

8. Large samples



Le Cam

Le Cam (1953) was the first to formally demonstrate the asymptotic normality of the posterior distribution.

Objective

Illustrate the limiting properties of Bayesian distributions.

Recommended Reading

- Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*, Section 5.3.
- Gelman et al (2003), Chapter 4, Sections 4.1 – 4.3 and Appendix B.

If the sample size is very large, it seems obvious that the prior parameter values will have very little influence.

Example 60

$X|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$. Suppose that we use a conjugate prior distribution. Then, for example:

$$E[\mu|\mathbf{x}] = \frac{cm + n\bar{x}}{c + n} \rightarrow \bar{x}$$

when $n \rightarrow \infty$.

In fact, we should usually expect that the properties of Bayesian posterior distributions will be similar to those of maximum likelihood estimators in the limit. The following results illustrate this.

Asymptotic results when Θ is discrete

The following theorem demonstrates that, in the discrete case, as long as the prior probability of the true value, θ_t , is positive, then the posterior probability density of θ converges to a point mass at θ_t .

Theorem 30

Let $X|\theta \sim f(\cdot|\theta)$ where the parameter space $\Theta = \{\theta_1, \theta_2, \dots\}$ is countable. Suppose that $\theta_t \in \Theta$ is the true value of θ .

Suppose that the prior distribution is $P(\theta)$ where $P(\theta_i) > 0 \forall i$ and we assume that

$$\int f(x|\theta_t) \log \frac{f(x|\theta_t)}{f(x|\theta_i)} dx > 0 \quad \forall i \neq t. \quad \text{Then}$$

$$\lim_{n \rightarrow \infty} P(\theta_t|\mathbf{x}) = 1 \quad \text{and} \quad \lim_{n \rightarrow \infty} P(\theta_i|\mathbf{x}) = 0 \quad \forall i \neq t.$$

An interesting extension of this result is that if $\theta_t \notin \Theta$, then the posterior distribution converges to a point mass at the point that gives a parametric model closest (in the sense of Kullback-Liebler distance) to the true model.

Proof Let $\mathbf{x} = (x_1, \dots, x_n)$.

$$\begin{aligned} P(\boldsymbol{\theta}_i | \mathbf{x}) &= \frac{P(\boldsymbol{\theta}_i) f(\mathbf{x} | \boldsymbol{\theta}_i)}{\sum_i P(\boldsymbol{\theta}_i) f(\mathbf{x} | \boldsymbol{\theta}_i)} \\ &= \frac{P(\boldsymbol{\theta}_i) f(\mathbf{x} | \boldsymbol{\theta}_i) / f(\mathbf{x} | \boldsymbol{\theta}_t)}{\sum_i P(\boldsymbol{\theta}_i) f(\mathbf{x} | \boldsymbol{\theta}_i) / f(\mathbf{x} | \boldsymbol{\theta}_t)} \\ &= \frac{\exp(\log P(\boldsymbol{\theta}_i) + S_i)}{\sum_i \exp(\log P(\boldsymbol{\theta}_i) + S_i)} \quad \text{where } S_i = \sum_{j=1}^n \log \frac{f(x_j | \boldsymbol{\theta}_i)}{f(x_j | \boldsymbol{\theta}_t)}. \end{aligned}$$

Conditional on $\boldsymbol{\theta}_t$, S_i is the sum of n i.i.d. random quantities and therefore, by the strong law of large numbers, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} S_i = \int f(x | \boldsymbol{\theta}_t) \log \frac{f(x | \boldsymbol{\theta}_i)}{f(x | \boldsymbol{\theta}_t)} dx.$$

This quantity is negative if $i \neq t$ and zero if $i = t$. Therefore, when $n \rightarrow \infty$, $S_t \rightarrow 0$ y $S_i \rightarrow -\infty$ if $i \neq t$, which proves the theorem. ■

The continuous case

The previous arguments cannot be used in the continuous case, as now, the probability at any particular value of θ is 0. Instead, we now define θ_t to be the value of θ that maximizes the Kullback-Liebler information

$$H(\theta) = \int \log \frac{f_t(x)}{f(x|\theta)} f_t(x) dx$$

of the distribution $f(\cdot|\theta)$ with respect to the true distribution of X , say $f_t(\cdot)$. Now we can demonstrate the following theorem.

Theorem 31

If θ is defined on a compact set and A is a neighbourhood of θ_t with non-zero prior probability, then

$$P(\theta_t \in A|\mathbf{x}) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

Proof See Gelman et al (2003). ■

Convergence to normality

Theorem 32

Under certain regularity conditions, the posterior distribution of $\boldsymbol{\theta}$ tends to a normal distribution with mean $\boldsymbol{\theta}_t$ and variance $nJ(\boldsymbol{\theta}_t)^{-1}$ where $J(\boldsymbol{\theta})$ is the Fisher information.

