Handout 18: Asymptotic behavior of the posterior distribution

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim: We examine the properties of the posterior distribution $\Pi(\theta|y)$, under different sets of conditions, as the number of observations n increases $n \to \infty$, as well as their implications in inference.

References:

- Ferguson, T. S. (1996, Section 21). A course in large sample theory. Chapman and Hall/CRC.
- Chen, C. F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. Journal of the Royal Statistical Society: Series B (Methodological), 47(3), 540-546.
- Van der Vaart, A. W. (2000, Chapter 10). Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics.

Web-applets

- `https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/`
- `https://georgios-stats-1.shinyapps.io/demo_conjugatejeffreyslaplacepriors/`
- `https://georgios-stats-1.shinyapps.io/demo_mixturepriors/`

What is about?

Notation 1. Consider the Bayesian model $(F(x_{1:n}|\theta), \Pi(\theta))$ as

$$\begin{cases} x_{1:n}|\theta & \sim F(\cdot|\theta) \\ \theta & \sim \Pi(\cdot) \end{cases} \tag{1}$$

where a sequence of observables $x_{1:n} = (x_1, ..., x_n)$ are drawn from the parametric model $F(\cdot|\theta)$ admitting a pdf/pmf $f(\cdot|\theta)$ with unknown parameter $\theta \in \Theta$. The prior $\Pi(\cdot)$ of θ admits pdf/pmf $\pi(\cdot)$.

Question 2. *We study the behavior of the posterior distribution $\Pi(\theta|x_{1:n})$ with respect to the number of observables n, first when θ is a discrete parameter, and then when θ is a continuous one.*

Note 3. All the theorems in this chapter are frequentist in character, namely we study the posterior laws under the assumption that the observables $x_{1:n}$ is a random sample from the sampling distribution $F(\cdot|\theta^*)$ for some fixed non-random true value $\theta^* \in \Theta$.

Notation 4. We denote the likelihood function as $L_n(\theta) := f(x_{1:n}|\theta)$ and the posterior distribution $\Pi_n(\theta)$ with pdf/pmf $\pi_n(\theta) := \pi(\theta|x_{1:n}) \propto L_n(\theta)\pi(\theta)$, to easy the notation and make clear Note 3.

1 Discrete θ: Asymptotic consistency

Note 5. Given the Bayesian model (1), we consider cases where $\theta \in \Theta$ is a discrete parameter, and Θ is a countable space.

Note 6. The theorem below implies that, if Θ is countable, under conditions, the posterior distribution function of $\theta \in \Theta$ ultimately degenerates to a step function with a single (unit) step at $\theta = \theta^*$, where θ^* is the true value of the unknown discrete parameter θ.

Theorem 7. *Assume the Bayesian model (1), let $x_{1:n} = (x_1, ..., x_n)$ be a sequence of IID observables, $\theta \in \Theta$ be the unknown parameter with prior distribution mass $\pi(\theta)$, and posterior distribution mass $\pi_n(\theta)$, where Θ is a countable parametric space. Suppose $\theta^* \in \Theta$ is the (only) true value of θ such that $\pi(\theta^*) > 0$, and $-KL(f(\cdot|\theta^*)\,||\,f(\cdot|\theta)) := \int \log \frac{f(x|\theta)}{f(x|\theta^*)} dF(x|\theta^*) < 0$ for all $\theta \neq \theta^*$. Then*

$$\lim_{n\to\infty} \pi_n(\theta) = \begin{cases} 1 & , \theta = \theta^* \\ 0 & , \theta \neq \theta^* \end{cases}.$$

Proof. Due to exchangeability of $x_{1:n}$, it is

$$\pi_n(\theta) = \frac{\frac{L_n(\theta)}{L_n(\theta^*)}\pi(\theta)}{\sum_{\forall\theta\in\Theta} \frac{L_n(\theta)}{L_n(\theta^*)}\pi(\theta)} = \frac{\exp\left(\sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)}\right)\pi(\theta)}{\sum_{\forall\theta\in\Theta} \exp\left(\sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)}\right)\pi(\theta)} = \frac{\exp(S_n(\theta))\pi(\theta)}{\sum_{\forall\theta\in\Theta}\exp(S_n(\theta))\pi(\theta)}$$

where$S_n(\theta) = \sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)}$. From the SLLN, as $n \to \infty$, it is

$$\lim_{n\to\infty} \frac{1}{n} S_n(\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)} = \mathrm{E}_F\left(\log \frac{f(x|\theta)}{f(x|\theta^*)}|\theta^*\right), \qquad \text{a.s.} \tag{2}$$

By using Jensen's inequality and the fact that log is concave, it is

$$\mathrm{E}_{x\sim F(\cdot|\theta^*)}\left(\log \frac{f(x|\theta)}{f(x|\theta^*)}\right) \leq \log \mathrm{E}_{x\sim F(\cdot|\theta^*)}\left(\frac{f(x|\theta)}{f(x|\theta^*)}\right) = \log(1) = 0 \quad \Longrightarrow \mathrm{E}_{x\sim F(\cdot|\theta^*)}\left(\log \frac{f(x|\theta)}{f(x|\theta^*)}\right) \leq 0 \tag{3}$$

