L6: Bayesian Computation: Numerical Methods

1 General Problem: Integration

Integration, in the continuous parameter case, or summation, in the discrete parameter case, is central to Bayesian inference, and arises in several areas.

1. Calculating the normalising constant, $m(\boldsymbol{y})$.

   The posterior distribution, $p(\theta|\boldsymbol{y})$:

   $$p(\theta|\boldsymbol{y}) = \frac{p(\theta, \boldsymbol{y})}{m(\boldsymbol{y})} = \frac{\pi(\theta) f(\boldsymbol{y}|\theta)}{\int \pi(\theta) f(\boldsymbol{y}|\theta) \mathrm{d}\theta} \tag{1}$$

   where $\boldsymbol{y}$ is the observed sample data, which could be a scalar, a vector, a matrix, etc, and θ could be a scalar, a vector, a matrix, etc.

   For a given value of θ, calculation of the numerator is often feasible as that involves evaluating the prior distribution at θ, $\pi(\theta)$, and the likelihood at θ, $f(\boldsymbol{y}|\theta)$ ($\equiv L(\theta|\boldsymbol{y})$).

   Calculation of the denominator, however, can be a difficult integration problem, particularly when θ is multidimensional, e.g., $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$.

2. Calculating marginal posterior distributions. Given $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$, the posterior distribution for a single component, θ_i, requires integrating over the other q-1 components.

   $$p(\theta_i|\boldsymbol{y}) = \int p(\theta_1, \theta_2, \ldots, \theta_q|\boldsymbol{y} \mathrm{d}\theta_1 \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_{i-1} \mathrm{d}\theta_{i+1} \ldots \mathrm{d}\theta_q \tag{2}$$

3. Numerical summaries of the posterior distribution. For example, letting θ be a scalar, to calculate the posterior mean, $E[\theta|\boldsymbol{y}]$:

   $$E[\theta|\boldsymbol{y}] = \int \theta p(\theta|\boldsymbol{y}) d\theta \tag{3}$$

   Calculation of the probabilities such as $\Pr(\theta > \theta^*)$, where θ is a scalar, also requires integration

   $$\Pr(\theta > \theta^*) = \int_{\theta^*}^{\infty} p(\theta|\boldsymbol{y}) d\theta \tag{4}$$

   Or with a continuous random variable, for arbitrarily small $\epsilon > 0$, "$\Pr(\theta = \theta^*)$":

   $$\Pr(\theta^* - \epsilon \leq \theta \leq \theta^* + \epsilon) = \int_{\theta^* - \epsilon}^{\theta^* + \epsilon} p(\theta|\boldsymbol{y}) d\theta \tag{5}$$

4. The posterior predictive distribution for future or unobserved sample data. Letting y^{new} denote a future or unobserved value, the distribution is found by integration:

   $$p(y^{new}|\boldsymbol{y}^{old}) = \int p(y^{new}|\theta, \boldsymbol{y}^{old}) p(\theta|\boldsymbol{y}^{old}) \mathrm{d}\theta \tag{6}$$

   Reminder: In many cases, y^{new} is conditionally independent of $\boldsymbol{y}$: $p(y^{new}|\theta, \boldsymbol{y}^{old}) = f(y^{new}|\theta)$.

In most of the examples given so far, the integration can be carried out *analytically*, i.e., there were exact analytic solutions for the posterior distribution. This has been the case with the conjugate prior distributions:

- Beta prior for θ when $y \sim$ Binomial(n,θ),
- Dirichlet prior for $\theta_1, \theta_2, \ldots, \theta_k$ when $y_1, y_2, \ldots, y_k \sim$ Multinomial($n, \theta_1, \theta_2, \ldots, \theta_k$).
- Gamma prior for θ when $y \sim$ Poisson(θ).
- Gamma prior for θ when $y \sim$ Exponential(θ).
- Pareto prior for θ when $y \sim$ Uniform(0,θ).
- Normal prior for μ when $y \sim$ Normal(μ, σ^2), when σ^2 is known.
- Inverse Gamma prior for σ^2 when $y \sim$ Normal(μ, σ^2), when μ is known.
- Inverse Gamma-Conditional Normal prior for σ^2 and μ when $y \sim$ Normal(μ, σ^2).

However, for many of the models being used, exact solutions are the exception not the rule, and approximate solutions are used.

2 Overview of integration methods

We will denote the integral to be evaluated generically as $I(\cdot)$, where $\cdot$ is some value; e.g., $I(\boldsymbol{y}) = \int \pi(\theta) p(\boldsymbol{y}|\theta) d\theta$.

In most of the methods to be discussed below, one will arrive at an estimate of $I(\cdot)$, which will be denoted $\hat{I}(\cdot)$, thus $\hat{I}(\cdot) \approx I(\cdot)$.

