

Large Sample Properties of Bayesian Inference

A fundamental result in Bayesian analysis is that, under general conditions, the posterior distribution is approximately multivariate normal for large n .

Let $\hat{\boldsymbol{\theta}}$ be the mode of the posterior distribution, let $\pi^*(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, and define the matrix $H(\boldsymbol{\theta})$ as follows:

$$H(\boldsymbol{\theta})_{ij} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log \pi^*(\boldsymbol{\theta}|\mathbf{y}).$$

The matrix $-H(\boldsymbol{\theta})$ is called the *Hessian*.

Under certain regularity conditions (to be discussed later) the posterior distribution is approximately $N(\hat{\boldsymbol{\theta}}, H(\hat{\boldsymbol{\theta}})^{-1})$ when n is sufficiently large.

The essential part of proving this asymptotic normality result is a Taylor series expansion. We expand $\log \pi^*(\boldsymbol{\theta}|\mathbf{y})$ in a Taylor series about $\hat{\boldsymbol{\theta}}$.

$$\begin{aligned} \log \pi^*(\boldsymbol{\theta}|\mathbf{y}) &= \log \pi^*(\hat{\boldsymbol{\theta}}|\mathbf{y}) \\ &\quad + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi^*(\boldsymbol{\theta}|\mathbf{y}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &\quad - \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T H(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &\quad + R_n, \end{aligned}$$

where R_n is negligible relative to the other terms as $n \rightarrow \infty$.

The i th element of $\frac{\partial}{\partial \boldsymbol{\theta}} \log \pi^*(\boldsymbol{\theta}|\mathbf{y})$ is

$$\frac{1}{\pi^*(\boldsymbol{\theta}|\mathbf{y})} \cdot \frac{\partial}{\partial \theta_i} \pi^*(\boldsymbol{\theta}|\mathbf{y}).$$

The last quantity evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ is 0 since $\hat{\boldsymbol{\theta}}$ maximizes π^* . It follows that

$$\begin{aligned}\pi^*(\boldsymbol{\theta}|\mathbf{y}) &= \pi^*(\hat{\boldsymbol{\theta}}|\mathbf{y}) \\ &\times \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T H(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \\ &\times \exp(R_n).\end{aligned}$$

So, for large n , the posterior is approximately proportional to a multivariate normal density with mean $\hat{\boldsymbol{\theta}}$ and covariance $H(\hat{\boldsymbol{\theta}})^{-1}$.

An obvious regularity condition required for the previous result is that the second partial derivatives of $\pi^*(\boldsymbol{\theta}|\mathbf{y})$ exist and be continuous throughout a neighborhood of the true parameter vector $\boldsymbol{\theta}_0$.

For a complete set of sufficient conditions, see LeCam and Yang (1990).

Example 15 *Asymptotic normality in binomial experiment*

Suppose Y_1, \dots, Y_n are i.i.d. Bernoulli(θ), and suppose the prior is Beta(α, β). If $y = \sum_{i=1}^n y_i$, then

$$\pi^*(\theta|\mathbf{y}) = C\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}.$$

$$\frac{\partial}{\partial \theta} \log \pi^*(\theta|\mathbf{y}) = \frac{(y + \alpha - 1)}{\theta} - \frac{n - y + \beta - 1}{1 - \theta}$$

$$-\frac{\partial^2}{\partial \theta^2} \log \pi^*(\theta|\mathbf{y}) = \frac{(y + \alpha - 1)}{\theta^2} + \frac{(n - y + \beta - 1)}{(1 - \theta)^2}$$

The mode of the posterior is

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

Verify that the Hessian evaluated at $\hat{\theta}$ is

$$H(\hat{\theta}) = \frac{(n + \alpha + \beta - 2)}{\hat{\theta}(1 - \hat{\theta})}.$$

So, the posterior is approximately normal with mean $\hat{\theta}$ and variance $\hat{\theta}(1 - \hat{\theta})/(n + \alpha + \beta - 2)$.

Recall that the MLE of θ is $\hat{\theta}_n = y/n$. We have seen previously that

$$\hat{\theta} = \hat{\theta}_n + O(1/n),$$

where $O(1/n)$ denotes a quantity that converges to 0 at the rate n^{-1} .

The asymptotic posterior variance is

$$\frac{\hat{\theta}(1 - \hat{\theta})}{(n + \alpha + \beta - 2)} = \frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n} + O_p(n^{-2}),$$

where $O_p(n^{-2})$ is a random variable that converges to 0 at the rate n^{-2} .

Note that $I(\theta) = [\theta(1 - \theta)/n]^{-1}$, and hence we may conclude that the posterior is approximately $N(\hat{\theta}_n, I(\hat{\theta}_n)^{-1})$.

The last result mentioned in Example 15 is not particular to the binomial model.

Under general conditions, the posterior is (for large n) approximately multivariate normal with mean vector equal to the MLE $\hat{\theta}_n$ and covariance matrix $I(\hat{\theta}_n)^{-1}$.

