Bayesian Statistics
Asymptotic Theory

Shaobo Jin

Department of Mathematics

Bayesian Data Generating Process

In a Bayes model, the parameter θ is a random variable with known distribution π.

- Finding the true parameter makes no sense in a Bayes model.

Data that we observe are generated in a hierarchical manner:

$$\theta \sim \pi(\theta), \qquad X \mid \theta \sim f(x \mid \theta).$$

Given the data x, we can make inference for the data generating process.

In the usual frequentist statistics, we can find consistent estimators such that

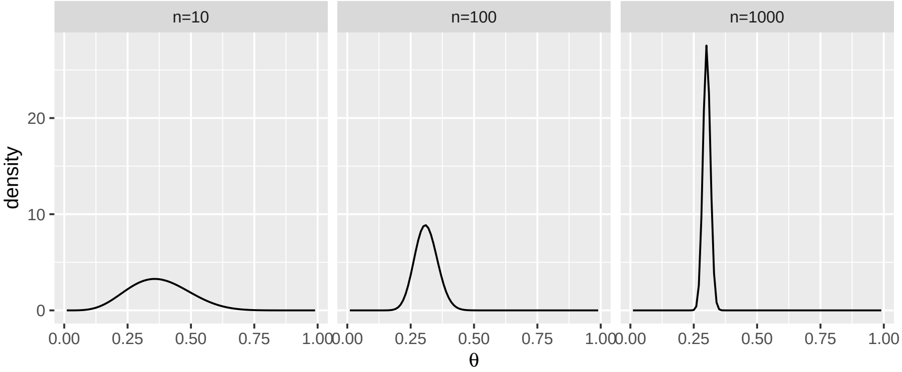
$$\mathrm{P}\left(\left\|\hat{\theta} - \theta\right\| < \epsilon\right) \rightarrow 1,$$

as $n \rightarrow \infty$. We also want the Bayes procedure to enable us to know θ with almost complete accuracy.

Example of Concentration

Example

Consider $X \mid \theta \sim \text{Binomial}\,(n, \theta)$ and $\theta \sim \text{Beta}\,(a_0, b_0)$. The posterior is $\theta \mid x \sim \text{Beta}\,(a_0 + x, b_0 + n - x)$, which concentrates around the true θ_0 as $n \to \infty$.



Convergence in Probability

Let $X \in \mathbb{R}^p$ be a $p \times 1$ random vector of random variables.

Definition (Convergence in probability)

X_n converges in probability to X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\left((X_n - X)^T (X_n - X) > \epsilon^2 \right) = 0.$$

It is denoted by $X_n \xrightarrow{P} X$. If X is a constant, then we also say X_n is consistent for X.

Example

Consider a sequence of independent random variables $\{X_n\}$, where $X_n \sim N\left(0,\ n^{-1}\right)$. Show that $X_n \xrightarrow{P} 0$.

Convergence Almost Surely

Definition (Convergence almost surely)

X_n converges almost surely to X if

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) \;=\; 1,$$

or equivalently, for every $\epsilon > 0$,

$$\mathrm{P}\left(\sqrt{\left(X_k - X\right)^T \left(X_k - X\right)} < \epsilon,\ \text{for all } k \geq n\right) \to 1$$

It is denoted by $X_n \overset{a.s.}{\to} X$.

Example

Consider a sequence of independent random variables $\{X_n\}$, where $X_n \sim N\left(0,\ n^{-1}\right)$. Show that $X_n \overset{a.s.}{\to} 0$.

Some Useful Results for Us

Theorem

$X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{P} X$, but not the reverse.

Theorem (Slutsky Theorem)

1. *If $X_n \xrightarrow{P} X$ and $X_n - Y_n \xrightarrow{P} 0$, then $Y_n \xrightarrow{P} X$.*
2. *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{P} \begin{bmatrix} X \\ Y \end{bmatrix}$.*

The theorem is also valid if every $\xrightarrow{P}$ is replaced by $\xrightarrow{a.s.}$.

Theorem (Continuous Mapping Theorem)

Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $P(X \in C) = 1$. If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$. The theorem is also valid if $\xrightarrow{P}$ is replaced by $\xrightarrow{a.s.}$.

