Integration: It is performed to evaluate integrals of the type

$$E(g(\theta)|\boldsymbol{x}) = \int_\Theta g(\theta)p(\theta|\boldsymbol{x})d\theta$$

Examples of commonly considered functions g are: A power of θ, for the posterior moments; An indicator function, for posterior probabilities and HPDs; A loss function, in decision theoretic settings; $p(z|\boldsymbol{x}, \theta)$ for predictions.

The most common methods are: Numerical approximations; Direct sampling; Rejection sampling; Markov Chain Monte Carlo Methods.

A Bayesian Central Limit Theorem, allowing normal approximations to the posterior distribution, is given by the following result.

Result: (Berger, p. 224) Let $X_1, \ldots, X_n$ be an i.i.d. sample from the density $p(X|\theta)$ and $\pi(\theta)$ the prior. Then for large n and some regularity conditions, $p(\theta|\boldsymbol{x})$ can be approximated as

1. $N(\mu(\boldsymbol{x}), V(\boldsymbol{x}))$ where $\mu(\boldsymbol{x})$ and $V(\boldsymbol{x})$ are the posterior mean and covariance matrix, or

2. $N(\hat{\theta}, \mathbf{I}(\boldsymbol{x})^{-1})$ where $\hat{\theta}$ is the posterior mode and

$$\mathbf{I}_{ij}(\boldsymbol{x}) = -\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log(p(\boldsymbol{x}|\theta)p(\theta))\right]_{\theta=\hat{\theta}}$$

A Taylor series expansion gives

$$\log p(\theta|\boldsymbol{x}) = \log p(\hat{\theta}|\boldsymbol{x}) + \frac{1}{2}(\theta - \hat{\theta})' \left[\frac{d^2}{d\theta^2} \log p(\theta|\boldsymbol{x})\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \dots$$

The first term in the Taylor expansion is a constant. The second term vanishes and the third term corresponds to the log of a normal density.

Approximation are usually improved by transforming the parameters to have support in $\mathbb{R}$. They are also improved by looking at marginal distributions or even conditional distributions. See Appendix B for the scheme of a proof.

There are a number of cases when a normal approximation would not work:

- Non identified parameters.
- Number of parameters increasing with sample size.
- Aliasing.
- Unbounded likelihoods.
- Improper posterior distributions.
- Prior distribution that exclude the MLE.
- $\hat{\theta}$ on the boundary of the domain.
- Tail of the distribution.

Obtaining the normal approximation for the posterior distribution requires: (a) transforming the parameters; (b) calculating the posterior mode. So we face an optimization problem. For parameter spaces of large dimension, this is can be a difficult problem in itself. For small to moderate dimensionality the method of choice is Newton's method.

Given an initial iterate θ^0, we compute the sequence

$$\theta^t = \theta^{t-1} - \left[\frac{d^2}{d\theta^2}\log p(\theta|\boldsymbol{x})\right]^{-1}_{\theta=\theta^{t-1}} [\nabla \log p(\theta|\boldsymbol{x})]_{\theta=\theta^{t-1}}$$

until convergence. Notice that scheme will provide both the estimate of the maximum and its covariance matrix.

Example: beta-binomial model

Table 5.1 of Albert (2007) contains 20 pairs of observations, one for each of 20 cities. These correspond to (y_j, n_j) being the numbers of cancer deaths and the numbers at risk. The data are available from the `R` package `LearnBayes`.

An initial model is $y_j \sim Bin(n_j, \theta_j)$. From this we can fit a Bayesian model using a beta prior. With a uniform prior, the posterior mean of θ_j would be $(y_j + 1)/(n_j + 2)$.

Note that in a binomial model

$$E(y_j) = n_j\theta_j \quad \text{and} \quad \mathrm{var}(y_j) = n_j\theta_j(1 - \theta_j).$$

So the variance depends on the mean.

Example: beta-binomial model

If we clump the data together we have $\hat{\theta} = 9.92 * 10^{-4}$. This implies that the variance is estimated as 70.92. But an empirical estimate of the variance of $Y = (y_1, \ldots, y_{20})$ yields 141.94, twice as much.

