Introduction to Bayesian Computation

Computing Integrals

The Bayesian recipe for inference is conceptually simple. If we observe data y from a sampling density $f(y|\theta)$, where θ is a vector of parameters and one assigns θ a prior $g(\theta)$, then the posterior density of θ is proportional to

$$g(\theta|y) \propto g(\theta)f(y|\theta).$$

The computational problem is to summarize this multivariate probability distribution to perform inference about functions of θ.

Many of the posterior summaries are expressible in terms of integrals. Suppose we are interested in the posterior mean of a function $h(\theta)$. This mean is expressible as a ratio of integrals,

$$E(h(\theta)|y) = \frac{\int h(\theta)g(\theta)f(y|\theta)d\theta}{\int g(\theta)f(y|\theta)d\theta}.$$

Computing Integrals

If we are interested in the posterior probability that $h(\theta)$ falls in a set A, we wish to compute

$$P(h(\theta) \in A|y) = \frac{\int_{h(\theta)\in A} g(\theta)f(y|\theta)d\theta}{\int g(\theta)f(y|\theta)d\theta}.$$

Integrals are also involved when we are interested in obtaining marginal densities of parameters of interest. Suppose we have the parameter $\theta = (\theta_1, \theta_2)$, where θ_1 are the parameters of interest and θ_2 are so-called nuisance parameters. One obtains the marginal posterior density of θ_1 by integrating out the nuisance parameters from the joint posterior:

$$g(\theta_1|y) \propto \int g(\theta_1, \theta_2|y)d\theta_2.$$

Computing Integrals to Estimate a Probability: Sleepless in Seattle

Let A and S denote respectively Annie's and Sam's arrival times, where we measure the arrival time as the number of hours after noon. We are assuming A and S are independent, where A is uniformly distributed on (10.5, 12) and S is uniformly distributed on (10, 11.5). We wish to compute the probability $P(A < S)$ which is expressed as the integral

$$P(A < S) = \int_{a<s} f_A(a) f_S(s) da ds,$$

where f_A and f_S denote the uniform densities for A and S. (By the independence assumption, the joint density $f_{A,S}(a, s) = f_A(a) f_S(s)$.) One can represent this probability as the expectation

$$P(A < S) = E[I(A < S)] = \int I(a < s) f_A(a) f_S(s) da ds,$$

where $I(a < s)$ is the indicator function that is equal to one when $a < s$, and zero otherwise.

Setting up a Bayesian Problem

One common situation is where one observes a random sample $y_1, ..., y_n$ from a sampling density $f(y|\theta)$ and one assigns θ the prior density $g(\theta)$. The logarithm of the posterior density of θ is given, up to an additive constant, by

$$\log g(\theta|y) = \log g(\theta) + \sum_{i=1}^{n} \log f(y_i|\theta).$$

Suppose we are sampling from a normal distribution with mean μ and standard deviation σ, the parameter vector $\theta = (\mu, \log \sigma)$, and we place an $N(10, 20)$ prior on μ and a flat prior on $\log \sigma$. The log posterior would have the form

$$\log g(\theta|y) = \log \phi(\mu; 10, 20) + \sum_{i=1}^{n} \log \phi(y_i; \mu, \sigma),$$

where $\phi(y; \mu, \sigma)$ is the normal density with mean μ and standard deviation σ. To program this function, we first write the simple function that evaluates the log likelihood of (μ, σ) for a component of y:

A Beta-Binomial Model for Overdispersion

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. Table 5.1 displays the mortality rates for 20 of these cities, where a cell contains the number n_j at risk and the number of cancer deaths y_j for a given city.

Table 5.1. Cancer mortality data. Each ordered pair represents the number of cancer deaths y_j and the number at risk n_j for an individual city in Missouri.

(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)				

A Beta-Binomial Model for Overdispersion

A first modeling attempt might assume that the $\{y_j\}$ represent independent binomial samples with sample sizes $\{n_j\}$ and common probability of death p. But it can be shown that these data are overdispersed in the sense that 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 that y_j is distributed from a beta-binomial model with mean η and precision K:

$$f(y_j|\eta, K) = \binom{n_j}{y_j} \frac{B(K\eta + y_j, K(1-\eta) + n_j - y_j)}{B(K\eta, K(1-\eta))}.$$

A Beta-Binomial Model for Overdispersion

Suppose we assign the parameters the vague prior proportional to

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

Then the posterior density of (η, K) is given, up to a proportionality constant, by

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

where $0 < \eta < 1$ and $K > 0$.