Law of Large Numbers

Theorem

Let X_1, X_2,... be iid random vectors, and let $\bar{X}_n = n^{-1}\sum_i X_i$. Then,

1. *Weak law of large numbers: If $E\left[\sqrt{X^T X}\right] < \infty$, then $\bar{X}_n \xrightarrow{P} \mu = E(X)$.*
2. *Strong law of large numbers: $\bar{X}_n \xrightarrow{a.s.} \mu$ for some μ if and only if $E\left[\sqrt{X^T X}\right] < \infty$ and $\mu = E(X)$.*

Consistency of MLE

Theorem

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta_0)$*. Suppose that the density is identified such that* $f(x \mid \theta) = f(x \mid \theta_0)$ *implies* $\theta = \theta_0$*. Assume*

C1 Θ *is an open set in* $\mathbb{R}^p$*, where* θ_0 *is an interior point,*

Then, under some other assumptions, the maximizer $\hat{\theta}$ *of* $\sum_{i=1}^n \log f(x_i \mid \theta)$ *is consistent, i.e.,* $\hat{\theta}_n \xrightarrow{P} \theta_0$*.*

If we change C1 to

C1' Θ *is a compact set in* $\mathbb{R}^p$*, where* θ_0 *is an interior point,*

then, under additional assumptions, $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$*.*

MAP Estimator

Suppose that data are generated from $X_i \mid \theta = \theta_0 \sim f(x \mid \theta_0)$, $i = 1, ..., n$. The MAP estimator essentially maximizes

$$\sum_{i=1}^{n} \log f(x_i \mid \theta) + \log \pi(\theta).$$

- Since the MLE of θ is a consistent estimator of the true value θ_0, the MAP should also be consistent if $n^{-1} \log \pi(\theta) \to 0$ as $n \to \infty$.
- We should also expect the MAP estimator to be strongly consistent, converging almost surely to θ_0.

Posterior Consistency

Apart from consistency of the estimator, we can also introduce consistency of the posterior distribution, as a frequentist evaluation of Bayesian posterior.

Definition

Suppose that data are generated from $X \mid \theta = \theta_0 \sim f(x \mid \theta_0)$.

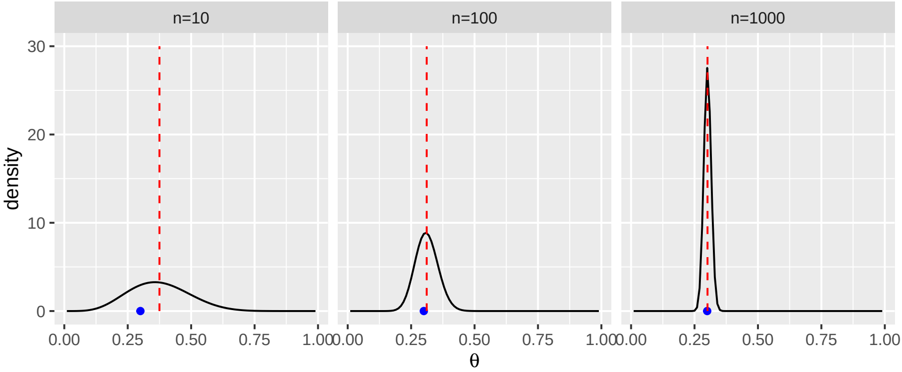
- the posterior is consistent at θ_0 if $\mathrm{P}(O \mid x)$ converges in probability to 1 under $f(x \mid \theta_0)$ as $n \to \infty$, for every open subset $O \subset \Theta$ with $\theta_0 \in O$.
- the posterior is strongly consistent at θ_0 if the convergence is almost surely. That is, for every open subset $O \subset \Theta$ with $\theta_0 \in O$, it holds that

$$\mathrm{P}(O \mid x) \to 1, \quad \text{as } n \to \infty, \quad \text{with probability 1.}$$

One Implication of Posterior Consistency

Posterior consistency suggests that, even though $\theta \sim \pi(\theta)$, the posterior $\pi(\theta \mid x)$ should concentrate around the θ_0 that generates the observed data.

- If $\pi(\theta \mid x)$ contracts to θ_0, we expect the Bayes estimator $\delta_B(x)$ should converge to θ_0.



Regularity Conditions

To establish such result, we need some regularity conditions. Let $L(\theta, d)$ be a loss function.

R1 There exists a constant $c_0 > 0$ for all d such that

$$c_0 \|d - \theta_0\| \leq L(\theta_0, d) - L(\theta_0, \theta_0).$$

- This condition implies that the loss function $L(\theta_0, \cdot)$ as a function of d has a minimum at $d = \theta_0$.