In (3), the equality holds for $\theta = \theta^*$ a.s., and the inequality holds for $\theta \neq \theta^*$ a.s., since Θ is a countable space and $\theta^* \in \Theta$, θ^* is "distinguishable" from the others, according to Theorem 39. Notice that, for any $\theta \neq \theta^*$, (2) and (3) imply that

$$\lim_{n\to\infty} \frac{1}{n} S_n(\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)} < 0, \qquad \text{a.s.}$$

which implies that

$$\lim_{n\to\infty} S_n(\theta) = \lim_{n\to\infty} n\frac{1}{n} S_n(\theta) = -\infty, \qquad \text{as}$$

Therefore,

- for any $\theta \neq \theta^*$, it is

$$\lim_{n\to\infty} \pi_n(\theta) = \lim_{n\to\infty} \frac{\exp(S_n(\theta))\pi(\theta)}{\sum_{\forall\theta\in\Theta}\exp(S_n(\theta))\pi(\theta)} = 0, \qquad \text{a.s.}$$

- for $\theta = \theta^*$, it is

$$\lim_{n\to\infty} \pi_n(\theta^*) = 1 - \sum_{\forall\theta\neq\theta^*} \underbrace{\lim_{n\to\infty} \pi_n(\theta)}_{=0,\text{ for }\theta\neq\theta^*} = 1, \qquad \text{a.s.}$$

□

Remark 8. Theorem 7 relies on the condition that the true parameter value θ^* is unique. If there was another θ^{**} such that $f(x|\theta^{**}) = f(x|\theta^*)$, we would observe IID data when θ equaled θ^* or θ^{**}, and hence the data could not discriminate between the two values.

Fact 9. *It can be shown that if $\theta^* \notin \Theta$, the posterior degenerates onto the value in Θ which gives the parametric model closest θ^*.*

2 Continuous θ: Asymptotic consistency and normality under Cramer's conditions

Note 10. Given the Bayesian model (1), we consider cases that $\theta \in \Theta$ is a continuous parameter, that $\Theta \subset \mathbb{R}^k$ is compact with $k \geq 1$, and that observables $\{x_i\}$ are IID.

Note 11. We show that when θ is continuous and under regularity conditions :

1. The posterior PDF of θ becomes more and more concentrated above an area around the true value θ^* as data size increases beyond a number $n \to \infty$.
2. The limiting posterior distribution of θ is close to a normal density $\mathrm{N}\left(\theta|\hat{\theta}_n, \frac{1}{n}\mathscr{I}(\theta^*)^{-1}\right)$ centered at $\hat{\theta}_n$ (the MLE of (19)), with variance $\frac{1}{n}\mathscr{I}(\theta^*)^{-1}$, Here, $\mathscr{I}(\cdot)$ is the Fisher information, where
$$\mathscr{I}(\theta) = \mathrm{E}_{x\sim F(\cdot|\theta)}\left((\nabla_\theta \log f(x|\theta))^\top (\nabla_\theta \log f(x|\theta))\right) = -\mathrm{E}_{x\sim F(\cdot|\theta)}\left(\nabla_\theta^2 \log f(x|\theta)\right).$$
3. These conclusions do not depend on the choice of the prior distribution provided that $\pi(\theta^*) > 0$.

Remark 12. The following version of the theorem, by Le Cam (1953), equivalently states that the posterior PDF of (the linear transformation) $\vartheta = \sqrt{n}(\theta - \hat{\theta}_n)$
$$\pi_n(\vartheta) = \frac{L_n(\vartheta\frac{1}{\sqrt{n}} + \hat{\theta}_n)\pi(\vartheta\frac{1}{\sqrt{n}} + \hat{\theta}_n)}{\int L_n(\vartheta\frac{1}{\sqrt{n}} + \hat{\theta}_n)\pi(\vartheta\frac{1}{\sqrt{n}} + \hat{\theta}_n)\mathrm{d}\vartheta}$$
approaches the PDF of $\mathrm{N}(0, \mathscr{I}(\theta^*)^{-1})$ as $n \to \infty$.

Condition 13. (Cramer conditions) Consider the following regular conditions.

d1 Θ is an open subset of $\mathbb{R}^k$

d2 second partial derivatives of $f(x|\theta)$ with respect to θ exist and are continuous for all x, and may be passed under the integral operator in $\int f(x|\theta)\mathrm{d}x$

d3 there is a function $K(x)$ such that $\mathrm{E}_{x\sim F(x|\theta^*)}(K(x)) < \infty$ and each component of $\nabla_\theta^2 \log(f(x|\theta))$ is bounded in absolute value by $K(x)$ uniformly in some neighborhood of θ^*

d4 $\mathscr{I}(\theta^*) = -\mathrm{E}_{x\sim F(\cdot|\theta^*)}\left(\nabla_{\theta^*}^2 \log f(x|\theta^*)\right)$ is positive definite