A broad categorization of methods for calculating $I(\cdot)$ is *Deterministic* and *Stochastic*. With the deterministic methods, the result is a single constant value. If you carry out the method and get $\hat{I}(\boldsymbol{y})$ =3, and if someone else carries out the same deterministic method, they will get the same $\hat{I}(\boldsymbol{y}) = 3$. With stochastic methods, however, you might get $\hat{I}(\boldsymbol{y}) = 3.02$, and another person will get $\hat{I}(\boldsymbol{y})$=2.93, and if you were to repeat the method yourself a second time, you might get $\hat{I}(\boldsymbol{y}) = 3.04$.

Our focus is on integration for Bayesian inference where the integrals involve probability distributions. Methods used for Bayesian inference (but not just Bayesian inference), that lie within each of these broad categories, include the following.

- Deterministic
  - Numerical integration (quadrature)
  - Asymptotic approximations: Bayesian Central Limit Theorem, Laplace Approximation
- Stochastic
  - Monte Carlo integration with Independent samples, both sampling from the "right" distribution (e.g., Direct sampling) and sampling from the "wrong" distribution (e.g., Rejection sampling, Importance sampling, Sampling Importance Resampling (SIR), Sequential Importance Sampling)
  - Monte Carlo integration with Dependent samples from "wrong" distribution, especially Markov Chain Monte Carlo (MCMC) (e.g., Metropolis-Hastings, Gibbs sampling)

3 Example: Modelling time between hurricanes

(Example from Gordon Ross.) An island community regularly experiences large and damaging hurricanes. The government wishes to build a statistical model that can quantify the probability of such hurricanes occurring in the future. Over the last 50 years, there have been 21 recorded large hurricanes. The number of years which occur between each hurricane (known as the inter-event times, of which there are n=20) are:

```
hurricane.gaps <- c( 0.30, 4.61, 5.75, 0.24, 0.09, 0.18, 7.38, 1.20, 2.40, 0.18,
  0.02, 10.07, 0.23, 0.44, 3.34, 0.06, 0.01, 0.71, 0.06, 0.42)
```

The goal is to estimate the distribution of the time between hurricanes.

Let y_i denote the time between hurricane i and hurricane $i+1$. A simple sampling model for the y_i's, which assumes independence between the times till the event occurs, is the Exponential distribution, a commonly used model for times between events[1]. Letting $\boldsymbol{y}$=$(y_1, \ldots, y_n)$, the sampling distribution is

$$f(\boldsymbol{y}|\lambda) = \prod_{i=1}^{n} \lambda \exp(-y_i\lambda) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} y_i) = \lambda^n \exp(-\lambda n \bar{y}) \tag{7}$$

which is the kernel for a Gamma distribution. Thus the conjugate prior for the exponential distribution is Gamma(α,β). The posterior distribution is then Gamma($\alpha + n$, $\beta + n\bar{y}$). For a "relatively uninformative" prior, let $\alpha = 0.01$ and $\beta = 0.01$, and with n=20 and $\bar{y}$=1.8845, the posterior is Gamma(20.01,37.70).

A diagnostic for goodness of fit is to compare the posterior predictive distribution to the observed distribution, e.g., the empirical density. The posterior predictive density for $\tilde{y}$, letting α'=$\alpha + n$ and β'=$\beta + n\bar{y}$:

$$p(y^{new}|\boldsymbol{y}^{old}) = \int \lambda \exp(-\lambda y^{new}) \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} \lambda^{\alpha'-1} \exp(-\lambda\beta') \mathrm{d}\lambda = \frac{\beta'^{\alpha'}\alpha'}{(\beta' + y^{new})^{\alpha'+1}}$$

Figure 1 compares the empirical density and the posterior predictive densities. Also plotted is the density using the mle for λ ($1/\bar{y}$= 0.5306), which is nearly identical to the posterior predictive density. The most important thing to note is the mismatch between the empirical and the modelled results, indicating poor goodness of fit.

An alternative distribution that is more flexible and may fit the data better is the *Weibull*(α, β) distribution:

$$f(y|\alpha, \beta) = \frac{\alpha}{\beta} \left(\frac{y}{\beta}\right)^{\alpha-1} \exp\left(-(y/\beta)^{\alpha}\right) \tag{8}$$

The Weibull is often used to model between-event times. The exponential distribution is a special case with α=1 and β=$1/\lambda$. The parameter α in the Weibull thus provides more flexibility than the exponential.