A Beta-Binomial Model for Overdispersion

Following the general guidance in Section 5.3, suppose we transform each parameter to the real line by using the reexpressions

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

The posterior density of (θ_1, θ_2) is given by

$$g_1(\theta_1, \theta_2|\text{data}) = g\left(\frac{e^{\theta_1}}{1+e^{\theta_1}}, e^{\theta_2}\right)\frac{e^{\theta_1+\theta_2}}{(1+e^{\theta_1})^2},$$

where the right term in the product is the Jacobian term in the transformation. The log posterior density of the transformed parameters is programmed in the function `betabinexch`. Note the change in the next-to-last line of the function that accounts for the logarithm of the Jacobian term.

A Beta-Binomial Model for Overdispersion

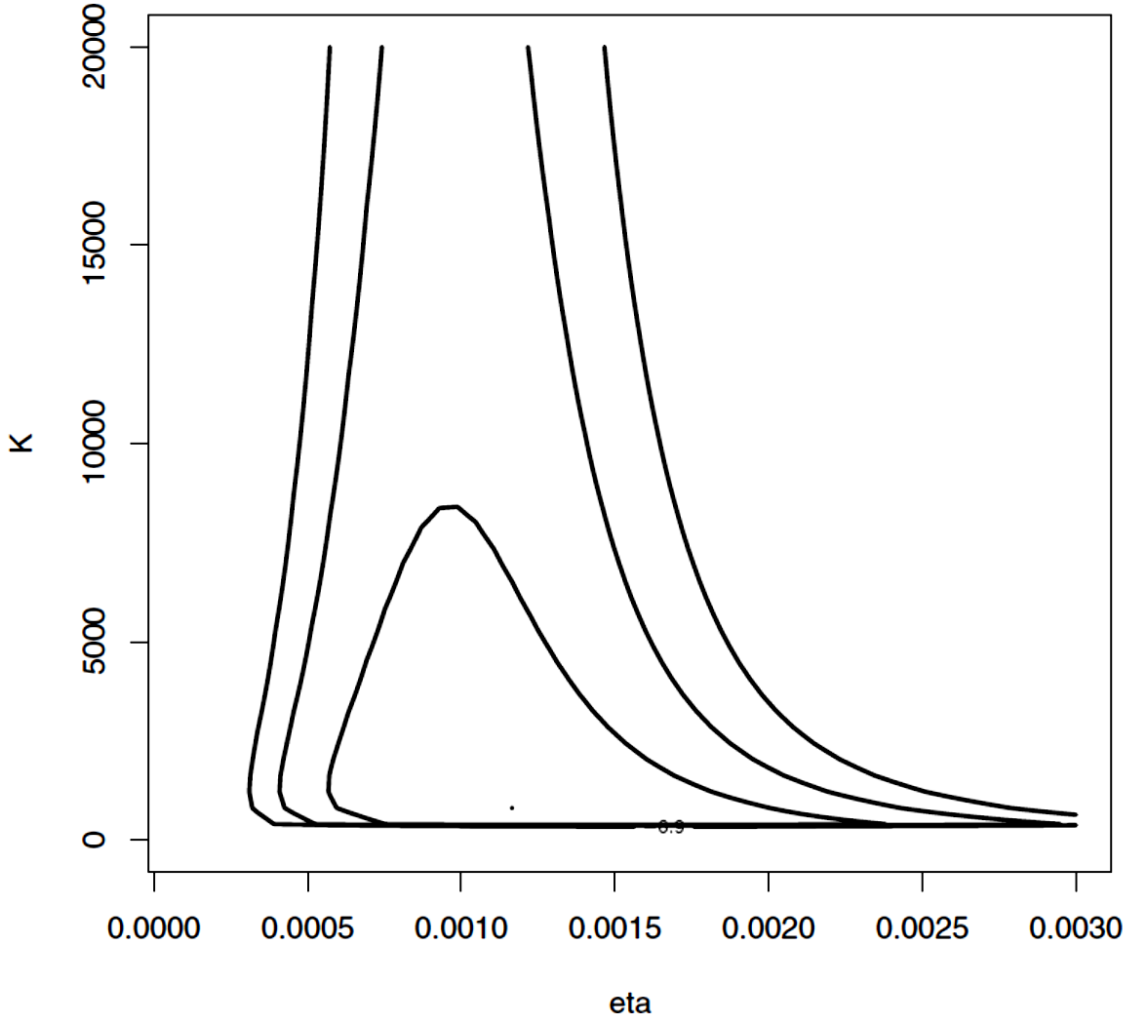

Fig. 5.1. Contour plot of parameters η and K in the beta-binomial model problem.