d5 (identifiability) $f(x|\theta) = f(x|\theta^*)$ a.s. then $\theta = \theta^*$

Theorem 14. *(Bernstein-von Mises) Let x_1, x_2,... be IID random variables drawn from a sample distribution with density $f(x|\theta)$, $\theta \in \Theta$, and let $\theta^* \in \Theta$ denote the true value of θ. Let $L_n(\theta) = f(x_{1:n}|\theta)$ denote the likelihood. Assume that the prior density $\pi(\theta)$ is continuous and $\pi(\theta) > 0$ for all $\theta \in \Theta$. Assume Conditions 13 hold. Then it is*
$$\frac{L_n\left(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}}\right)}{L_n(\hat{\theta}_n)}\pi\left(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}}\right) \xrightarrow{a.s.} \exp\left(-\frac{1}{2}\vartheta^\top \mathscr{I}(\theta_0)\vartheta\right)\pi(\theta^*), \tag{4}$$

where $\hat{\theta}_n$ is the strongly consistent sequence of roots of the likelihood equation (19) of Theorem 41. If, additionally,

$$\int_{\Theta} \frac{L_n(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}})}{L_n(\hat{\theta}_n)} \pi(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}}) d\vartheta \xrightarrow{a.s.} \int_{\Theta} \exp(-\frac{1}{2}\vartheta^{\top}\mathscr{I}(\theta^*)\vartheta)\pi(\theta^*) d\vartheta \tag{5}$$

then

$$\int_{\Theta} |\pi_n(\vartheta) - N(\vartheta|0, \mathscr{I}(\theta^*)^{-1})| d\theta \xrightarrow{a.s.} 0. \tag{6}$$

Proof. We prove: fist the existence of the the MLE, then (4), and finally (6).

Existence of consistent roots: I gonna use Theorem 41 (in Appendix) to prove that there exists a consistent sequence $\hat{\theta}_n$ of roots of (Eq. 19 in Appendix), and hence I need to show that its conditions are satisfied. Let $S_\rho = \{\theta : |\theta - \theta^*| \leq \rho\}$, with $\rho > 0$, be a neighborhood of θ^* on which (d3) is satisfied. So for $\Theta = S_\rho$ (in Theorem 41). Conditions (c1), (c2), (c5) of Theorem 41 (in Appendix) are automatic! Condition (c4) follows from continuity of $f(x|\theta)$ at θ. Condition (c3), ok.... By Taylor's theorem, I expand $D(x, \theta) = \log(f(x|\theta)) - \log(f(x|\theta^*))$ around θ^* as

Can be skipped

$$\begin{aligned} D(x, \theta) = D(x, \theta^*) &+ \nabla_\theta \log(f(x|\theta^*))(\theta - \theta^*) \\ &+ (\theta - \theta^*) \int_0^1 \int_0^1 v\nabla^2_\theta \log(f(x|\theta_0 + uv(\theta - \theta^*)))\mathrm{d}u\mathrm{d}v\, (\theta - \theta^*) \end{aligned}$$

So because $D(x, \theta^*) = 0$, $\nabla_\theta \log(f(x|\theta^*))$ is integrable, and the components of $\nabla^2_\theta \log(f(x|\theta))$ are bounded by $K(x)$ uniformly on S_ρ, we get that $D(x, \theta)$ is bounded on S_ρ. So (c3) holds.

Asymptotic Normality: Let

$$\ell_n(\theta) = \log(L_n(\theta))\,; \qquad \dot{\ell}_n(\theta) = \nabla_\theta \log(L_n(\theta))\,; \qquad \ddot{\ell}_n(\theta) = \nabla^2_\theta \log(L_n(\theta))$$

By Taylor's Theorem 38, we expand $\ell_n(\theta)$ around $\hat{\theta}_n$ as

$$\ell_n(\theta) = \ell_n(\hat{\theta}_n) + \dot{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + (\theta - \hat{\theta}_n)^{\top} I_n(\theta)(\theta - \hat{\theta}_n)$$

where

$$I_n(\theta) = -\frac{1}{n}\int_0^1 \int_0^1 v\ddot{\ell}_n(\hat{\theta}_n + uv(\theta - \hat{\theta}_n)\mathrm{d}u\mathrm{d}v) \tag{7}$$

Because, it is $\dot{\ell}_n(\hat{\theta}_n) = 0$ a.s., we get:

$$\begin{aligned} \ell_n(\theta) = \ell_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^{\top} I_n(\theta)(\theta - \hat{\theta}_n) &\iff \\ \frac{L_n(\theta)}{L_n(\hat{\theta}_n)} = \exp(-(\theta - \hat{\theta}_n)^{\top} I_n(\theta)(\theta - \hat{\theta}_n)), \quad \text{a.s.} \end{aligned}$$

Let's work on the asymptotics of (7); it is:

$$\begin{aligned} \frac{1}{n}\ddot{\ell}_n(\theta) = \frac{1}{n}\nabla^2_\theta \log(L_n(\theta)) = \frac{1}{n}\nabla^2_\theta \log(\prod_{i=1}^{n} f(x_i|\theta)) = \frac{1}{n}\sum_{i=1}^{n}\nabla^2_\theta \log(f(x_i|\theta)) \\ \xrightarrow{\text{a.s.}} \mathrm{E}_F\left(\nabla^2_\theta \log f(x|\theta)|\theta_0\right) \end{aligned} \tag{8}$$

as $n \to \infty$ by SLLN. Also, it is

$$\mathrm{E}_F\left(\nabla^2_\theta \log f(x|\theta)|\theta_0\right) = -\mathscr{I}(\theta_0) \tag{9}$$