Proof Suppose that θ is univariate and let $\hat{\theta}$ be the posterior mode. Then a Taylor expansion of $\log p(\theta|\mathbf{x})$ around the mode is

$$\log p(\theta|\mathbf{x}) = \log p(\hat{\theta}|\mathbf{x}) + \frac{1}{2}(\theta - \hat{\theta})^2 \frac{d^2}{d\theta^2} \log p(\theta|\mathbf{x})|_{\theta=\hat{\theta}} + \dots$$


The first term in this expression is constant and the second term is

$$\begin{aligned} (\theta - \hat{\theta})^2 \frac{d^2}{d\theta^2} \log p(\theta|\mathbf{x})|_{\theta=\hat{\theta}} &= (\theta - \hat{\theta})^2 \frac{d^2}{d\theta^2} \log \frac{p(\theta)f(\mathbf{x}|\theta)}{f(\mathbf{x})} \Big|_{\theta=\hat{\theta}} \\ &= (\theta - \hat{\theta})^2 \left(\frac{d^2}{d\theta^2} \log p(\hat{\theta}) + \sum_{i=1}^n \frac{d^2}{d\theta^2} \log f(x_i|\theta) \Big|_{\theta=\hat{\theta}} \right). \end{aligned}$$

Now the first bracketed term, $\frac{d^2}{d\theta^2} \log p(\hat{\theta})$ is constant, whereas the second term (thinking of the x_i as variables) is the sum of n i.i.d. random variables with negative mean.

As the posterior mode is a consistent estimator, it follows that if $f_t(x) = f(x|\theta_t)$ is the true distribution, then the mean is $-J(\theta_t)$. Otherwise, the mean is $E_{f_t} \left[\frac{d^2}{d\theta^2} \log f(x|\theta) \right]$ evaluated at $\theta = \theta_t$ which is also negative by definition of θ_t .

Therefore, the coefficient of the second term in the Taylor series increases with order n . Similarly, the higher order terms can also be shown to increase no faster than order n .

Letting $n \rightarrow \infty$, we thus have that the importance of the higher order terms of the Taylor expansion fades relative to the quadratic term as the mass of the posterior concentrates around θ_t and the normal approximation grows in precision. 

Here, we used the mode as a consistent estimator of θ_t . We could equally use the mean or the classical MLE.

Theorem 33

Let $X_i|\boldsymbol{\theta} \sim f(\cdot|\boldsymbol{\theta})$ with prior distribution $f(\boldsymbol{\theta})$. Given data \mathbf{x} , when $n \rightarrow \infty$,

1. $\boldsymbol{\theta}|\mathbf{x} \approx \mathcal{N}(E[\boldsymbol{\theta}|\mathbf{x}], V[\boldsymbol{\theta}|\mathbf{x}])$, supposing that the mean and variance of $\boldsymbol{\theta}$ exist,
2. $\boldsymbol{\theta}|\mathbf{x} \approx \mathcal{N}(\hat{\boldsymbol{\theta}}, I(\hat{\boldsymbol{\theta}})^{-1})$ where $\hat{\boldsymbol{\theta}}$ is the mode. $I(\boldsymbol{\theta})$ is *the observed information*

$$I(\boldsymbol{\theta}) = -\frac{d^2}{d\boldsymbol{\theta}^2} \log(f(\boldsymbol{\theta}|\mathbf{x})).$$

3. $\boldsymbol{\theta}|\mathbf{x} \approx \mathcal{N}(\hat{\boldsymbol{\theta}}, I^*(\hat{\boldsymbol{\theta}})^{-1})$ where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, supposing this exists and

$$I^*(\boldsymbol{\theta}) = -\frac{d^2}{d\boldsymbol{\theta}^2} \log(f(\mathbf{x}|\boldsymbol{\theta}))$$

4. $\boldsymbol{\theta}|\mathbf{x} \approx \mathcal{N}(\hat{\boldsymbol{\theta}}, I^{**}(\hat{\boldsymbol{\theta}})^{-1})$ where $I^{**}(\boldsymbol{\theta}) = -nE_X \left[\frac{d^2}{d\boldsymbol{\theta}^2} \log(f(X|\boldsymbol{\theta})) \right]$.

Proof See Bernardo and Smith (1994) or Gelman et al (2003). ■

Approximating a beta posterior distribution

Normally, the first approximation will be better than the second and so on. In many cases, the posterior mean and variance are difficult to evaluate but it is much easier to calculate the mode and observed information.

Example 61

Let $X|\theta \sim \mathcal{BI}(n, \theta)$ and $\theta \sim \mathcal{B}(\alpha, \beta)$. Then, $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + x, \beta + n - x)$. If n is large, we can approximate the posterior distribution of θ . Here we compare the four approximations given earlier.