A Beta-Binomial Model for Overdispersion

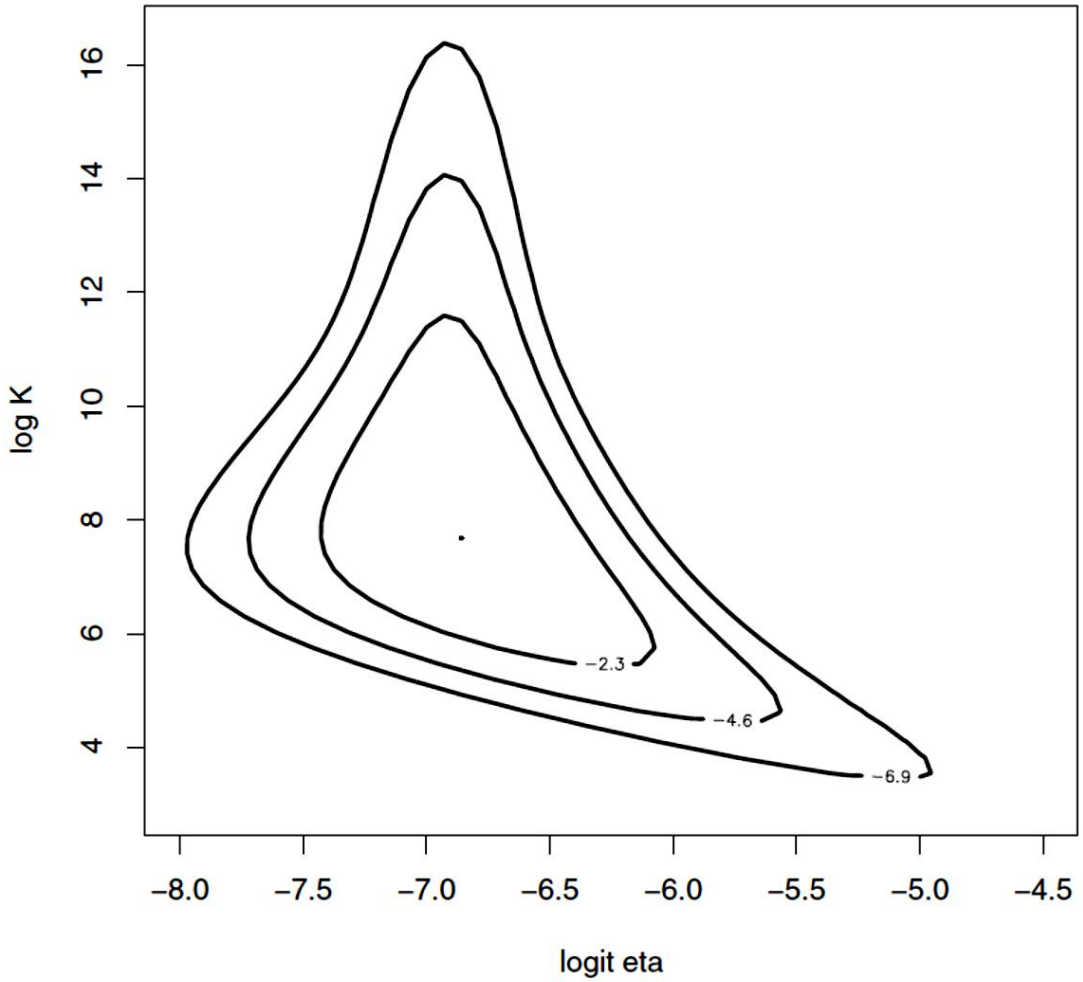

Fig. 5.2. Contour plot of transformed parameters $\text{logit}(\eta)$ and $\log K$ in the beta-binomial model problem.

Approximations Based on Posterior Modes

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 $g(\theta)$. If we observe data y with sampling density $f(y|\theta)$, then consider the logarithm of the joint density of θ and y,

$$h(\theta, y) = \log(g(\theta)f(y|\theta)).$$

In the following, we write this log density as $h(\theta)$ since after the data are observed θ is the only random quantity. Denoting the posterior mode of θ by $\hat{\theta}$, we expand the log density in a second-order Taylor series about $\hat{\theta}$. This gives the approximation

$$h(\theta) \approx h(\hat{\theta}) + (\theta - \hat{\theta})' h''(\hat{\theta})(\theta - \hat{\theta})/2,$$

where $h''(\hat{\theta})$ is the Hessian of the log density evaluated at the mode. Using this expansion, the posterior density is approximated by a multivariate normal density with mean $\hat{\theta}$ and variance-covariance matrix

$$V = (-h''(\hat{\theta}))^{-1}.$$

Approximations Based on Posterior Modes

In addition, this approximation allows one to analytically integrate out θ from the joint density and obtain the following approximation to the prior predictive density,

$$f(y) \approx (2\pi)^{d/2} g(\hat{\theta}) f(y|\hat{\theta})| - h''(\hat{\theta})|^{1/2},$$

where d is the dimension of θ.