For a Bayesian analysis, the most general approach is to specify a joint prior for (α, β). If α is fixed, Γ^{-1} is conjugate for β. There is a conjugate prior for α, but it is awkward to work with. For simplicity (and to demonstrate some points about integration), we assume that β=1 and specify a non-conjugate prior for α, Gamma(κ, λ). The likelihood for α:

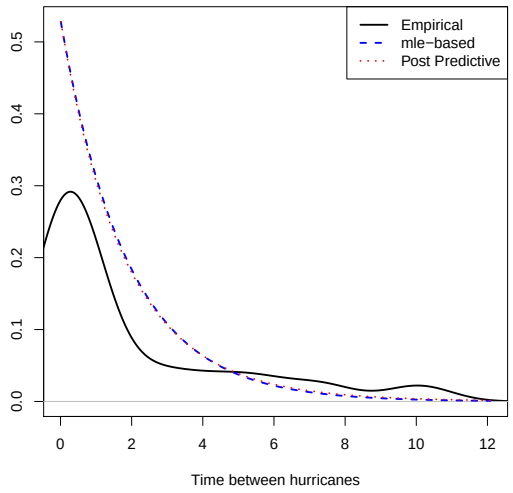
$$f(\boldsymbol{y}|\alpha, \beta = 1) \equiv L(\alpha|\boldsymbol{y}, \beta = 1) = \prod_{i=1}^{n} \alpha \left(y_i^{\alpha-1}\right) \exp\left(-y_i^{\alpha}\right) = \alpha^n \exp\left(-\sum_{i=1}^{n} y_i^{\alpha}\right) \prod_{i=1}^{n} y_i^{\alpha-1}$$

and the prior for α:

$$\pi(\alpha|\kappa, \lambda) = \frac{\lambda^{\kappa}}{\Gamma(\kappa)} \alpha^{\kappa-1} \exp(-\lambda\alpha)$$

[1]Recall the example from L1 notes for waiting time to see a teller in a bank.

Figure 1: Empirical, (mle) fitted, and post predictive densities for between hurricane times. (The mle fitted and posterior predictive are nearly exactly aligned.)



Then the posterior:

$$
\begin{aligned}
p(\alpha|\boldsymbol{y}) &= \frac{p(\alpha,\boldsymbol{y})}{m(\boldsymbol{y})} = \frac{\pi(\alpha)f(\boldsymbol{y}|\alpha)}{\int \pi(\alpha)f(\boldsymbol{y}|\alpha)\mathrm{d}\alpha} \\
&= \frac{\frac{\lambda^\kappa}{\Gamma(\kappa)}\alpha^{\kappa-1}\exp(-\lambda\alpha)\times\alpha^n\exp\left(-\sum_{i=1}^n y_i^\alpha\right)\prod_{i=1}^n y_i^{\alpha-1}}{\int \frac{\lambda^\kappa}{\Gamma(\kappa)}\alpha^{\kappa-1}\exp(-\lambda\alpha)\times\alpha^n\exp\left(-\sum_{i=1}^n y_i^\alpha\right)\prod_{i=1}^n y_i^{\alpha-1}\mathrm{d}\alpha}
\end{aligned}
\tag{9}
$$

where the marginal distribution for $\boldsymbol{y}$, the denominator of (9) is a complicated integral. Such complex integrals often occur for nonconjugate priors.

4 Deterministic Numerical Integration

While stochastic methods for estimating integrals are sometimes called "numerical" methods, here we will use the term "numerical" for non-stochastic, or deterministic methods. These methods are also called quadrature. Such methods are a focus of a numerical methods course and we only briefly discuss basic and elementary forms, refer to some R software, and discuss some limitations of such methods.

4.1 Grid-based methods for calculating posterior expectations

Grid-based integration methods are a numerical method for calculating expectations. Consider a one-dimensional parameter, θ, where the objective is to calculate the expected value of a function of the parameter, $g(\theta)$.

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

For simplicity, assume that the support (domain) of θ is a finite interval, $[\theta_L, \theta_U]$. Partition the interval into m equal length intervals, let θ_i^* be the midpoint of the ith interval. The expected value can be estimated as follows.

$$\widehat{E}[g(\theta)|\boldsymbol{y}] = \sum_{i=1}^{m} g(\theta_i^*) W_i$$

where

$$W_i = \frac{\pi(\theta_i^*) f(\boldsymbol{y}|\theta_i^*)}{\sum_{j=1}^{m} \pi(\theta_j^*) f(\boldsymbol{y}|\theta_j^*)}$$

4.2 Simple numerical integration example

Consider a one-dimensional integral of the form:

$$\int_a^b f(x)dx$$

which may or may not have an analytic solution. The gist of many of the numerical solutions is to first partition $[a, b]$ into n subintervals:

$$[x_0 = a, x_1], [x_2, x_3], \ldots, [x_{n-1}, x_n = b]$$

Thus breaking the integration problem into n components:

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx$$

Then the integral over each subinterval is approximated using some relatively simple formula or rule. For example, the integral within a subinterval is estimated by the area of a rectangle with width equal to the width of the subinterval and height equal to the average of the function evaluated at the endpoints of the interval:

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \text{ width * } \overline{\text{height}} = (x_{i+1} - x_i) \times \frac{f(x_i) + f(x_{i+1})}{2}$$

This particular approach is called the Trapezoid rule.

