat{\theta})]^{-1}$.

Step 5 : Normal approximation of $p(\theta|\boldsymbol{y})$: $\mathrm{N}(\hat{\theta}, -[h''(\hat{\theta})]^{-1})$.