Thus, the data show more dispersion than is obtained by fitting such a binomial model. Write he beta density as

$$p(\theta) = \frac{1}{B(\mu\tau, \tau(1-\mu))} \theta^{\mu\tau - 1}(1-\theta)^{\tau(1-\mu)-1} \quad \theta \in (0,1)$$

Here $E(\theta) = \mu \in (0,1)$, $\tau > 0$ and

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Using the beta prior and the binomial likelihood we obtain

$$p(y_j|\mu,\tau) = \binom{n_j}{y_j} \frac{B(\mu\tau + y_j, \tau(1-\mu) + n_j - y_j)}{B(\mu\tau, \tau(1-\mu))}$$

This is the marginal density of the observations. It depends of two parameters: μ and τ. We can use it as the likelihood for the data. Thus,

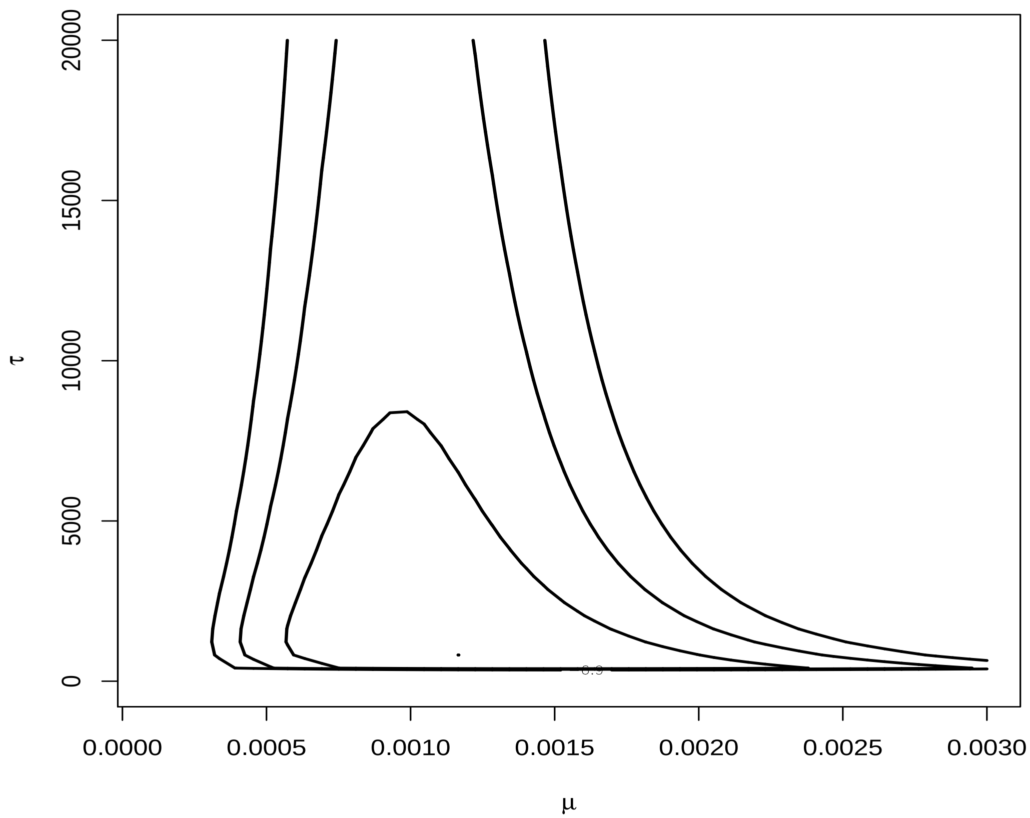
$$\prod_{j=1}^{20} \frac{B(\mu\tau + y_j, \tau(1-\mu) + n_j + y_j)}{B(\mu\tau, \tau(1-\mu))}$$

This is known as the beta-binomial model.