Hence, from (8) and (9), I get

$$\frac{1}{n}\ddot{\ell}_n(\theta) \xrightarrow{\text{a.s.}} -\mathscr{I}(\theta_0) \tag{10}$$

Therefore,

$$I_n(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}}) = -\frac{1}{n}\int_0^1\int_0^1 v\ddot{\ell}_n(\hat{\theta}_n + uv(\theta-\hat{\theta}_n)\mathrm{d}u\mathrm{d}v \xrightarrow{\text{a.s.}} \frac{1}{2}\mathscr{I}(\theta^*) \tag{11}$$

because of (10) and because of $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta^*$ from Theorem 41 (in Appendix).

So back to what we wish to prove, and putting all these together, it is

$$\begin{aligned}\frac{L_n(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}})}{L_n(\hat{\theta}_n)}\pi(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}}) &= \exp(-\vartheta^\top I_n(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}})\vartheta)\pi(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}}) \\ &\xrightarrow{\text{a.s.}} \exp(-\frac{1}{2}\vartheta^\top\mathscr{I}(\theta^*)\vartheta)\pi(\theta^*)\end{aligned}$$

because of (11) and $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta^*$.

Now, about the second part of the proof. If (5) then by dividing (4) and (5), I get

$$\pi_n(\vartheta) \xrightarrow{\text{a.s.}} \mathrm{N}(\vartheta|0, \mathscr{I}(\theta^*)^{-1})$$

for all $\theta \in \Theta$. Hence By Scheffe's Theorem 35 (in Appendix) we get (6).

□

Remark 15. Note that Bernstein-von Mises Theorem 14 implies that the posterior distribution of $\vartheta = \sqrt{n}(\theta - \hat{\theta})$ given the data converges to the Normal distribution $\mathrm{N}(0, \mathscr{I}(\theta_0)^{-1})$ in Total Variation Norm, namely

$$\sup_{\forall A\subset\Theta}\left|\pi_n(\vartheta\in A) - \mathrm{N}(\vartheta\in A|0,\mathscr{I}(\theta^*)^{-1})\right|\mathrm{d}x \to 0, \qquad \text{as } n\to\infty$$

Corollary. *If the conditions of (Bernstein-von Mises) Theorem 14 hold, and if $\mathscr{I}(\theta)$ is continuous at Θ, then*

$$\sqrt{n}\mathscr{I}(\hat{\theta}_n)^{-1/2}(\theta-\hat{\theta}_n) \xrightarrow{D} z, \quad \text{where } z \sim N(0, I_k) \tag{12}$$

this is the result stated in Stat Concepts II notes (Term 2, 2017).

Proof. Bernstein-von Mises Theorem implies $\sqrt{n}(\theta-\hat{\theta}_n) \xrightarrow{D} \mathrm{N}(0,\mathscr{I}(\theta^*)^{-1})$ or equiv.

$$Y_n = \sqrt{n}\mathscr{I}(\theta^*)^{1/2}(\theta-\hat{\theta}_n) \xrightarrow{D} Z, \tag{13}$$

with $Z \sim \mathrm{N}(0, I_k)$. From Theorem 41 (in Appendix) I get $\hat{\theta}_n \to \theta^*$ a.s.. Due to continuity of $\mathscr{I}(\theta)$, it is

$$X_n = \mathscr{I}(\hat{\theta}_n)^{1/2}\mathscr{I}(\theta^*)^{-1/2} \xrightarrow{\text{a.s.}} I_k \tag{14}$$

According to Slutsky's theorem[1] by multiplying (13), (14), I get $X_nY_n \xrightarrow{D} Z$, i.e., $\sqrt{n}\mathscr{I}(\hat{\theta}_n)^{1/2}(\theta-\hat{\theta}_n) \xrightarrow{D} \mathrm{N}(0, I_k)$.

□

Example 16. Consider a Bayesian model

$$\begin{cases} x_i|\theta & \overset{\text{iid}}{\sim} \mathrm{Bn}(\theta), \qquad i = 1, ..., n \\ \theta & \sim \mathrm{Be}(a, b) \end{cases}$$

where $a > 0$, $b > 0$, and $n > 2$. Find the asymptotic posterior distribution of θ as $n \to \infty$, given Cramer's conditions.