For example consider the integral $\int_0^3 \frac{1}{2} e^{-x/2} \mathrm{d}x$, which does have an analytic solution (the integrand is the pdf for an Exponential(1/2) random variable):

$$\int_0^3 \frac{1}{2} e^{-x/2} dx = -e^{-x/2}|_0^3 = 1 - \exp(-1.5) = 0.7768698$$

Breaking [0,3] into four intervals of equal length (3/4=0.75) and applying the trapezoid rule:

$$\begin{aligned}
\int_0^3 \frac{1}{2} e^{-x/2} dx &= \int_0^{0.75} \frac{1}{2} e^{-x/2} dx + \int_{0.75}^{1.50} \frac{1}{2} e^{-x/2} dx + \int_{1.5}^{2.25} \frac{1}{2} e^{-x/2} dx + \int_{2.25}^{3} \frac{1}{2} e^{-x/2} dx \\
&\approx \frac{3}{4} \frac{0.5(e^{-0.75/2} + e^{-0/2})}{2} + \frac{3}{4} \frac{0.5(e^{-1.5/2} + e^{-0.75/2})}{2} + \frac{3}{4} \frac{0.5(e^{-2.25/2} + e^{-1.5/2})}{2} + \frac{3}{4} \frac{0.5(e^{-3.0/2} + e^{-2.25/2})}{2} \\
&= 0.3163667 + 0.2174355 + 0.1494411 + 0.1027092 = 0.7859525
\end{aligned}$$

Increasing the number of subintervals improves the approximation, e.g., with 20 subintervals the estimate is 0.777234.

The general solution is to

- Select $m+1$ points within each subinterval; for interval i, $[x_i, x_{i+1}]$: the m points are x_{ij}, $j = 0, 1, 2, \ldots, m$:

$$[x_i = x_{i0} \leq x_{i1} < x_{i2} < \ldots < x_{i,m-1} \leq x_{i,m} = x_{i+1}]$$

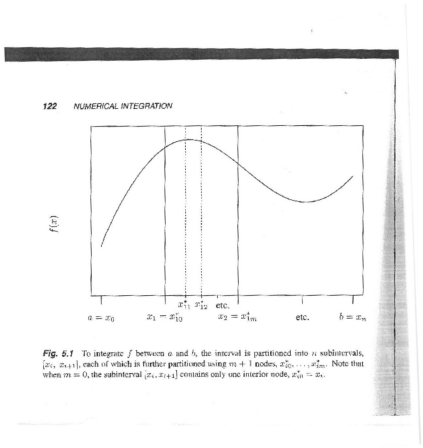
  Figure 2 shows the subintervals and points within an interval.

- Evaluate the function at each of these points, $f(x_{ij})$.

- Estimate the integral with a weighted combination of the function evaluations

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \sum_{j=1}^{m} w_{ij} f(x_{ij})$$

Figure 2: Partitions of an integral with points within an interval. From Givens and Hoeting, Computational Statistics, 2013, p130.

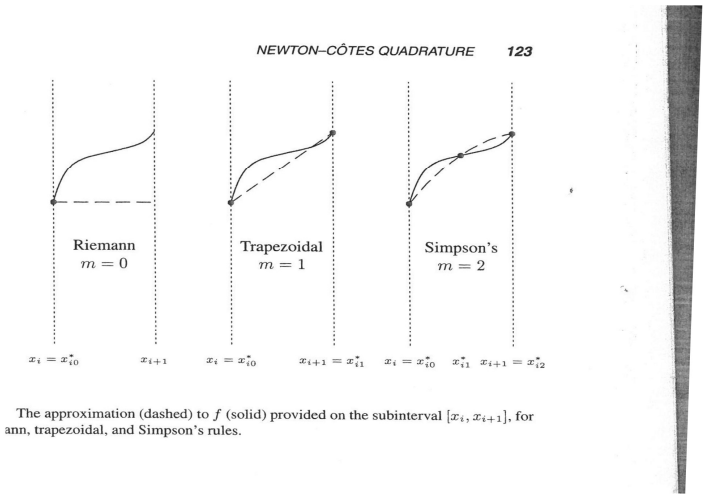


With the Trapezoid rule, m=1, and the $m+1$=2 selected points are the endpoints of the subinterval. Two other simple rules are the Riemann rule, m=0, where one of the endpoints of the subinterval is selected, and Simpson's rule, m=2, where the endpoints x_i and x_{i+1} are used as well as the midpoint, $(x_i + x_{i+1})/2$. Figure 3 shows how the three rules differ.

Demonstration with R. The above methods, while simple to understand, are relatively slow, and inefficient compared to other more sophisticated methods, where, for example, subinterval widths need not be equal[2]. R has a built-in function called `integrate` that can be applied to one dimensional integrals, which uses a more sophisticated procedure. The R package `pragca` has a similar function called `integral`.

[2]A highly readable chapter on numerical integration can be found in Computational Statistics, 2*nd* Edition by Givens and Hoeting (2013).

Figure 3: Three rules: Reimann, Trapezoid, and Simpson, for evaluation of the integral over a subinterval. From Givens and Hoeting, Computational Statistics, 2013, p131.

