Final Project: Laplace Approximation

Thi Thu Ha Vo

December 01, 2024

1 Introduction

In many problems involving posterior distributions, we often prefer conjugate priors because they simplify computations. Conjugate priors ensure that the posterior distribution belongs to the same family as the prior. However, conjugate priors are not always available, and even when they exist, they may not always be suitable for the model at hand. This creates a challenge, as the posterior distribution may not have a standard form, making it difficult to compute. Several numerical integration methods have been developed to address this issue, with Markov Chain Monte Carlo (MCMC) being one of the most widely used. However, MCMC can be computationally expensive. In such cases, Laplace approximation offers an efficient alternative, providing a simple and deterministic method for approximating these posterior distributions.

Notice that posterior distributions are usually bell-shaped and symmetric, which makes them similar to the normal distribution. In most cases, the posterior becomes approximately normal as the number of data points increases, according to the Bernstein-von Mises Theorem. Laplace approximation uses this idea to simplify complex posterior distributions, which are typically smooth and peaked around their maximum. The method works by finding the mode of the posterior, using it as the mean of the normal distribution, and calculating the variance by examining the curvature of the logarithm of the posterior at the mode.

2 Mathematical Intuition

The idea behind the Laplace approximation is straightforward. It uses the second-order Taylor expansion of the log of the probability density function to show that the density function can be approximated by

a normal distribution.

Suppose we have an probability density $p(x)$ that achieves its maximum at x_0. Since the logarithm is an increasing function, $\log p(x)$ also reaches its maximum at x_0. Let $q(x) = \log p(x)$, using the second-order Taylor expansion around the point x_0, we can write,

$$q(x) \approx q(x_0) + q'(x_0)(x - x_0) + \frac{1}{2}q''(x_0)(x - x_0)^2$$

Notice that x_0 is the maximum of the function $q(x) = \log p(x)$, so $q'(x_0) = 0$. Therefore, we can simplify the above expression to,

$$q(x) \approx q(x_0) + \frac{1}{2}q''(x_0)(x - x_0)^2$$

Then, if we take the exponent on both sides, we get,

$$\exp\left(q(x)\right) \approx \exp\left(q(x_0) + \frac{1}{2}q''(x_0)(x - x_0)^2\right)$$

$$p(x) \approx p(x_0) \exp\left(\frac{1}{2}q''(x_0)(x - x_0)^2\right)$$

For simplicity, if we denote $A = -q''(x_0) = -\left.\frac{d^2 \log p(x)}{dx^2}\right|_{x=x_0}$, we obtain,

$$p(x) \approx p(x_0) \exp\left(\frac{-A}{2}(x - x_0)^2\right)$$

Now the exponential part on the right hand-side looks like the kernel of the normal distribution with mean x_0 and variance $\frac{1}{A}$. Since x_0 is a maximum, it means $q''(x_0) < 0$, so $A = -q''(x_0) > 0$. Therefore, the pdf $p(x)$ is approximately $\mathcal{N}(x_0, \frac{1}{A})$.

The Laplace approximation is a straightforward method that can be applied when the log-probability density function (log-pdf) is smooth and has a well-defined maximum. To use this method, we need to identify the point of maximum x_0, and its corresponding curvature of the log-pdf value. We are also interested in determining the normalizing constant, $Z = \int p(x)\, dx$.

$$Z = \int p(x_0) \exp\left(\frac{-A}{2}(x - x_0)^2\right) dx = p(x_0) \int \exp\left(\frac{-A}{2}(x - x_0)^2\right) dx = p(x_0)\sqrt{\frac{2\pi}{A}}$$

Then the normalized version of the approximated density $p(x)$ is,

$$p^*(x) = \frac{1}{Z}p(x) = \left(\frac{A}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-A}{2}(x - x_0)^2\right) \tag{1}$$

Additionally, we can extend the Laplace approximation for a multivariate distribution. Assume that $\mathbf{x} \in \mathbf{R}^d$, denote the matrix of the second derivative $-\log p(\mathbf{x})$ at the maximum $\mathbf{x}_0$ is A (here A is the Hessian matrix of $\log p(\mathbf{x})$ at $\mathbf{x}_0$ but with the minus sign),

$$A_{ij} = -\frac{\partial^2 \log p(\mathbf{x})}{\partial x_i \partial x_j}\Big|_{\mathbf{x}=\mathbf{x}_0}$$

So the Taylor expansion of the $\log p(\mathbf{x})$ at the maximum value $\mathbf{x}_0$ is,

$$\log p(\mathbf{x}) \approx \log p(\mathbf{x}_0) - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T A(\mathbf{x} - \mathbf{x}_0)$$