[1]Sluky's theorem: If $Y_n \xrightarrow{D} Z$ and $X_n \xrightarrow{\text{a.s.}} c$, where $c \in \mathbb{R}^k$ is a constant, then $X_nY_n \xrightarrow{D} cZ$

Solution. I will find the MLE $\hat{\theta}_n$ of θ. The likelihood is $L_n(\theta) = \prod_{i=1}^{n} \mathrm{Bn}(x_i|\theta)$. Then

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(x_{1:n}|\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{i=1}^{n} \log(\mathrm{Bn}(x_i|\theta)) = \frac{n\theta - \sum_{i=1}^{n} x_i}{\theta(1-\theta)} \Longrightarrow$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(x_{1:n}|\theta)|_{\theta=\hat{\theta}_n} = 0 \quad \Longrightarrow \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

I will find the Fisher Information; it is

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log(f(x|\theta)) = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log(\mathrm{Bn}(x|\theta)) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \Longrightarrow$$

$$\mathscr{I}(\theta) = -\mathrm{E}_{\mathrm{Bn}(\theta)} \left(-\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \right) = \frac{1}{\theta(1-\theta)}$$

According to Bernstein-von Mises Theorem 14, it is $\theta|x_{1:n} \dot\sim \mathrm{N}\left(\hat{\theta}_n, \frac{1}{n}\mathscr{I}(\theta^*)^{-1}\right)$, where θ^* is the true value of θ. According to Corollary 2, it is $\theta|x_{1:n} \dot\sim \mathrm{N}\left(\hat{\theta}_n, \frac{1}{n}\mathscr{I}(\hat{\theta}_n)^{-1}\right)$ as well.

3 Continuous θ: Asymptotic distribution under Chen (1985) conditions

Notation 17. Let $U_n(\theta) = \log(\pi_n(\theta))$. Let $|\theta| = \sqrt{\theta^\top \theta}$. Let $B_\delta(\theta^*) = \{\theta \in \Theta; |\theta - \theta^*| < \delta\}$.

Assumption 18. *Assume that the posterior density has maximum at $\theta = m_n$ such that $\dot{U}_n(m_n) = 0$, and $\Sigma_n = -(\ddot{U}_n(m_n))^{-1}$ where $\Sigma_n > 0$ is positive-definite.*

Note 19. We consider weaker conditions which guarantee that the posterior density can be approximated by a Normal distribution around a small neighborhood of a posterior density maximum m_n as $n \to \infty$.

Remark 20. Assumption 18 is so general that (i.) $\{m_n\}$ is not assumed to converge, (ii.) $\pi_n(\theta)$ can be multimodal for each n, (iii.) m_n need not be the global maximum point of $\pi_n(\theta)$ for each n.

Condition 21. Consider the following regularity conditions

e1 (Steepness) $\bar{\sigma}_n^2 \to 0$ as $n \to \infty$ where $\bar{\sigma}_n^2$ is the largest eigenvalue of Σ_n

e2 (Smoothness) For any $\epsilon > 0$ there exists N and $\delta > 0$ such that, for any $n > N$ and $\theta \in B_\delta(m_n)$, $\ddot{U}_n(\theta)$ exists and satisfies

$$I - A(\epsilon) \leq \ddot{U}_n(\theta) (\ddot{U}_n(m_n))^{-1} \leq I + A(\epsilon),$$

where I is the $k \times k$ identity matrix, and $A(\epsilon)$ is a $k \times k$ symmetric positive-semidefinete matrix whose largest eigenvalue tends to zero as $\epsilon \to 0$

e3 (Concentration) For any $\delta > 0$, as $n \to \infty$.

$$Q_n := \int_{B_\delta(m_n)} \pi_n(\theta) \mathrm{d}\theta \to 1. \tag{15}$$

Remark 22. Conditions 21 are weaker than Cramer Conditions 13. Conditions (e1 & e2) imply that $\pi_n(\theta)$ becomes pick around m_n and behave like a normal kernel inside a neighborhood of m_n. Condition (e3) ensures that the mass outside that neighborhood is negligible. No IID sampling is assumed.

Lemma 23. *If conditions (e1) and (e2) hold then*

$$\lim_{n\to\infty} \pi_n(m_n) |\Sigma_n|^{1/2} \leq (2\pi)^{-k/2}. \tag{16}$$

The equality holds when condition (e3) is satisfied.

Proof. Omitted but provided in the Exercise sheet. □

Theorem 24. *Assume posterior density $\pi_n(\theta)$ has maximum at $\theta = m_n$ such that $\dot{U}_n(m_n) = 0$, $\Sigma_n > 0$ where $\Sigma_n = -(\ddot{U}_n(m_n))^{-1}$, and $U_n(\theta) = \log(\pi_n(\theta))$. Let $\phi_n = \Sigma_n^{-1/2}(\theta - m_n)$, where $\Sigma_n^{-1/2}$ is the inverse of the lower matrix from the Cholesky decomposition of Σ_n. Given (e1), and (e2), (e3) is necessary and sufficient condition so that*

$$\Sigma_n^{-1/2}(\theta - m_n) \underset{D}{\longrightarrow} Z; \quad \textit{where} \quad Z \sim N(0, 1).$$

Proof. Define $Z_n = \Sigma_n^{-1/2}(\theta - m_n)$. Assume $a, b \in \Theta$, such that $a \leq b$. It is sufficient to show that for any $a \leq 0$ and $b \geq 0$, it is $\lim_{n\to\infty} P_n(a, b) = P(a, b)$, where $P_n(a, b) = \mathsf{P}(a \leq Z_n \leq b)$ and $P(a, b) = \mathsf{P}(a \leq Z \leq b)$, if and only if C-3 holds.