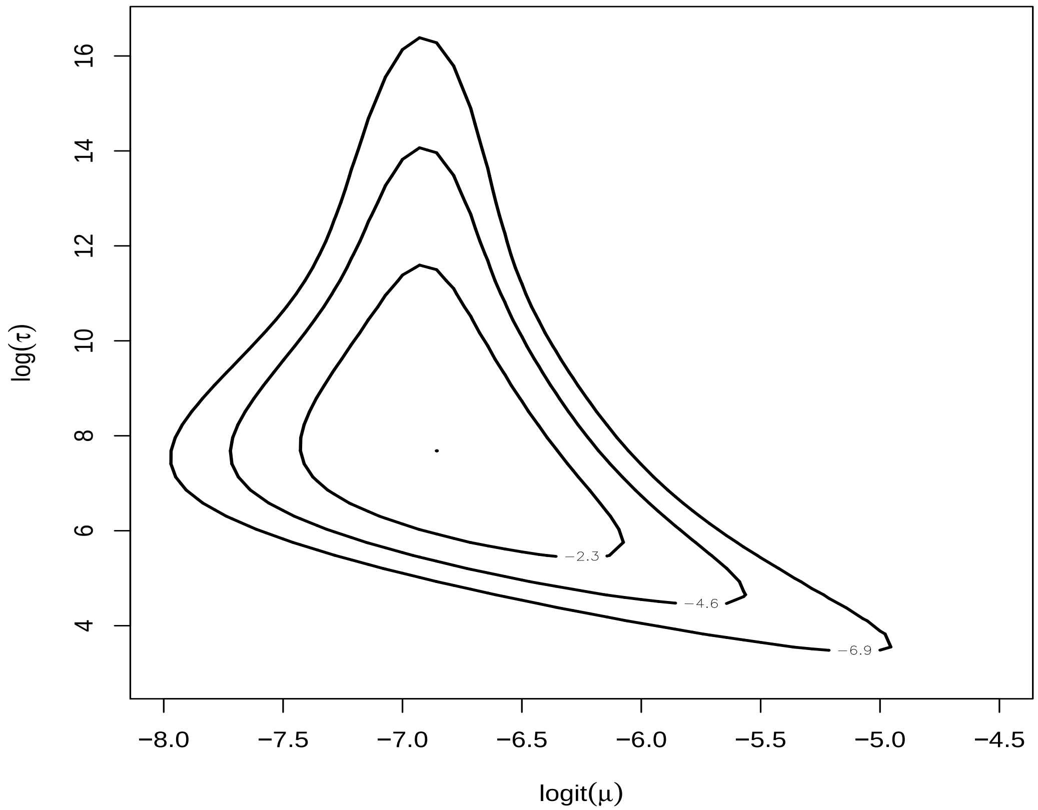
To complete the model we need to specify a prior for the two parameters μ and τ. We use the prior

$$p(\mu,\tau) \propto \frac{1}{\mu(1-\mu)} \frac{1}{(1+\tau)^2} \quad \mu \in (0,1) \quad \tau > 0$$