An interesting aspect of this result is that the approximation to the posterior depends in no way on the prior.

The validity of the result rests on the fact that, for large n , the likelihood dominates the prior, in the sense that the likelihood is sharply peaked relative to the prior.

Another interesting aspect of the result on p. 158N is how it parallels the asymptotic normality of MLEs.

Frequentists make heavy use of the following result:

Under general conditions, the sampling distribution of the MLE $\hat{\theta}_n$ is approximately $N(\theta_0, I(\theta_0)^{-1})$ for large n .

Therefore, as we have seen in other cases, the frequency distribution of a statistic for a given parameter value has the same form as the distribution of a parameter conditional on the data.

Example 16 *Logistic regression*

Suppose $\mathbf{Y} = (Y_1, \dots, Y_n)$, where Y_1, \dots, Y_n are independent with $Y_i \sim \text{Bin}(1, \theta_i)$, $i = 1, \dots, n$.

The elements of $\boldsymbol{\theta}$ are modeled as functions of the covariates $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$, via the *logistic regression* function:

$$\theta_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters.

The likelihood function is

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}) &= \prod_{i=1}^n \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \\ &= \prod_{i=1}^n \left[\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right]^{y_i} \\ &\quad \times \left[\frac{1}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right]^{1-y_i} \end{aligned}$$

Hence,

$$f(\mathbf{y}|\boldsymbol{\beta}) = \exp \left\{ \sum_{i=1}^n \left[y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log \left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right) \right] \right\}.$$

Suppose that we use a uniform (improper) prior for $\boldsymbol{\beta}$. Then

$$\pi^*(\boldsymbol{\beta}|\mathbf{y}) \equiv f(\mathbf{y}|\boldsymbol{\beta}),$$

and the posterior mode is the MLE $\hat{\boldsymbol{\beta}}$. We now compute the Hessian.

$$\frac{\partial}{\partial \beta_j} \log \pi^*(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^n \left(y_i - \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) x_{ij}$$

$$-\frac{\partial^2}{\partial \beta_j \partial \beta_k} \log \pi^*(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^n \left[\frac{x_{ij} x_{ik} e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{\left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right)^2} \right]$$

Define \mathbf{X} to be the $n \times p$ matrix of covariates and $\mathbf{S}^T = (y_1 - \theta_1, \dots, y_n - \theta_n)$. Then we may write

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log \pi^*(\boldsymbol{\beta}|\mathbf{y}) = \mathbf{X}^T \mathbf{S}$$

and

$$\begin{aligned} H(\boldsymbol{\beta}) &= \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \log \pi^*(\boldsymbol{\beta}|\mathbf{y}) \\ &= \mathbf{X}^T \mathbf{V} \mathbf{X}, \end{aligned}$$

where \mathbf{V} is an $n \times n$ diagonal matrix with i th diagonal element

$$v_{ii} = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{\left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)^2}, \quad i = 1, \dots, n.$$

For large n , the posterior distribution is approximately $N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T \hat{\mathbf{V}} \mathbf{X})^{-1})$, where $\hat{\mathbf{V}}$ has the same form as \mathbf{V} but with v_{ii} replaced by its MLE.

Some situations where the asymptotic normality result *does not* hold.

- *Nonidentifiable models.* A model is *non-identifiable* when the distribution of \mathbf{Y} is the same throughout a set of θ values. In this case, the data cannot distinguish between such θ values.
- *Improper posteriors.* The posterior distribution must be proper in order for asymptotic normality to hold. An improper posterior will occur only when the *prior* is improper.
- *Priors that are 0 in a neighborhood of θ_0 .* The prior probability that θ is in $(\theta_0 - \epsilon, \theta_0 + \epsilon)$ should be larger than 0 for each $\epsilon > 0$.

- *True parameter is on boundary of Θ .* Suppose for example that $\Theta = [0, \infty)$ and $\theta_0 = 0$. Then the posterior will not be asymptotically normal because it will have no mass less than θ_0 .

Consistency of Bayes estimators

Bayesians are not nearly as concerned with consistency as are frequentists. Worrying about what happens at $n = \infty$ is a bit like "considering data sets that might have been, but were not, observed."

Nonetheless, it is somewhat reassuring to know that, under general conditions, Bayes estimators are consistent. In particular, the mode of the posterior is consistent under general conditions.

Consistency of the mode follows from the following facts:

- (i) The MLE and posterior mode are asymptotically the same (generally speaking), and
- (ii) MLEs are generally consistent.

The inverse of the information matrix is generally of order n^{-1} , and hence the posterior will be highly concentrated near the MLE.

Since the posterior mode is asymptotically equivalent to the MLE, it also inherits the *efficiency* properties of the MLE.

See Lehmann and Casella, *Theory of Point Estimation*, 2nd edition, for a full treatment of asymptotic properties of MLEs.