Now we can approximate $p(\mathbf{x})$ with $\mathcal{N}(\mathbf{x}_0, A^{-1})$ and the normalizing constant can be computed as,

$$\begin{aligned}
Z &= \int \cdots \int_d p(\mathbf{x}_0) \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T A(\mathbf{x} - \mathbf{x}_0)\right) dx_1 \dots dx_d \\
&= p(\mathbf{x}_0) \int \cdots \int_d \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T A(\mathbf{x} - \mathbf{x}_0)\right) dx_1 \dots dx_d \\
&= p(\mathbf{x}_0)\sqrt{(2\pi)^d |A^{-1}|} \\
&= p(\mathbf{x}_0)\sqrt{\frac{(2\pi)^d}{|A|}}
\end{aligned}$$

Finally, the normalized pdf for the multivariate case is,

$$p^*(\mathbf{x}) = \frac{1}{Z}p(\mathbf{x}) = \frac{|A|^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T A(\mathbf{x} - \mathbf{x}_0)\right)$$

3 Applying the Laplace Approximation

In this section, we will apply the Laplace approximation method to compute the posterior distribution of a given statistical model. For our model, we assume that the likelihood follows a Poisson distribution,

$$X|\lambda \sim \text{Poisson}(\lambda)$$

Additionally, we assume an improper prior distribution $P(\lambda) = \frac{1}{\lambda}$. Now, we will attempt to derive the formula for the posterior distribution. The posterior distribution can be expressed as,

$$p(\lambda|x) \propto f(x|\lambda)p(\lambda) = \frac{1}{\lambda}\prod_{i=1}^{n}\exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = \frac{\exp(-n\lambda)}{\lambda}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!} = \text{Const} \times \exp(-n\lambda)\lambda^{\sum_{i=1}^{n}x_i-1} \quad (2)$$

Notice that this posterior distribution is Gamma($\sum_{i=1}^{n} x_i, \frac{1}{n}$) (here the scale is $\frac{1}{n}$). Now we want to use Laplace approximation to approximate this posterior distribution. In order to do that, first we need to take the log of this pdf,

$$\log p(\lambda|x) = -n\lambda + \log\lambda\left(\sum_{i=1}^{n}x_i - 1\right) + \log \text{Const}$$

Then, we need to compute the value of λ such that $p(\lambda|x)$ is maximized. It is definitely the mode of the posterior distribution in this case and we can compute it via the maximum a posterior (MAP),

$$\lambda_{\text{MAP}} = \max_{\lambda} \log p(\lambda|x)$$

We can find MAP by deriving $\log p(\lambda|x)$ and setting it to zero, and finally solving the equation for λ,

$$\frac{d\log p(\lambda|x)}{d\lambda} = -n + \frac{\sum_{i=1}^{n}x_i - 1}{\lambda} = 0$$

$$\lambda_{\text{MAP}} = \frac{\sum_{i=1}^{n}x_i - 1}{n}$$

Now, compute the second order derivative of $\log p(\lambda|x)$ in order to obtain the variance of the normal approximation,

$$A = -\left.\frac{d^2\log p(\lambda|x)}{d\lambda^2}\right|_{\lambda=\lambda_{\text{MAP}}} = \frac{\sum_{i=1}^{n}x_i - 1}{\lambda_{\text{MAP}}^2} = \frac{n^2}{\sum_{i=1}^{n}x_i - 1}$$

We can also refer to (1) in order to obtain the approximated distribution of $p(\lambda|x)$ but with normalizing constant. Finally, the Laplace approximation of the posterior distribution is $\mathcal{N}(\frac{\sum_{i=1}^{n}x_i-1}{n}, \frac{\sum_{i=1}^{n}x_i-1}{n^2})$.

In order to estimate the quality of the approximation, we are interested in drawing the true distribution of the posterior distribution (2) and the approximated density in the same graph. First, suppose we have only one observation from the Poisson(5). Here are the code in R that can be used to generate and plot the result,

```
# Set the seed in order to obtain the same random values
set.seed(42)

# Initialize the number of sampling
n <- 1

# Sample from the poisson distribution with prior value of lambda
lambda_0 <- 5
X <- rpois(n, lambda_0)

# Set the range of lambda
lambda <- seq(0, 20, 0.1)

# Get the true pdf of posterior, which is gamma distribution
gamma_pdf <- dgamma(lambda, shape = sum(X), scale = 1/n)

# Mean and standard deviation obtained from Laplace approximation
mean <- (sum(X) - 1) / n
sd <- sqrt((sum(X) - 1) / n^2)

# Compute the density of normal distribution, which is Laplace approximation
laplace_approx <- dnorm(lambda, mean = mean, sd = sd)

# Plot the result
plot(lambda, laplace_approx, type = "l", lwd = 2, xlab = "lambda", ylab = "Posterior density")
lines(lambda, gamma_pdf, col = "red", lwd = 2)
legend("topright", legend = c("Laplace approximation", "The true pdf"),
       col = c("black", "red"), lty = 1, lwd = 2, cex = 0.5)
```