R2 There exists a constant K for all $X \sim f(x \mid \theta_0)$ such that

$$\int L^2(\theta, \theta_0) \pi(\theta \mid x) d\theta \leq K^2 \text{ almost sure.}$$

Consistency of Bayes Estimator

Theorem

Suppose that the loss function fulfills the conditions R1 and R2. Assume that

1. *for all* $\epsilon > 0$ *and all open sets* $O \subset \Theta$ *with* $\theta_0 \in O$, *it holds for*

$$B_\epsilon(\theta_0) = \{\theta : \theta \in O, |L(\theta, d) - L(\theta_0, d)| < \epsilon, \text{ for all } d\}$$

   that the prior probability $P(B_\epsilon(\theta_0)) > 0$ *and there is an open subset in* $B_\epsilon(\theta_0)$ *such that* θ_0 *is an interior point.*

2. *Let* $X \sim f(x \mid \theta_0)$ *and the sequence of posteriors* $\pi(\theta \mid x)$ *be strongly consistent at* θ_0.

Then, for $n \to \infty$, $\delta_B(x) \to \theta_0$ *almost surely.*

Consistency of General Estimator

Theorem

Suppose that the sequence of posteriors is strongly consistent at θ_0. Define the estimator $\hat{\theta}$ as the center of a ball of minimal radius that has posterior mass at least 0.5. Then $\hat{\theta}$ is consistent at θ_0.

Influence of Prior: Posterior

Theorem (Posterior robustness)

Consider $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta_0)$. *Let* θ_0 *be an interior point of* Θ, *and* π_1 *and* π_2 *be two prior densities, which are positive and continuous at* θ_0. *Let* $\pi_1(\theta \mid x)$ *and* $\pi_2(\theta \mid x)$ *be the respective posterior densities of* θ. *If* $\pi_1(\theta \mid x)$ *and* $\pi_2(\theta \mid x)$ *are both strongly consistent at* θ_0, *then*

$$\lim_{n\to\infty} \int \left|\pi_1(\theta \mid x) - \pi_2(\theta \mid x)\right| d\theta = 0, \text{ almost surely under } P_{\theta_0}.$$

Influence of Prior: Predictive Distribution

Theorem (Predictive robustness)

Assume that $\theta \mapsto P_\theta$ is one-to-one and continuous. Assume also that there is a compact set K such that $P(X \in K \mid \theta) = 1$ for all θ. Suppose that the posteriors $\pi_1(\theta \mid x)$ and $\pi_2(\theta \mid x)$ are both strongly consistent at θ_0, then the predictive distributions $\lambda_1(x^* \mid x)$ and $\lambda_2(x^* \mid x)$ satisfy

$$\lim_{n\to\infty} \left| \int \phi(x^*) \lambda_1(x^* \mid x) \, dx^* - \int \phi(x^*) \lambda_2(x^* \mid x) \, dx^* \right| = 0$$

for all bounded continuous functions ϕ.

Doob Consistency

Theorem (Doob's theorem for posterior consistency)

Suppose that $\theta \mapsto P(X \in A \mid \theta)$ *is one-to-one. Then, there exists a* $\Theta_0 \subseteq \Theta$ *with prior probability* $P(\Theta_0) = 1$ *such that, for every* $\theta_0 \in \Theta_0$, *if* $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta_0)$, *we have*

$$\lim_{n\to\infty} P(\theta \in O \mid X_1, \ldots, X_n) = 1, \text{ almost surely under } P_{\theta_0}$$

for any open set O *with* $\theta_0 \in O$.

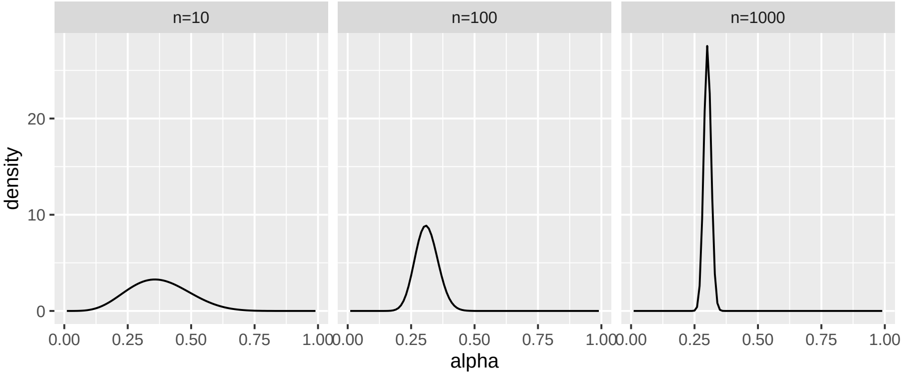
Doob's theorem says that the posterior will concentrate in a neighborhood, as long as

1. the statistics model is identified,
2. Θ_0 has strictly positive measure under the prior.