Write

$$P_n(a, b) = \int_{R_n} \pi_n(\theta)\mathrm{d}\theta,$$

where $R_n = \{\theta | \, \Sigma_n^{1/2} a \leq (\theta - m_n) \leq \Sigma_n^{1/2} b\} \subseteq B_\delta(m_n)$, for any $\delta > 0$ and sufficiently large n, by (e1).

for every $\epsilon > 0$, $P_n(a, b) \in [P_n^-(a, b, \epsilon), P_n^+(a, b, \epsilon)]$, where

$$P_n^+(a, b, \epsilon) = \pi_n(m_n)|\Sigma_n|^{1/2}|I - A(\epsilon)|^{-1/2} \int_{R(\epsilon)} \exp(-\frac{1}{2}z^\top z)\mathrm{d}z;$$

$$P_n^-(a, b, \epsilon) = \pi_n(m_n)|\Sigma_n|^{1/2}|I + A(\epsilon)|^{-1/2} \int_{R(\epsilon)} \exp(-\frac{1}{2}z^\top z)\mathrm{d}z,$$

where $R(\epsilon) = \{z | \, [I - A(\epsilon)]^{-1/2} a \leq z \leq [I - A(\epsilon)]^{-1/2} b\}$.

By letting $\epsilon \to 0$, and under (e1), (e2), we get

$$\lim_{n\to\infty} P_n(a, b) = \lim_{n\to\infty} \pi_n(m_n)|\Sigma|^{1/2} \int_R \exp(-\frac{1}{2}z^\top z)\mathrm{d}z,$$

where $R = \{z | \, a \leq z \leq b\}$. According to Lemma 23, $\lim_{n\to\infty} P_n(a, b) = P(a, b)$ if and only if (e3) holds. □

Remark 25. Conditions (el) and (e2) in Theorem 24 are relatively easy to check in practice. Condition (e3) maybe be a bit tricky, hence, two alternative conditions (e3.1) and (e3.2) for the tail behaviors of $\pi_n(\theta)$ are provided. They are especially useful when m_n is the global maximum point of $\pi_n(\theta)$ for all n, such as in the unimodal case.

Proposition 26. *Assume that (el) and (e2) hold. Then, either (e3.1) or (e3.2) implies (e3).*

e3.1 *For any $\delta > 0$, there exists an integer N, and true numbers $c > 0$, $p > 0$ such that, for any $n > N$ and $\theta \notin B_\delta(m_n)$,*

$$U_n(\theta) - U_n(m_n) < -c((\theta - m_n)^\top \Sigma_n^{-1}(\theta - m_n))^p$$

e3.2 *For any $\delta > 0$, there exists an integer N, and real numbers $c > 0$, $p > 0$ such that, for any $n > N$ and $\theta \notin B_\delta(m_n)$,*

$$U_n(\theta) - U_n(m_n) < -c/|\Sigma_n|^p + \log(g(\theta)),$$

for some integrable function $g(\theta)$, i.e. $\int g(\theta)d\theta < \infty$.

Proof. Under (e3.1)

$$Q_n < \pi_n(m_n)|\Sigma_n|^{1/2} \int_{|z|>\delta/\sigma_n} \exp(-c(z^\top z)^d)\mathrm{d}z \tag{17}$$

Under (e3.2)

$$Q_n < \pi_n(m_n)|\Sigma_n|^{1/2}|\Sigma_n|^{-1/2} \int_{|z|>\delta/\sigma_n} \exp(-c|\Sigma_n|^{-d})\mathrm{d}z \tag{18}$$

where Q_n in (15). From Lemma 23 $\lim_{n\to\infty} \pi_n(\theta)|\Sigma_n|^{1/2}$ is bounded by $(2\pi)^{-k/2}$. Also rest terms in (17) and (18) tend to zero as $n \to \infty$. Then $\lim_{n\to\infty} Q_n = 0$, and e3 is implied. □

Remark 27. Assumptions (e3.1) and (e3.2) do not require the computation of the, often unknown, normalizing constant because it is simplified,

$$U_n(\theta) - U_n(m_n) = \log(f(x_{1:n}|\theta)) - \log(f(x_{1:n}|m_n)) + \log(\pi(\theta)) - \log(\pi(m_n)).$$

Remark 28. When the sample size is large enough, most priors will lead to the same inference and this inference will be equivalent to the one based only on the likelihood function.

Example 29. (Cont. Example 16) Consider the posterior distribution is $\theta|x_{1:n} \sim \text{Be}(a_n, b_n)$, where $a_n = a + n\bar{x}$, and $b_n = b + n - n\bar{x}$. Find the asymptotic distribution of θ as $n \to \infty$.