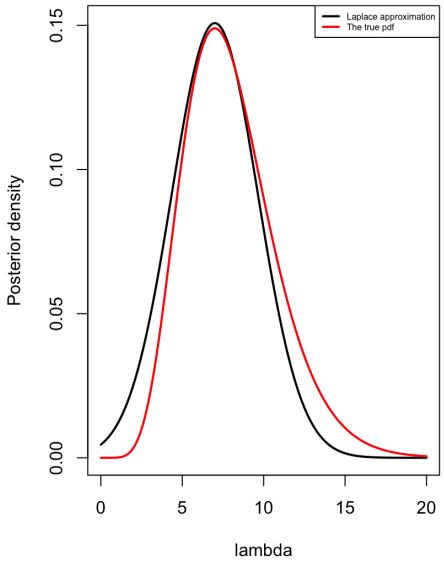

Figure 1: The Laplace approximation of posterior density with n = 1

Here, the solid black curve is the Laplace approximation while the red one is the true distribution, which is Gamma distribution. From the plot, we can observe that the Laplace approximation works quite well in the neighborhood of the mode. However, if we move far away from the mode, the tail of Gamma is heavier on the right. This can be explained since we are working with only a single observation. Now consider the graph when the number of observation $n = 30$.

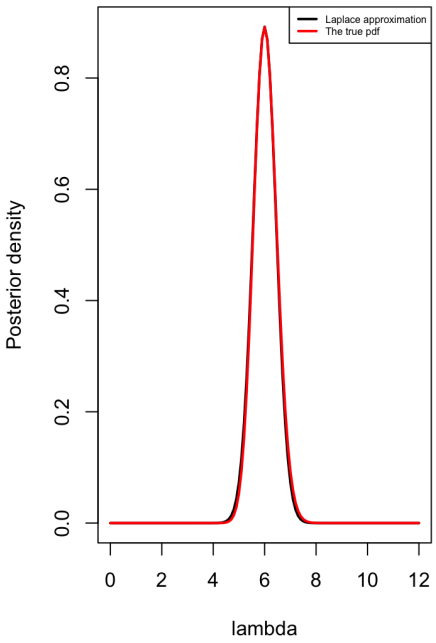

Figure 2: The Laplace approximation of posterior density with n = 30

It looks better now, the Laplace approximation almost coincides with the true pdf (2). It means that we would expect the approximation to improve as the sample size increases. This is somehow proved by the Bernstein-von Mises Theorem which is mentioned below.

Theorem 1 *Let*

$$B \mapsto \Pi_n(B \mid X_1, \ldots, X_n)$$

be the posterior distribution of a parameter θ *based on observations* $X_1, \ldots, X_n$ *sampled from a density* p_θ *and a prior measure* Π *on the parameter set* $\Theta \subset \mathbf{R}^m$*. The Bernstein–von Mises theorem asserts that if* $X_1, \ldots, X_n$ *is a random sample from the density* p_{θ_0}*, the model* $\theta \mapsto p_\theta$ *is appropriately smooth and identifiable, and the prior puts positive mass around the parameter* θ_0*, then*

$$\sup_B \left| \Pi_n(B \mid X_1, \ldots, X_n) - \mathcal{N}_{\left(\hat{\theta}_n, \frac{1}{n} I_{\theta_0}^{-1}\right)}(B) \right| \to 0,$$

where $\mathcal{N}(x, \Sigma)$ *denotes the (multivariate) normal distribution centered at* x *with covariance matrix* Σ*,* $\hat{\theta}_n$

may be any efficient estimator sequence of the parameter, and I_θ is the Fisher information matrix of the model at θ. Under regularity conditions, ones can define $\hat{\theta}_n$ as the maximum likelihood estimator of θ.

This theorem somehow guarantees that the Laplace approximation is actually good if the model satisfies the conditions of the theorem and the value of n is large. In this case, the posterior distribution should look like a bell curve, and a normal distribution is actually a good approximation.

4 Conclusion

Calculating the exact posterior probability distribution is crucial when we need to estimate quantiles or draw samples for prediction. Unfortunately, many posteriors do not have a simple form, especially when the prior distribution is not conjugate. While Markov Chain Monte Carlo (MCMC) is a powerful numerical method to handle these complex posteriors, it can be computationally expensive and time-consuming. To address this, we can use the Laplace approximation as a faster and simpler alternative. This method approximates the complex posterior distribution with a normal distribution, which works well when the posterior is smooth, relative symmetric and has a clear peak. The Bernstein-von Mises Theorem further supports this approach by showing that, under most conditions, the posterior distribution becomes increasingly normal as the amount of data grows.

References

[1] J. A. Hartigan. *Asymptotic Normality of Posterior Distributions.* Wiley, 1983.

[2] Robert E. Kass, Luke Tierney, and Joseph B. Kadane. A note on asymptotic optimality of bayesian estimators. *Journal of the American Statistical Association*, 85(410):1202–1213, 1990.

[3] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, Cambridge, UK, 2003.

[4] A.W. van der Vaart. *10.2 Bernstein–von Mises Theorem.* Publisher Name, City, 1998.

[4, 3, 1, 2]