Example of Consistency

Example

Consider $X \mid \theta \sim \text{Binomial}\,(n, \theta)$ and $\theta \sim \text{Beta}\,(a_0, b_0)$. The posterior is $\theta \mid x \sim \text{Beta}\,(a_0 + x, b_0 + n - x)$. Show that posterior distributions concentrates around θ_0 as $n \to \infty$.



Positive Prior Assumption

The positive prior assumption plays a very important role in consistency. The posterior is

$$\pi(\theta \mid x) \quad \propto \quad f(x \mid \theta)\, \pi(\theta),$$

where $\pi(\theta \mid x) > 0$ only if $\pi(\theta) > 0$. We should never exclude any possible value from the prior.

Example

If $\pi(\theta) > 0$ for $\theta > 0$ and $\pi(\theta) = 0$ otherwise, then posterior can be better expressed as

$$\pi(\theta \mid x) \quad \propto \quad f(x \mid \theta)\, \pi(\theta)\, 1(\theta > 0).$$

The posterior is always zero for $\theta < 0$.

Doob's Theorem For Estimators

Theorem

Suppose that $\theta \mapsto P(X \in A \mid \theta)$ is one-to-one. Then, there exists a $\Theta_0 \subseteq \Theta$ with prior probability $P(\Theta_0) = 1$ such that, for every $\theta_0 \in \Theta_0$, if $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta_0)$, we have

$$\lim_{n\to\infty} E[g(\theta) \mid X_1, ..., X_n] = g(\theta_0), \text{ almost surely under } P_{\theta_0}$$

for any function $g(\theta)$ such that

$$\int g(\theta)\pi(\theta)\, d\theta < \infty.$$

For $g : \Theta \mapsto \mathbb{R}^p$, the theorem holds to each component of $g(\theta)$.

Doob's Theorem: Example

Example

Show that the posterior mean is strongly consistent.

1. Consider $X \mid \theta \sim \text{Binomial}\,(n, \theta)$ and $\theta \sim \text{Beta}\,(a_0, b_0)$. The posterior is $\theta \mid x \sim \text{Beta}\,(a_0 + x, b_0 + n - x)$.
2. Let X_1, ..., X_n be iid $N\left(\mu, \sigma^2\right)$, where σ^2 is known. Consider the prior $\mu \sim N\left(\mu_0,\ \sigma_0^2\right)$. The posterior is

$$\theta \mid x \quad \sim \quad N\left(\frac{\sigma_0^2 \sum_{i=1}^n x_i + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2}, \quad \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}\right).$$

Limitations of Doob Consistency

Doob's theorem establishes consistency under quite mild conditions. However, it has been criticized based on various grounds.

1. It only guarantees consistency on set Θ_0 with prior probability $\mathrm{P}\left(\Theta_0\right)=1$, not specific points θ_0.
2. It is less useful if θ is of infinite dimension, as the null set can be very large.

An alternative general theory is the Schwartz' theorem.

Distance/Divergence Between Two Distributions

Let P and Q be two probability measures with respective densities $p(x)$ and $q(x)$.

- Kullback-Leibler divergence (aka entropy loss):

$$\mathrm{KL}(\mathrm{P}, \mathrm{Q}) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx.$$

- Hellinger distance:

$$\mathrm{H}^2(\mathrm{P}, \mathrm{Q}) = \frac{1}{2}\int \left[\sqrt{p(x)} - \sqrt{q(x)}\right]^2 dx = 1 - \mathrm{H}_{\frac{1}{2}}(\mathrm{P}, \mathrm{Q}),$$

where $\mathrm{H}_{\frac{1}{2}}(\mathrm{P}, \mathrm{Q}) = \int \sqrt{p(x)\, q(x)} dx$ is the Hellinger transform.