Approximations Based on Posterior Modes

To apply this approximation, one needs to find the mode of the posterior density of θ. One general-purpose optimization algorithm for finding this mode is provided by Newton's method. Suppose one has a guess at the posterior mode θ^0. If θ^{t-1} is the estimate at the mode at the $t-1$ iteration of the algorithm, then the next iterate is given by

$$\theta^t = \theta^{t-1} - [h''(\theta^{t-1})]^{-1} h'(\theta^{t-1}),$$

where $h'(\theta^{t-1})$ and $h''(\theta^{t-1})$ are the gradient and Hessian of the log density evaluated at the current guess at the mode. One continues these iterations until convergence. 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` in the R base package. This algorithm is an iterative method based on the evaluation of the objective function over vertices of a simplex (a triangle for two variables). For the examples described in this book, the Nelder-Mead algorithm appears to be preferable to Newton's method since it is less sensitive to the choice of starting value.

The Example

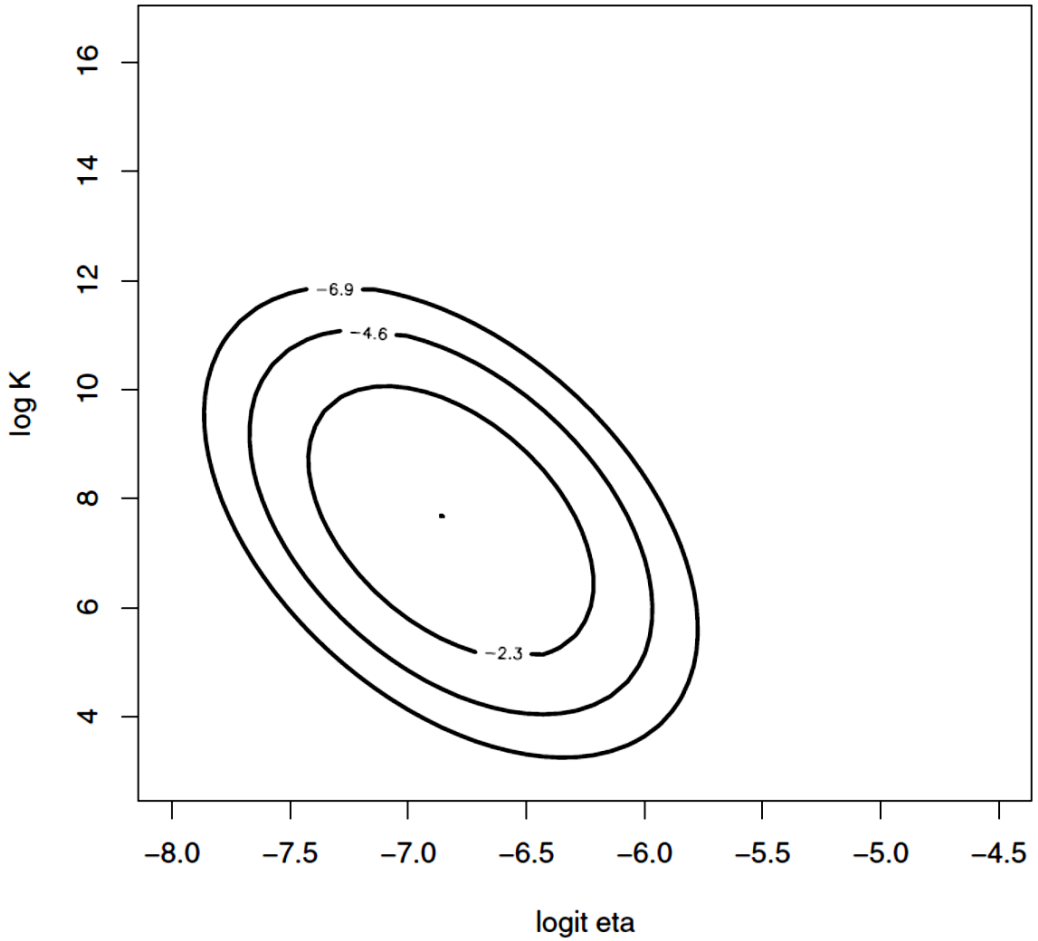

Fig. 5.3. Contour plot of normal approximation of logit(η) and log K in the beta-binomial model problem.