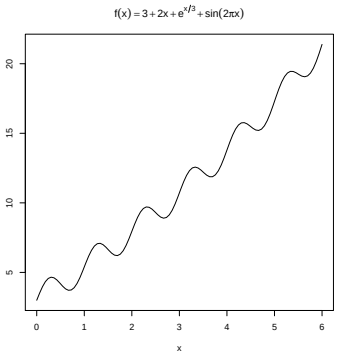


To demonstrate the usage of the R functions, consider the integral:

$$\int_0^5 2 + 2x + \exp(x/3) + \sin(2\pi x)dx$$

which can be evaluated analytically and equals 47.88347. The function is plotted in Figure 4.

Figure 4: Function, $f(x) = 2+2x+\exp(x/3)+\sin(2\pi x)$, to demonstrate R's integrate and integral functions.



The R code that describes the function, plots it over (0,6), and estimates the integral using `integral` and `integrate` functions is shown below.

```
# example function to be integrated over (0,5)
```

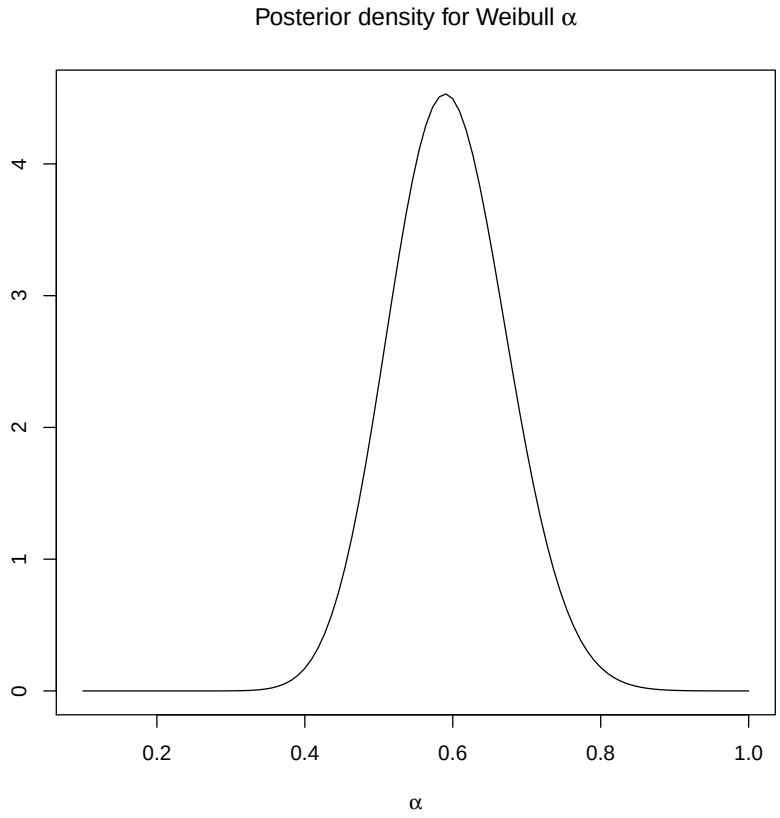
```
f <- function(x) {
   2+2*x+exp(x/3) + sin(2*pi*(x))
}
x.seq <- seq(0,6,length=100)
plot(x.seq,f(x.seq),xlab="x",ylab="",main= expression(f(x)==paste(3+2*x+e^{x/3}+sin(2*pi*x))),
     type="l")

antideriv <- function(x) 2*x+x^2+3*exp(x/3)-(1/2*pi)*cos(2*pi*x)
true.value <- antideriv(5)-antideriv(0)

cat("true value=",true.value,
    "pracma:",integral(f=f,0,5),
    "base:",integrate(f=f,0,5)$value,"\n")
#  true value= 47.88347 pracma: 47.88347 base: 47.88347
```

Demonstration with Weibull example. R code using the `pracma` integral function to estimate both the normalising constant as well as the posterior distribution for the Weibull example is given in Appendix A.2. The normalising constant was estimated to be 2.008407e-14. The posterior distribution for α is shown in Figure 5.

Figure 5: Posterior distribution for the α parameter, $p(\alpha|\boldsymbol{y})$, in the Weibull example of times between hurricanes based on R's integral function.



4.3 Multiple integrals

To numerically integrate a double integral where the region of integration is a rectangle, $(a \leq x \leq b)$ and $(c \leq y \leq d)$, i.e.,

$$\int_a^b \int_c^d f(x,y)dydx,$$

the double integral can be rewritten as:

$$\int_a^b g(x)dx$$

where

$$g(x) = \int_c^d f(x,y)dy$$

Quadrature methods can be applied in an iterative manner: estimating $g(x)$ by quadrature over $[c,d]$, then given a set of $g(x)$ values (over the sub-intervals of $[a,b]$) carry out quadrature over $[a,b]$. If there are n subintervals for x and y, then there are n^2 evaluations. Two practical problems:

- With higher dimension integrals the number of evaluations can become infeasible. Imagine a joint posterior $(\theta_1, \theta_2, \ldots, \theta_9)$ with n=5 subintervals: 5^9 = 1,953,125 evaluations are needed.
- The integration region will often not be rectangular (or a hyper-rectangle, $[a,b],[c,d],[e,f]$), and the problem becomes even more difficult[3].