Firstly, approximating with a normal using the beta mean and variance we have

$$\theta|\mathbf{x} \approx \mathcal{N} \left(\frac{\alpha + x}{\alpha + \beta + n}, \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \right).$$

Secondly, we can calculate the mode of the beta distribution.

$$\begin{aligned}\log p(\theta|\mathbf{x}) &= c + (\alpha + x - 1) \log(\theta) + (\beta + n - x - 1) \log(1 - \theta) \\ \frac{d}{d\theta} &= \frac{\alpha + x - 1}{\theta} - \frac{\beta + n - x - 1}{1 - \theta} \\ \hat{\theta} &= \frac{\alpha + x - 1}{\alpha + \beta + n - 2}\end{aligned}$$

is the mode. Also:

$$\begin{aligned}\frac{d^2}{d\theta^2} \log(f(\theta|\mathbf{x})) &= -\frac{\alpha + x - 1}{\theta^2} - \frac{\beta + n - x - 1}{(1 - \theta)^2} \\ I(\hat{\theta}) &= \frac{\alpha + x - 1}{\hat{\theta}^2} + \frac{\beta + n - x - 1}{(1 - \hat{\theta})^2} \\ &= \frac{(\alpha + \beta + n - 2)^3}{(\alpha + x - 1)(\beta + n - x - 1)} \quad \text{and therefore} \\ \theta|\mathbf{x} &\approx \mathcal{N}\left(\frac{\alpha + x - 1}{\alpha + \beta + n - 2}, \frac{(\alpha + x - 1)(\beta + n - x - 1)}{(\alpha + \beta + n - 2)^3}\right)\end{aligned}$$

Thirdly, note that the MLE is $\hat{\theta} = \frac{x}{n}$ and $I^*(\hat{\theta}) = \frac{n^3}{x(n-x)}$. Thus

$$\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{x}{n}, \frac{x(n-x)}{n^3}\right)$$

and finally, note that

$$\begin{aligned} -\frac{d^2}{d\theta^2} \log(f(X|\theta)) &= \frac{X}{\theta^2} + \frac{n-X}{(1-\theta)^2} \\ I^{**}(\theta) &= \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} \\ &= \frac{n}{\theta} + \frac{n}{(1-\theta)} \\ I^{**}(\hat{\theta}) &= \frac{n^3}{x(n-x)} \end{aligned}$$

and we have the same approximation.

When $n \gg \alpha + \beta$ the approximations will give similar results but in small samples, the results could be quite different.

Example 62

Suppose that $\alpha = \beta = 2$ and $x = 20$, $n = 30$. We shall use the different approximations and estimate $P(\theta > 0.5|\text{data})$.

We have $\theta|\mathbf{x} \sim \mathcal{B}(22, 12)$ and using Matlab, we can show that $P(\theta > 0.5|\mathbf{x}) = 0.95993$ is the exact probability.

Now, using the first approximation, we have $\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{22}{34}, \frac{22 \times 12}{34^2 \times 35}\right) = \mathcal{N}(0.64706, 0.006525)$ and we find $P(\theta > 0.5|\mathbf{x}) \approx P(Z > -1.8206) = 0.9660$.

Approximating using the mode, we have $\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{21}{32}, \frac{21 \times 11}{32^3}\right) = \mathcal{N}(0.65625, 0.00705)$ and thus $P(\theta > 0.5|\mathbf{x}) \approx P(Z > -1.8610) = 0.9686$.

Using the classical approximations, $\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{20}{30}, \frac{20 \times 10}{30^3}\right) = \mathcal{N}(0.66667, 0.00741)$ and thus, $P(\theta > 0.5|\mathbf{x}) \approx P(Z > -1.9365) = 0.9735$.

When the theorem cannot be applied

In some situations we cannot apply the results of the theorem. For example

- If θ_t is a boundary point of Θ ,
- If the prior mass density around θ_t is 0,
- If the posterior density is improper,
- If the model is not identifiable.

Example 63

Suppose that we have the model

$$f(x|\theta_1, \dots, \theta_k) = w_1 f(x|\theta_1) + \dots + w_k f(x|\theta_k)$$

i.e. a mixture of k densities from the same family.

Then given the data, the likelihood will be multimodal because the model is not identifiable. Thus, we need to restrict the parameter space Θ in order to identify the model.

One possibility is to order the parameters, $\theta_1 < \dots < \theta_k$, to identify the model.

See Gelman et al (2003) for more examples.

The Laplace approximation

Tierney and Kadane (1996) introduced this generalization of the normal approximation in order for the problem of estimating posterior moments.