Solution. It is

$$U_n(\theta) = \log(\pi(\theta|x_{1:n})) = (a_n - 1)\log(\theta) + (b_n - 1)\log(1-\theta) - \log f(x_{1:n})$$

So

$$\dot{U}_n(\theta) = \frac{a_n - 1}{\theta} - \frac{b_n - 1}{1-\theta}; \qquad \ddot{U}_n(\theta) = \frac{a_n - 1}{\theta^2} - \frac{b_n - 1}{(1-\theta)^2};$$

Then

$$m_n := \frac{a_n - 1}{a_n + b_n - 2}; \qquad \Sigma_n := (-U_n''(m_n))^{-1} = \frac{(a_n-1)(b_n-1)}{(a_n+b_n-2)^3}.$$

Condition (e1) holds because $\lim_{n\to\infty}(-\ddot{U}_n(m_n))^{-1} = 0$. Condition (e2) holds because $\ddot{U}_n(\theta)$ is a continuous with respect of θ. Condition (e3) holds by using the same arguments as in Theorem **??**. Therefore, θ has asymptotic posterior distribution $\theta|x_{1:n} \dot{\sim} \text{N}(m_n, \Sigma_n)$.

4 Continuous θ: Asymptotic efficiency of Bayes Estimates

Remark 30. Consider the squared error loss $\ell(\theta, \delta) = (\theta - \delta)^\top(\theta - \delta)$ which implies the posterior expectation $\delta^\pi = \text{E}_\Pi(\theta|x_{1:n})$ as Bayes point estimator. Given that we can interchange the limit and the expectation operator of $\vartheta = \sqrt{n}(\theta - \hat{\theta}_n)$, we get $\sqrt{n}(\delta^\pi - \hat{\theta}_n) \xrightarrow{\text{P}} 0$ meaning that δ^π and $\hat{\theta}_n$ are asymptotically equivalent; i.e. $\delta^\pi - \hat{\theta}_n \xrightarrow{\text{P}} 0$. Hence

$$\sqrt{n}(\delta^\pi - \theta^*) = \sqrt{n}(\delta^\pi - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\text{D}} \text{N}(0, \mathscr{I}(\theta^*)^{-1})$$

which means that the Bayes estimator is asymptotically efficient.

Example 31. (Cont. Example 29) How the Bayes estimators under the square loss and under the 0-1 loss behave as $n \to \infty$?

Solution. The exact posterior distr. is $\theta|x_{1:n} \sim \text{Be}(a + n\bar{x}, b + n - n\bar{x})$. The squared loss and the $0-1$ loss imply the posterior mean $\delta_1(x_{1:n}) = \frac{n\bar{x}+a}{n+a+b}$ and posterior mode $\delta_2(x_{1:n}) = \frac{n\bar{x}+a-1}{n+a+b-2}$ as Bayes estimators correspondingly. Both converge to the MLE $\hat{\theta}_n = \bar{x}$ since $\lim_{n\to\infty}\delta_1(x_{1:n}) = \bar{x}$ and $\lim_{n\to\infty}\delta_2(x_{1:n}) = \bar{x}$.

5 Exercises

Try the Exercises 81, 82 from the Exercise sheet.

Appendix

A An inventory of definitions

Definition 32. (Types of converge) Assume a probability triplet $\{\Omega, \mathcal{F}, P\}$, and a sequence of random quantities $\{x_n; n = 1, 2, ...\}$, such that $x_n : \Omega \to \mathbb{R}^d$, $d > 0$. Then

- $\{x_n\}$ converges almost surely to a random quantify x if and only if
$$P(\lim_{n\to\infty} x_n = x) = 1.$$
It is demoted as $x_n \xrightarrow{\text{a.s.}} x$.
- $\{x_n\}$ converges in distribution to a random quantify x if and only if for all $\epsilon > 0$,
$$\lim_{n\to\infty} F_{x_n}(t) = F_x(t)$$
for all continuity points t of $F_{\cdot}$ in $\mathbb{R}$, where $F_{x_n}(t) = P(x_n \le t)$ and $F_x(t) = P(x \le t)$ are CDFs of $x_n$and x. It is demoted as $x_n \xrightarrow{D} x$.
- $\{x_n\}$ converges in total variation a random quantity x if and only if
$$\lim_{n\to\infty} \sup_{\forall B\subset\Theta} |P(x_n \in B) - P(x \in B)| = 0,$$
It is demoted as $x_n \xrightarrow{\text{T.V.}} x$.

Definition 33. ~~(Upper semicontinuous) A real-valued function, $f(\theta)$, defined on Θ is said to be upper semicontinuous (u.s.c.) on Θ, if for all $\theta \in \Theta$ and for any sequence θ_n in Θ such that $\theta_n \to \theta$, we have~~

$$\cancel{\lim_{n\to\infty} \sup f(\theta_n) \le f(\theta)}$$

Proposition 34. *If $\pi_n(\cdot)$ and $\pi(\cdot)$ are the PDFs of $x_n$and x correspondingly, then*
$$\sup_{\forall B\subset\Theta} |P(x_n \in B) - P(x \in B)| = \int \frac{1}{2} |\pi_n(t) - \pi(t)| dt$$

Theorem 35. *(Scheffe convergence theorem[2]) If $f_n(\cdot)$ and $g(\cdot)$ are density functions such that for all $x \in \mathcal{X}$ $\lim_{n\to\infty} f_n(x) = g(x)$, then*
$$\lim_{n\to\infty} \int_{\mathcal{X}} |f_n(x) - g(x)| dx = 0$$
(that is a point-wise convergence of densities)

Theorem 36. *If random variables x_n has density $f_n(x)$ and random variable x has density $g(x)$, and if $\lim_{n\to\infty} \int_{\mathcal{X}} |f_n(x) - g(x)| dx = 0$ then*
$$\sup_{\forall A\subset\mathcal{X}} |\mathsf{P}(x_n \in A) - \mathsf{P}(x \in A)| dx \to 0, \qquad \text{as } n \to \infty$$
that is called convergence in Total Variation.