Thus while understanding the principles of numerical integration are useful, and are sometimes used for small dimension θ, they "do not scale well with the number of parameters k" (Reich and Ghosh, p 74) and are not as commonly used as Monte Carlo methods.

[3]See http://www.aip.de/groups/soe/local/numres/bookfpdf/f4-6.pdf for more discussion of these difficulties.

5 Normal approximation to posterior

Suppose that the data $y_i \overset{iid}{\sim} f(y|\theta)$. Letting $\boldsymbol{y} = (y_1, \ldots, y_n)$, the joint probability is $f(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$. If n is relatively large, the likelihood will be relatively peaked and the small changes in the prior will have little effect on the posterior. Assume that a mode for the posterior distribution, denoted $\hat{\theta}^p$, exists (where in the multivariate case θ is a vector). Further assume that $\pi(\theta)f(\boldsymbol{y}|\theta)$ is positive and twice differentiable at the mode. Then under suitable regularity conditions the posterior distribution will be approximately normal with mean $\hat{\theta}^p$ and covariance matrix equal to the negative of the inverse of the second derivative matrix of the log posterior distribution evaluated at $\hat{\theta}^p$:

$$p(\theta|\boldsymbol{y}) \approx \text{Multivariate Normal}\left(\hat{\theta}^p, (I^p(\boldsymbol{y}))^{-1}\right) \tag{10}$$

where $I^p(\boldsymbol{y})$ is the "generalised" observed Fisher information matrix:

$$I^p(\boldsymbol{y}) = -\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta_i \mathrm{d}\theta_j} \ln\left(\pi(\theta)f(\boldsymbol{y}|\theta)\right)\right]_{\theta=\hat{\theta}^p} \tag{11}$$

The result in (10) is sometimes referred to as the Bayesian Central Limit Theorem.

To provide some insight into this result (but not a rigorous proof[4]), consider the univariate θ case and let $l(\theta)$ denote $\ln(\pi(\theta)f(\boldsymbol{y}|\theta))$. The posterior distribution, $p(\theta|\boldsymbol{y})$, is approximated by a 2nd order Taylor series expansion of $l(\theta)$ around $\hat{\theta}^p$:

$$\begin{aligned}
p(\theta|\boldsymbol{y}) &\propto \pi(\theta)f(\boldsymbol{y}|\theta) = \exp\left[\ln(\pi(\theta)f(\boldsymbol{y}|\theta))\right] = \exp(l(\theta)) \\
&\approx \exp\left\{l(\hat{\theta}^p) + \frac{\mathrm{d}}{\mathrm{d}\theta}l(\theta)|_{\theta=\hat{\theta}^p}(\theta - \hat{\theta}^p) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}l_{\theta=\hat{\theta}^p}\left(\theta - \hat{\theta}^p\right)^2\right\} \\
&= \exp\left\{l(\hat{\theta}^p) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}l_{\theta=\hat{\theta}^p}\left(\theta - \hat{\theta}^p\right)^2\right\}, \quad \text{because the 2nd term is 0 at the mode} \\
&= \exp\left\{l(\hat{\theta}^p) - \frac{1}{2}I^p(\boldsymbol{y})\left(\theta - \hat{\theta}^p\right)^2\right\} = \exp\left\{l(\hat{\theta}^p) - \frac{1}{2}\frac{\left(\theta - \hat{\theta}^p\right)^2}{(I^p(\boldsymbol{y}))^{-1}}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\frac{\left(\theta - \hat{\theta}^p\right)^2}{(I^p(\boldsymbol{y}))^{-1}}\right\}
\end{aligned}$$

The last term is the kernel of Normal $\left(\hat{\theta}^p, (I^p(\boldsymbol{y}))^{-1}\right)$.

Remarks.

- The posterior mode, $\hat{\theta}^p$, is called the *Maximum a posteriori* or *MAP* estimator. Note that calculating its value typically requires differentiation. Thus the problem of integration has been replaced by a problem of differentiation, in particular an *optimization* problem.
- The process of solving an integral problem, $\int f(x)\mathrm{d}x$, where $f(x)$ is postive valued, by first rewriting the integrand as $\exp(\log(f(x))$, and then approximating $\log(f(x))$ by a second-order Taylor expansion at the mode of $f(x)$ is known as a *Laplace approximation*. The Laplace approximation method applies to higher dimensional integrals and is the basis of a popular model fitting program called Template Model Builder (TMB).

[4]Note in particular that the demonstration is working with a single parameter θ, not a vector of parameters.