Schwartz' Theorem

Theorem

Let X_1, ..., X_n be iid from P_θ, denoted by $P_\theta^{\otimes n}$. Let $f_n(x \mid \theta)$ be the density of $x = (x_1, ..., x_n)$.

1. *KL condition: Suppose that $P(K_\epsilon(\theta_0)) > 0$ for all $\epsilon > 0$, where $K_\epsilon(\theta_0) = \{\theta : KL(P_{\theta_0}, P_\theta) < \epsilon\}$.*
2. *Hellinger condition: For every open set $O \in \Theta$ with $\theta_0 \in O$, there exist constants D_0 and $q_0 < 1$, such that*

$$H_{\frac{1}{2}}\left(P_{\theta_0}^{\otimes n}, P_{n,O^c}\right) \quad \leq \quad D_0 q_0^n,$$

where P_{n,O^c} is defined by

$$P_{n,O^c}(A) \quad = \quad \int_A \int_{O^c} f_n(x \mid \theta) \frac{\pi(\theta)}{P(O^c)} d\theta dx.$$

Then, the sequence of posteriors is strongly consistent at θ_0.

Interpret the Conditions

1. The KL condition $\mathrm{P}\left(K_\epsilon\left(\theta_0\right)\right) > 0$ means that the prior does not exclude a neighborhood (in terms of the KL divergence) of θ_0.
2. The Hellinger condition means that the Hellinger distance

$$\mathrm{H}^2\left(\mathrm{P}_{\theta_0}^{\otimes n}, \mathrm{P}_{n,O^c}\right) = 1 - \mathrm{H}_{\frac{1}{2}}\left(\mathrm{P}_{\theta_0}^{\otimes n}, \mathrm{P}_{n,O^c}\right) \geq 1 - D_0 q_0^n.$$

   Intuitively speaking, we can distinguish between $\mathrm{P}_{\theta_0}^{\otimes n}$ and $\mathrm{P}_{\theta}^{\otimes n}$ if θ is not in O where $\theta_0 \in O$.

   - The Hellinger condition essentially replaces the identification condition ($\theta \mapsto \mathrm{P}\left(X \in A \mid \theta\right)$ is one-to-one) in Doob's theorem.

Hellinger Condition

Lemma

Let X_1, ..., X_n be iid from P_θ, denoted by $P_\theta^{\otimes n}$. Let $f_n(x \mid \theta)$ be the density of $x = (x_1, ..., x_n)$. Consider testing H_0: $P_{\theta_0}^{\otimes n}$ versus H_1: $\left\{P_\theta^{\otimes n}:\ \theta \in \Theta \setminus O\right\}$, where $O \in \Theta$ ise a neighborhood of θ_0. Suppose that there exists a nonrandomized test $\phi_n(x)$ and positive constants C and β such that

$$E[\phi_n(x) \mid \theta_0] + \sup_{\theta \in \Theta \setminus O} E[1 - \phi_n(x)] \quad \leq \quad C \exp(-n\beta).$$

Then the Hellinger condition holds.

Consistent Test

The condition in the lemma means that we can find a uniformly consistent test for testing H_0: $\theta = \theta_0$ versus H_1: $\theta \in O^c$, where $O \in \Theta$ ise a neighborhood of θ_0.

- That is, there exists a test $\phi_n(x)$ such that

$$\mathrm{E}\left[\phi_n(x) \mid \theta_0\right] \to 0, \qquad \sup_{\theta \in O^c} \mathrm{E}\left[1 - \phi_n(x) \mid \theta\right] \to 0.$$

- If we can find a uniformly consistent test, we can apply the Hoeffding's inequality to obtain the exponential rate.
- The existence of a uniformly consistent test only requires that we can find a uniformly consistent estimator of θ, i.e.,

$$\sup_{\theta} \mathrm{P}\left[\left(\hat{\theta} - \theta\right)^T \left(\hat{\theta} - \theta\right) > \epsilon \mid \theta\right] \ \to \ 0.$$

Schwartz' Theorem: Another Version

Theorem

Let X_1, ..., X_n be iid from P_θ, denoted by $P_\theta^{\otimes n}$. Let $f_n(x \mid \theta)$ be the density of $x = (x_1, ..., x_n)$.