Theorem 37. *(A Strong low of large numbers (SLLN)) Let $\{x_i\}_{i=1}^n$ be a sequence of IID random quantities, with $E(x_i) = \mu < \infty$, and $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ then $\bar{x}_n \xrightarrow{a.s.} \mu$.*

[2] This is not the original version, but it is what we need

Theorem 38. *(Taylor's theorem) If $f : \mathbb{R}^d \to \mathbb{R}$, and if $\nabla^2 f(x) = \nabla(\nabla f(x))^\top$ is continuous in the ball $\{x \in \mathcal{X} : |x - x_0| < r\}$, then for $|t| < r$, it is*

$$f(x_0 + t) = f(x_0) + \nabla f(x_0)t + t^\top \cdot \int_0^1 \int_0^1 u\nabla^2 f(x_0 + uvt)dudv \cdot t$$

Theorem 39. *(Shannon-Kolmogorov Information Inequality) Let $f_0(x)$ and $f_1(x)$ be densities with respect to Lesbeque measure dx. Then*

$$KL(f_0 \big|\big| f_1) = E_{F_0(x)}(\log \frac{f_0(x)}{f_1(x)}) = \int_{\mathcal{X}} \log \frac{f_0(x)}{f_1(x)} f_0(x)dx \geq 0,$$

with equality if and only if $f_1(x) = f_0(x)$ a.s.

Lemma 40. *(Passing the derivative under the integral operator) If $(\partial/\partial\theta)g(x,\theta)$ exists and is continuous in θ for all x and all θin an open interval $\mathcal{S}$ and if $|(\partial/\partial\theta)g(x,\theta)| \leq K(x)$ on $\mathcal{S}$ where $\int K(x)dx < \infty$ and if $\int g(x,\theta)dx$ exists on $\mathcal{S}$, then*

$$\frac{d}{d\theta} \int g(x,\theta)dx = \int \frac{d}{d\theta} g(x,\theta)dx$$

B Strong consistency of Maximum Likelihood Estimates

In frequentist statistics, given that $\nabla_\theta f(x|\theta)$ exists, one may seek to find the MLE $\hat{\theta}_n$ as the solution of the likelihood equation:

$$\hat{\theta}_n : \ \nabla_\theta \log f(x_{1:n}|\theta)|_{\theta=\hat{\theta}_n} = \sum_{i=1}^n \nabla_\theta \log f(x_i|\theta)|_{\theta=\hat{\theta}_n} = 0 \tag{19}$$

The following theorem states (more or less) that the MLE $\hat{\theta}_n$ in (19) is consistent.

Theorem 41. *(Strong consistency of MLE) Let x_1, x_2,... be IID random variables with density $f(x|\theta)$ (with respect to measure dx), $\theta \in \Theta$, and let θ^* denote the true value of θ. If the following conditions are satisfied:*

c1 *Θ is a closed and bounded set in $\mathbb{R}^k$*

c2 ~~*$f(x|\theta)$ is u.s.c. in θ for all $x \in \mathcal{X}$*~~ *$f(x|\theta)$ is continuous in θ for all x*

c3 *there is a function $K(x)$ such that $E^{f(x|\theta^*)}(|K(x)|) < \infty$ and*

$$\log(f(x|\theta)) - \log(f(x|\theta^*)) \leq K(x), \quad \forall x, \ \forall\theta$$

c4 *for all $\theta \in \Theta$ and sufficiency small $\rho > 0$, $\sup_{|\theta'-\theta|<\rho} f(x|\theta')$ is measurable in x*

c5 *(identifiability) $f(x|\theta) = f(x|\theta^*)$ a.s. then $\theta = \theta^*$*

then, for any sequence of maximum-likelihood estimates $\hat{\theta}_n$ of θ, it is

$$\hat{\theta}_n \xrightarrow{a.s.} \theta^* \tag{20}$$

The following theorem states (more or less) that the MLE is asymptotically normal.

Theorem 42. *(Cramer) Let x_1, x_2,... be IID random variables density $f(x|\theta)$ (with respect to some distribution $F(x|\theta)$), $\theta \in \Theta$, and let θ^* denote the true value of θ. If the conditions (d1)-(d5) stated in Theorem 14 (check in the next Theorem) are satisfied, then there exists a strongly consistent sequence $\hat{\theta}_n$ of roots of the likelihood equation (19) such that*

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{D} N(0, \mathscr{I}(\theta^*)^{-1})$$