Beta-Binomial example. The data y are Binomial(n,θ) and the prior for θ is Beta(α, β). Then

$$\ln(p(\theta|y)) \equiv l(\theta) \propto \ln\left(\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}\right) = (y+\alpha-1)\ln(\theta) + (n-y+\beta-1)\ln(1-\theta)$$

The mode of the posterior distribution is found by differentiating $l(\theta)$ with respect to θ:

$$\frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta} = \frac{y+\alpha-1}{\theta} - \frac{n-y+\beta-1}{1-\theta}$$

setting the result equal to 0 and solving for θ, yielding the MAP estimate

$$\hat{\theta}^p = \frac{y+\alpha-1}{n+\alpha+\beta-2}$$

Then $I^p(y)$ is:

$$\begin{aligned} I^p(y) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} l(\theta)|_{\hat{\theta}} &= \frac{\alpha+y-1}{\theta^2} + \frac{\beta+n-y-1}{(1-\theta)^2}|_{\hat{\theta}} \\ &= \frac{(\alpha+\beta+n-2)^3}{(\alpha+y-1)(\beta+n-y-1)} \end{aligned}$$

Then

$$p(\theta|y) \approx \text{Normal}\left(\frac{y+\alpha-1}{n+\alpha+\beta-2}, \frac{(\alpha+y-1)(\beta+n-y-1)}{(\alpha+\beta+n-2)^3}\right)$$

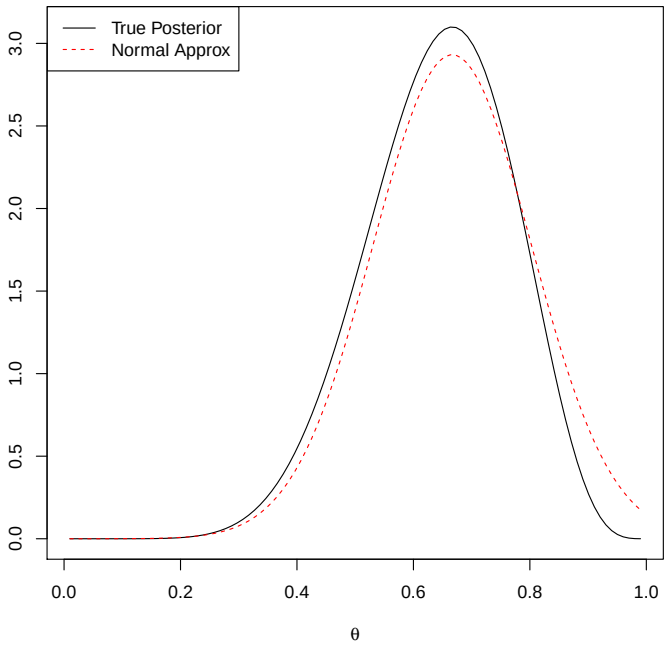
Note that if α=β=1,

$$p(\theta|y) \approx \text{Normal}\left(\frac{y}{n}, \frac{y(n-y)}{n^3}\right) \equiv \text{Normal}\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

where $\hat{p} = y/n$, the familiar sample proportion.

To examine the quality of the approximation, the prior for θ is Beta(1,3) (thus an expected value of 0.25) and n=10 with y=8 successes. The posterior for θ is Beta(9,5) (thus an expected value of 0.643). The mode, $\hat{\theta}$, is 0.667 and the approximate variance, $I^p(y)^{-1}$ is 0.0185. The true posterior density and the normal approximation are plotted in Figure 6. The modes are quite similar but the normal approximation attaches too much probability to $\theta > 0.8$. The differences are apparent in the lower and upper quantiles: e.g., the true 0.025 quantile is 0.386 compared to the normal approximation value of 0.400, and the true 0.975 quantile is 0.861 compared to 0.933 for the normal approximation. The R code is shown in Appendix A.3.

Figure 6: Posterior distribution for θ (Beta(9,5)) and Normal(0.667,0.0185) approximation.



A R Code

A.1 Trapezoid rule demonstration

```
#--- Trapezoid rule example
# true value pexp(3,1/2), 1-exp(-1.5) = 0.7768698
true.integral <- pexp(3,1/2)

exp.f <- function(x,lambda) {
  lambda*exp(-lambda*x)
}
trapezoid.int <- function(lb,ub,f,lambda) {
  width <- ub-lb
  height <- 0.5*(exp.f(x=ub,lambda=lambda)+exp.f(x=lb,lambda=lambda))
  out <- width*height
  return(out)
}
intervals <- seq(0,3,length=5)
#intervals <- seq(0,3,length=21)
result <- 0
verbose <- TRUE
for(i in 1:(length(intervals)-1)) {
  out <- trapezoid.int(lb=intervals[i],ub=intervals[i+1],
                       f=exp.f,lambda=0.5)
  if(verbose) cat("step i value=",out,"\n")
  result <- result + out
}
cat("true value=",true.integral,"trapezoidal rule approx=",result,"\n")
```