1. *KL condition: Suppose that $P(K_\epsilon(\theta_0)) > 0$ for all $\epsilon > 0$, where $K_\epsilon(\theta_0) = \{\theta : \ KL(P_{\theta_0}, P_\theta) < \epsilon\}$.*
2. *Uniformly consistent test condition: Let $O \in \Theta$ be a neighborhood of θ_0. Consider testing H_0: $\theta = \theta_0$ versus H_1: $\theta \in O^c$. There exists a test $\phi_n(x)$ such that*

$$E[\phi_n(x) \mid \theta_0] \to 0, \qquad \sup_{\theta \in O^c} E[1 - \phi_n(x) \mid \theta] \to 0.$$

Then, the sequence of posteriors is strongly consistent at θ_0.

Consistency and Normality of MLE

Theorem

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta)$. *Assume that*

- C1 Θ *is an open set in* $\mathbb{R}^p$, *where* θ_0 *is an interior point,*
- C2 $\{x : f(x \mid \theta) > 0\}$ *does not depend on* θ, *i.e., common support,*
- C3 $\int f(x \mid \theta)\, dx$ *can be twice differentiable under the integral sign,*
- C4 *The Fisher information* $\mathcal{I}(\theta)$ *satisfies* $0 < I(\theta) < \infty$.

If some other regularity conditions are satisfied, then there exists a strongly consistent sequence $\hat{\theta}$ *of roots of the likelihood equation*

$$\frac{\partial \sum_{i=1}^{n} \log f(x_i \mid \theta)}{\partial \theta} = 0,$$

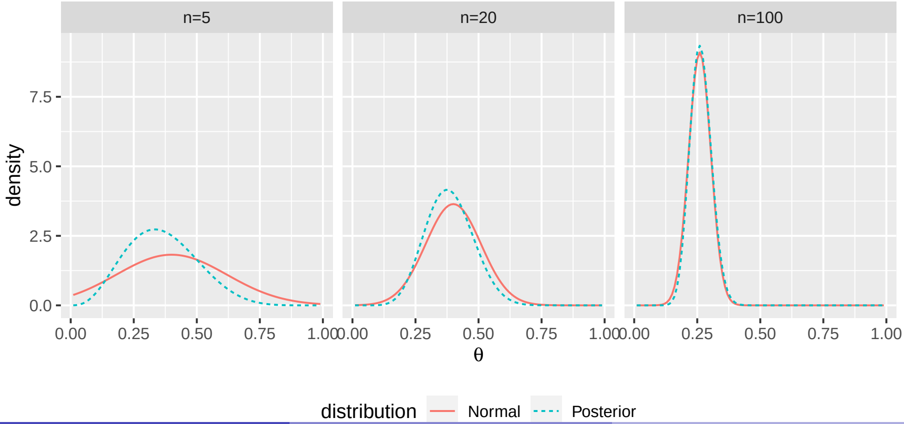
such that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \mathcal{I}^{-1}(\theta)\right).$$

Posterior Distribution

The posterior of a beta-binomial model is

$$\text{Beta}\left(a_0 + \sum_{i=1}^{n} x_i, b_0 + n - \sum_{i=1}^{n} x_i\right).$$



Normality of Posterior

The heuristic argument that we aim to conclude is that posterior distributions in differentiable parametric models converge to the Gaussian posterior distribution.

- If $\hat{\theta}$ is the MLE of θ, then

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \ \xrightarrow{d} \ N\left(0, \mathcal{I}^{-1}\left(\theta\right)\right).$$

- We want to claim that the difference between the posterior distribution $\pi\left(\theta \mid x_1, ..., x_n\right)$ and the normal distribution

$$\hat{\theta} \ \approx \ N\left(\theta, \frac{1}{n}\mathcal{I}^{-1}\left(\theta\right)\right)$$

  converge to zero.

Bernstein-von Mises Theorem

Let $\hat{\theta}$ be the strongly consistent sequence of roots of the likelihood equation. Define $t = \sqrt{n}\left(\theta - \hat{\theta}\right)$. Let $\pi\left(t \mid x_1, ..., x_n\right)$ be the posterior density of t.

Theorem

Let $X_1, \ldots, X_n \overset{iid}{\sim} f\left(x \mid \theta\right)$. Suppose that the assumptions C1 - C4 in the previous theorem hold. Assume that $\pi\left(\theta\right)$ is continuous and $\pi\left(\theta\right) > 0$ for all $\theta \in \Theta$.