Assume that we wish to estimate

$$E[g(\theta)|\mathbf{x}] = \frac{\int g(\theta)l(\theta|\mathbf{x})p(\theta) d\theta}{\int l(\theta|\mathbf{x})p(\theta) d\theta}$$

where it is supposed that $g(\cdot)$ is non negative.

Then we can write this expectation as

$$E[g(\theta)|\mathbf{x}] = \frac{\int \exp(-nh^*(\theta)) d\theta}{\int \exp(-nh(\theta)) d\theta} \quad \text{where}$$

$$-nh(\theta) = \log p(\theta) + \log l(\theta|\mathbf{x})$$

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

Then, we use the Taylor expansion of h (h^*) about the mode $\hat{\theta}$ ($\hat{\theta}^*$).

$$-h(\hat{\theta}) = \max_{\theta}(-h(\theta)) \quad -h^*(\hat{\theta}^*) = \max_{\theta}(-h^*(\theta))$$

and retain the quadratic terms. We estimate the denominator by

$$\int \exp(-nh(\theta)) d\theta \approx \sqrt{2\pi\sigma} n^{-1/2} \exp(-nh(\hat{\theta}))$$

where $\sigma = \left(\frac{d^2}{d\theta^2} h(\theta) \Big|_{\theta=\hat{\theta}} \right)^{-1/2}$ and similarly for the numerator.

This leads to the following estimate:

$$E[g(\theta|\mathbf{x})] \approx \left(\frac{\sigma^*}{\sigma} \right) \frac{g(\hat{\theta}) f(\hat{\theta}) l(\hat{\theta}|\mathbf{x})}{f(\hat{\theta}) l(\hat{\theta}|\mathbf{x})} \quad \text{where}$$

$$\sigma^* = \left(\frac{d^2}{d\theta^2} h^*(\theta) \Big|_{\theta=\hat{\theta}^*} \right)^{-1/2} .$$

Example 64

Return to Example 61. We have $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + x, \beta + n - x)$.

Without loss of generality, suppose that $0 \leq \alpha, \beta < 1$. If not, simply transform, $x \rightarrow x + [\alpha]$ and $n \rightarrow n + [\alpha] + [\beta]$.

Writing the beta density as above,

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \theta^{\alpha+x-1} (1-\theta)^{\beta+n-x-1} \\ &\propto \exp(-nh(\theta)) \quad \text{where} \\ h(\theta) &= -\frac{1}{n} ((\alpha + x - 1) \log \theta + (\beta + n - x - 1) \log(1 - \theta)) \end{aligned}$$

We can thus show (Exercise) that the Laplace estimate of the posterior mean will be

$$\frac{(\alpha + x)^{\alpha+x+1/2} (\alpha + \beta + n - 2)^{\alpha+\beta+n-1/2}}{(\alpha + x - 1)^{\alpha+x-1/2} (\alpha + \beta + n - 1)^{\alpha+\beta+n+1/2}}.$$

For example, if $\theta|\mathbf{x} \sim \mathcal{B}(8, 12)$, setting $\alpha = \beta = 0$ and $n = 20$ the Laplace estimate of the posterior mean is

$$E[\theta|\mathbf{x}] \approx \frac{8^{8.5} 18^{19.5}}{77.5 19^{20.5}} \approx .3994$$

The true value of the mean is $8/20 = 0.4$ and approximating the mean by the mode as in approximation 2 of the theorem, we have $E[\theta|\mathbf{x}] \approx 7/18 = .3889$. The Laplace approximation is somewhat better.

Approximating the Bayes factor

Consider the case of two composite hypotheses H_0 and H_1 . The Bayes factor is

$$B = \frac{\int f(\mathbf{x}|\boldsymbol{\theta}_0, H_0)f(\boldsymbol{\theta}_0|H_0) d\boldsymbol{\theta}_0}{\int f(\mathbf{x}|\boldsymbol{\theta}_1, H_1)f(\boldsymbol{\theta}_1|H_1) d\boldsymbol{\theta}_1}$$

and both numerator and denominator are positive functions. Therefore we can apply the Laplace approximation. See Kass and Raftery (1995) for details.

Properties and problems with the Laplace approximation

- The Laplace approximation is $O(1/n^2)$.
- If $\Theta \neq R$, then the model can be reparameterized in order to improve the approximation.
- The Laplace approximation can be extended to the multivariate situation.
- In order to implement the Laplace approximation, we need to be able to calculate the MLE of θ .

References

Gelman, A., Carlin, J.B., Stern, H. and Rubin, D.B. (2003). *Bayesian Data Analysis* (2'nd ed.), Chapman and Hall.

Kass, R. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Le Cam, L. (1953). On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates. *University of California Publications in Statistics*, **1**, 277–328.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.