A.2 Numerical integration for the Weibull example.

```
# Hurricane gap data
hurricane.gaps <- c( 0.30, 4.61, 5.75, 0.24, 0.09, 0.18, 7.38, 1.20, 2.40, 0.18,
                     0.02, 10.07, 0.23, 0.44, 3.34, 0.06, 0.01, 0.71, 0.06, 0.42)

#----------- Exponential model with Gamma prior  ---------------------------
cat("average gap=",mean(hurricane.gaps),"\n")
# average gap= 1.8845

n <- length(hurricane.gaps)
mle.lambda <- 1/mean(hurricane.gaps)
cat("mle for Exponential lambda=",mle.lambda,"\n") #mle for Exponential lambda= 0.5306447

# Hyperparameters for Gamma prior
prior.alpha <- 0.01
prior.beta  <- 0.01
posterior.alpha <- prior.alpha+n
posterior.beta <- prior.beta + sum(hurricane.gaps)
cat("posterior alpha==",posterior.alpha,"beta=",posterior.beta,"\n")
# posterior alpha== 20.01 beta= 37.7

# posterior predictive density for exponential data given Gamma prior
d.post.pred.exp.gamma <- function(y.New,a,b) {
  out <- (b^a*a)/((b+y.New)^(a+1))
  return(out)
}

#-- plot the empirical, the mle exponential, and posterior predictive densities
y.seq <- seq(0.01,max(hurricane.gaps)*1.2,length=100)
empirical.d <- density(hurricane.gaps)
mle.based.d    <-  dexp(y.seq,rate=mle.lambda)
post.pred.d    <- d.post.pred.exp.gamma(y.seq,a=posterior.alpha,b=posterior.beta)
my.ylim <- range(c(empirical.d$y,mle.based.d,post.pred.d))
plot(density(hurricane.gaps),type="l",xlab="Time between hurricanes",
     ylab="",main="",ylim=my.ylim,xlim=range(y.seq))
lines(y.seq,mle.based.d,col=2,lty=2)
lines(y.seq,post.pred.d,col=3,lty=3)
# title("Empirical, (mle) fitted, post predictive Densities")
legend("topright",legend=c("Empirical","mle-based","Post Predictive"),
       col=1:3,lty=1:3)

#-------- Weibull dist'n: numerical quadrature in R, using integrate --------
library(pracma)
kappa   <- lambda <- 0.01
lb      <- 0.01
ub      <- 10

#--- using Gordon Ross's function
posterior <- function(alpha,y,kappa,lambda,lb=lb,ub=ub,verbose=FALSE) {
  n <- length(y)
  f <- function(alpha) {
    num.vals <- length(alpha)
    results <- numeric(num.vals)
    for(i in 1:num.vals) {
      results[i] <- alpha[i]^(n+kappa-1)*
        exp(-lambda*alpha[i]-sum(y^alpha[i]))*
        prod(y^(alpha[i]-1))
    }
    return(results)
  }

  norm.constant <- integral(f,lb,ub)
  if(verbose) cat("pracma: norm constant=",(lambda^kappa)/(gamma(kappa))*norm.constant,"\n")
```

```
  numerator <- f(alpha)
  out <- numerator/norm.constant
  return(out)
}

alpha.seq <- seq(0,1,length=100)
post.density <- posterior(alpha=alpha.seq,y=hurricane.gaps,
                          kappa=kappa,lambda=lambda,lb=lb,ub=ub,verbose=TRUE)
#    pracma: norm constant= 2.008407e-14
plot(alpha.seq,post.density,type="l",xlab=expression(alpha),ylab="",
     main=expression(paste("Posterior density for Weibull ",alpha)))
```

A.3 Normal approximation example

```
 #-- normal approximation to Beta-Binomial
set.seed(832)
theta <- 0.6
n     <- 10
prior.alpha <- 1
prior.beta  <- 3
y   <- rbinom(n=1,size=n,prob=theta)
cat("#successes=",y,"in n=",n,"trials\n") # #successes= 8 in n= 10 trials
post.alpha <- prior.alpha+y
post.beta  <- prior.beta+n-y

post.mode <- (post.alpha-1)/(post.alpha+post.beta-2)
post.var  <- post.mode*(1-post.mode)/(post.alpha+post.beta-2)
cat("post mode=",post.mode,"post variance=",post.var,"\n")
# post mode= 0.6666667 post variance= 0.01851852

theta.seq <- seq(0.01,0.99,length=100)
true.post.density <- dbeta(theta.seq,post.alpha,post.beta)
approx.post.density <- dnorm(theta.seq,mean=post.mode,sd=sqrt(post.var))
my.ylim <- range(c(true.post.density,approx.post.density))
plot(theta.seq,true.post.density,type="l",xlab=expression(theta),ylab="")
lines(theta.seq,approx.post.density,type="l",col=2,lty=2)
legend("topleft",legend=c("True Posterior","Normal Approx"),col=1:2,lty=1:2)
```

October 23, 2020