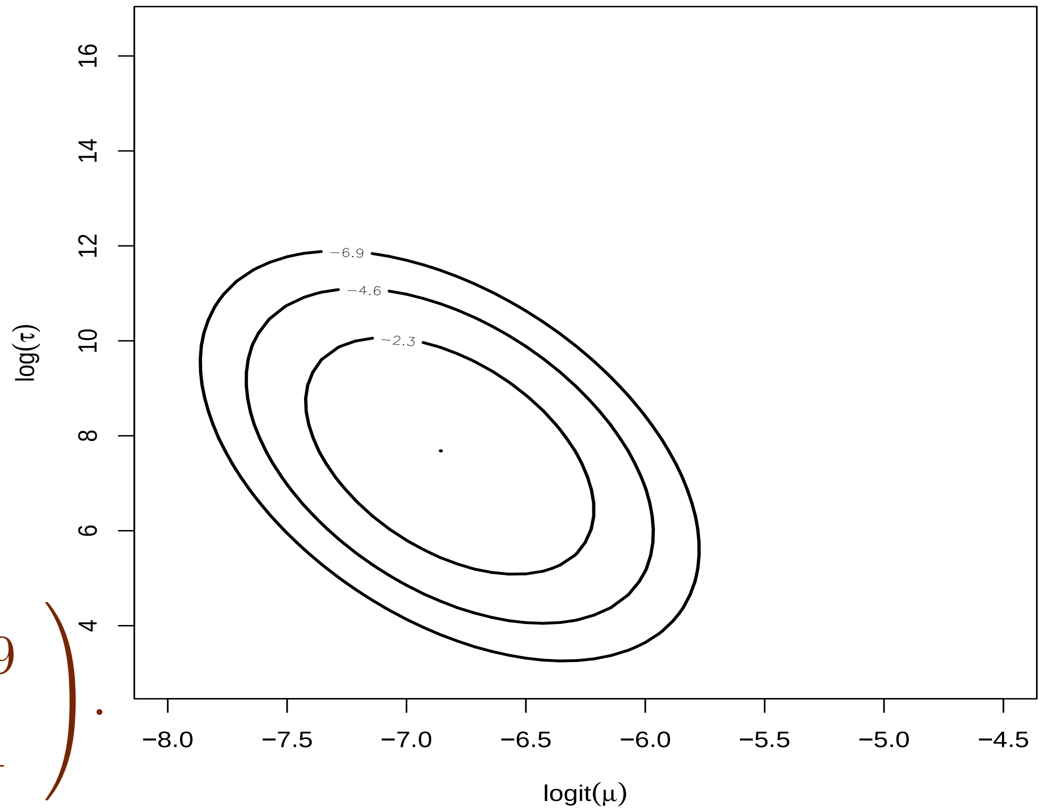
Contour of the log-posterior of μ and τ.



Contour of the log-posterior of $\theta_1 = \log\left(\frac{\mu}{1-\mu}\right)$ and $\theta_2 = \log(\tau)$.



Contour of the normal approximation to the log-posterior of θ_1 and θ_2. The mode is $(-6.82, 7.57)$ and the covariance matrix is
$$\begin{pmatrix} 0.079 & -0.149 \\ -0.149 & 1.3491 \end{pmatrix}.$$



Approximating with Mixtures

If the posterior distribution is multimodal, suppose that $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are the posterior modes, then the posterior can be approximated as

$$\sum_{k=1}^{K} \omega_k N(\hat{\theta}_k, \mathbf{I}_k(\boldsymbol{x}))$$

When the posterior modes are fairly widely separated, we use $\omega_k = q(\hat{\theta}_k|\boldsymbol{x}), k = 1, \ldots, K$, where q is the un-normalized posterior density.

For small samples, a more conservative approximation is given by a mixture of Student densities:

$$\sum_{k=1}^{K} q(\hat{\theta}_k|\boldsymbol{x}) \left[\alpha + (\theta - \hat{\theta}_k)' \mathbf{I}_k(\boldsymbol{x})^{-1} (\theta - \hat{\theta}_k)\right]^{-(d+\alpha)/2}$$

Laplace Approximation

Laplace approximations are useful to approximate expressions of the form

$$\int_{\Theta} g(\theta) \exp\{-nh(\theta)\} d\theta$$

The approximation is given by

$$g(\hat{\theta})(2\pi/n)^{d/2}|\hat{\Sigma}|^{1/2} \exp\{-nh(\hat{\theta})\}, \quad \hat{\Sigma} = \left[\frac{\partial^2}{\partial\theta\partial\theta'} h(\theta)\right]^{-1}_{\theta=\hat{\theta}}$$

and $\hat{\theta}$ is the value that minimizes h.

It can be seen that these approximations are $o(n^{-1})$.

For posterior inference we consider a non-negative $g(\theta)$ and define

$$-nh(\theta) = \log(p(\boldsymbol{x}|\theta)) + \log(p(\theta)) \quad \text{and} \quad nh^*(\theta) = nh(\theta) - \log(g(\theta))$$

Denoting by θ^* the value that minimizes h^* and Σ^* the corresponding matrix of second derivatives, we have that

$$E(g(\theta)) \approx \frac{g(\theta^*)|\Sigma^*|^{1/2} \exp\{-nh^*(\theta^*)\}}{|\hat{\Sigma}|^{1/2} \exp\{-nh(\hat{\theta})\}}$$

Due to cancellations between the numerator and denominator, this approximation is of order $o(n^{-2})$.