1. *If some other regularity conditions are satisfied, then*

$$\left|\pi\left(t \mid x_1, ..., x_n\right) - \phi\left(t, 0, \mathcal{I}^{-1}\left(\theta\right)\right)\right| \overset{a.s.}{\to} 0 \text{ under } P_\theta,$$

where $\phi\left(t, 0, \mathcal{I}^{-1}\left(\theta\right)\right)$ is the density of $N\left(0, \mathcal{I}^{-1}\left(\theta\right)\right)$.

2. *If, in addition, $\mathcal{I}\left(\theta\right)$ is continuous, then,*

$$\left|\pi\left(t \mid x_1, ..., x_n\right) - \phi\left(t, 0, \mathcal{I}^{-1}\left(\hat{\theta}\right)\right)\right| \overset{a.s.}{\to} 0 \text{ under } P_\theta.$$

Bernstein-Von Mises Theorem: Example

Example

Suppose that $X_1,\ldots, X_n$ are iid Bernoulli(θ). We consider a continuous prior $\pi(\theta) > 0$ for all $\theta \in \Omega$. Approximate the posterior of θ.

Total Variation Distance

Let P and Q be two probability measures. Then, their total variation distance is

$$\sup_{A} \left| \mathrm{P}\left(A\right) - \mathrm{Q}\left(A\right) \right|,$$

for all Borel sets A. If p and q are the respective densities, then,

$$\sup_{A} \left| \mathrm{P}\left(A\right) - \mathrm{Q}\left(A\right) \right| = \frac{1}{2} \int \left| p\left(x\right) - q\left(x\right) \right| dx.$$

The Bernstein-von Mises theorem indicates that

$$\begin{aligned} \sup_{A} \left| \mathrm{P}\left(t \in A \mid x_1, ..., x_n\right) - \mathrm{P}\left(t \in A \mid t \sim \phi\left(t, 0, \mathcal{I}^{-1}\left(\theta\right)\right)\right) \right| &\overset{a.s.}{\to} 0, \\ \text{and} \int \left| \pi\left(t \mid x_1, ..., x_n\right) - \phi\left(t, 0, \mathcal{I}^{-1}\left(\theta\right)\right) \right| dt &\overset{a.s.}{\to} 0. \end{aligned}$$

Bayesian Credible Set

For simplicity, the classic one dimensional MLE $\hat{\theta}$ satisfies that $\mathrm{P}\left(\theta \in C\left(\alpha\right)\right) \to 1-\alpha$,where

$$C\left(\alpha\right) = \left[\hat{\theta} - \lambda_{1-\alpha/2}\sqrt{\frac{\mathcal{I}^{-1}\left(\theta\right)}{n}}, \hat{\theta} + \lambda_{1-\alpha/2}\sqrt{\frac{\mathcal{I}^{-1}\left(\theta\right)}{n}}\right].$$

The Bernstein-von Mises theorem allows us to approximate the posterior probability. In particular, let

$$B\left(\alpha\right) = \left\{\theta : \pi\left(\theta \mid x\right) \geq c_n\right\}$$

such that $\mathrm{P}\left(B\left(\alpha\right) \mid x\right) = 1-\alpha$. Then, for any $\epsilon > 0$,

$$\mathrm{P}\left(C\left(\alpha+\epsilon\right) \subset B\left(\alpha\right) \subset C\left(\alpha-\epsilon\right)\right) \to 1.$$

Example

Approximate the Bayesian credible set in the beta-binomial model.

Asymptotic Efficiency of Bayes Estimators

Consider the squared loss. The Bayes estimator of θ is the posterior mean $\tilde{\theta} = \mathrm{E}\left[\theta \mid x\right]$. The Bernstein-von Mises theorem may also indicate that

$$\mathrm{E}\left[\sqrt{n}\left(\theta - \hat{\theta}\right) \mid x\right] \;\; \rightarrow \;\; 0.$$

This suggests that

$$\sqrt{n}\left(\tilde{\theta} - \hat{\theta}\right) \;\; \rightarrow \;\; 0.$$

- The Bayes estimator and the MLE are asymptotically equivalent.
- The Bayes estimator is asymptotically efficient since MLE is asymptotically efficient.

An Counterexample

Example

Suppose that X_1, ..., X_n are iid with density

$$f(x \mid \theta) = \exp\{-(x-\theta)\}, \quad x > \theta.$$

The prior is $\theta \sim \text{Gamma}(2, b_0)$. Find the posterior of θ and show that the Bernstein-von Mises theorem is not applicable.

Counterexample: Posterior of Shifted Exponential