Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2017

Handout 17: Asymptotic behavior of the posterior distribution [a]

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

Aim: We examine properties of the posterior distribution $\pi(\mathrm{d}\theta|x_{1:n})$ as the number of observations increases $n \rightarrow \infty$. We show that, as $n \rightarrow \infty$, :

- posterior beliefs about $\theta \in \Theta$ would become more and more concentrated around a "true" value θ^* (if exists)
- the unknown parameter $\theta \in \Theta$ has Normal distribution as asymptotic distribution (if $\Theta \subseteq \mathbb{R}^k$, $k \geq 1$).

References:

- Van der Vaart, A. W. (2000, Chapter 10). Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics.
- Ferguson, T. S. (1996, Section 21). A course in large sample theory. Chapman and Hall/CRC.
- DeGroot, M. H. (1970, Chapter 10). Optimal statistical decisions (Vol. 82). John Wiley & Sons.

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/`

[a] Author: Georgios P. Karagiannis.

What is about?

We consider the Bayesian model $(f(x_{1:n}|\theta), \pi(\mathrm{d}\theta))$ as

$$\begin{cases} x_i|\theta & \stackrel{\text{IID}}{\sim} f(\mathrm{d}\cdot|\theta), \qquad i=1,...,n \\ \theta & \sim \pi(\mathrm{d}\cdot) \end{cases} \tag{1}$$

where a sequence of observables $x_{1:n} = (x_1, ..., x_n)$ are drawn from the parametric model $f(\mathrm{d}x|\theta)$ with unknown parameter $\theta \in \Theta$. We study the behavior of the posterior distribution $\pi(\mathrm{d}\theta|x_{1:n})$ with respect to the number of observables n, first when θ is a discrete parameter, and them when it is a continuous one.

All the theorems in this chapter are frequentist in character, namely we study the posterior laws under the assumption that the observables $x_{1:n}$ are a random IID sample from the sampling distribution $f(\mathrm{d}x|\theta^*)$ for some fixed, non-random $\theta^* \in \Theta$.

1 Asymptotic consistency: when θ is discrete

In this section we focus in the case that $\theta \in \Theta$ is a discrete parameter, and Θ is a countable space, given the Bayesian model (1). In particular, the theorem below states that, for countable Θ, the posterior distribution function for $\theta \in \Theta$ ultimately degenerates to a step function with a single (unit) step at $\theta = \theta^*$, where θ^* is the real value of the unknown discrete parameter θ.

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

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

Proof. It is

$$\pi(\theta|x_{1:n}) = \frac{f(x_{1:n}|\theta)\pi(\theta)}{\sum_{\forall\theta\in\Theta} f(x_{1:n}|\theta)\pi(\theta)} = \frac{\frac{f(x_{1:n}|\theta)}{f(x_{1:n}|\theta^*)}\pi(\theta)}{\sum_{\forall\theta\in\Theta} \frac{f(x_{1:n}|\theta)}{f(x_{1:n}|\theta^*)}\pi(\theta)}.$$

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

$$\pi(\theta|x_{1:n}) = \frac{[\prod_{i=1}^n \frac{f(x_i|\theta)}{f(x_i|\theta^*)}]\pi(\theta)}{\sum_{\forall\theta\in\Theta}[\prod_{i=1}^n \frac{f(x_i|\theta)}{f(x_i|\theta^*)}]\pi(\theta)} = \frac{\exp(\overbrace{\sum_{i=1}^n \log\frac{f(x_i|\theta)}{f(x_i|\theta^*)}}^{=S_n(\theta)})\pi(\theta)}{\sum_{\forall\theta\in\Theta}\exp(\sum_{i=1}^n \log\frac{f(x_i|\theta)}{f(x_i|\theta^*)})\pi(\theta)} = \frac{\exp(S_n(\theta))\pi(\theta)}{\sum_{\forall\theta\in\Theta}\exp(S_n(\theta))\pi(\theta)}$$

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

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

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

$$\mathrm{E}^{f(\mathrm{d}x|\vartheta_T)}(\log\frac{f(x|\theta)}{f(x|\theta^*)}) \leq \log \mathrm{E}^{f(\mathrm{d}x|\theta^*)}(\frac{f(x|\theta)}{f(x|\theta^*)}) = \log(1) = 0$$

$$\implies \mathrm{E}^{f(\mathrm{d}x|\theta^*)}(\log\frac{f(x|\theta)}{f(x|\theta^*)}) \leq 0 \tag{3}$$

In the (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 15. 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

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

Therefore, □

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

$$\lim_{n\to\infty}\pi(\theta|x_{1:n}) = \lim_{n\to\infty}\frac{\exp(S_n(\theta))\pi(\theta)}{\sum_{\forall\theta\in\Theta}\exp(S_n(\theta))\pi(\theta)} = 0, \qquad \text{a.s.}$$

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

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

Remark 2. It can be shown that if $\theta^* \notin \Theta$, the posterior degenerates onto the value in Θ which gives the parametric model closest θ^*.

Remark 3. One important condition we considered was that the real 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.

Remark 4. If θ is a continuous parameter, then $\pi(\theta|x_{1:n})$ is always zero for any finite sample, and so the above theorem 1 apply.

2 Asymptotic consistency and normality: when θ is continuous

In this section, we focus in the case that $\theta \in \Theta$ may be a continuous parameter, $\Theta \subset \mathbb{R}^k$ is compact with $k \geq 1$, given the Bayesian model (1). In particular, under specific conditions, we prove that as $n \to \infty$:

1. the posterior PDF of θ becomes more and more concentrated above an area around the true value θ^*
2. the limiting posterior distribution of θ is a Normal distribution with specific parameters, and remarkably
3. these conclusions do not depend on the choice of the prior distribution provided that $\pi(\theta^*) > 0$.

We recall asymptotic results of MLE in Appendix B.

Consider the Bayesian model $(f(x_{1:n}|\theta), \pi(\mathrm{d}\theta))$ in (1), with $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^k$. Assume a prior distribution $\pi(\mathrm{d}\theta)$ with a PDF $\pi(\theta)$ where $\pi(\theta)$ continuous in Θ. To easy notation, we denote the likelihood as $L_n(\theta) := f(x_{1:n}|\theta) = \prod_{i=1}^n f(x_i|\theta)$. So the posterior PDF of θ is

$$\pi(\theta|x_{1:n}) = \frac{L_n(\theta)\pi(\theta)}{\int L_n(\theta)\pi(\theta)\mathrm{d}\theta} = \frac{\prod_{i=1}^n f(x_i|\theta)\pi(\theta)}{\int \prod_{i=1}^n f(x_i|\theta)\pi(\theta)\mathrm{d}\theta}$$

We present the Bernstein-von Mises theorem which (more-or-less) states that the posterior PDF $\pi(\theta|x_{1:n})$ is close to a normal density $\mathrm{N}(\theta|\hat{\theta}_n, \frac{1}{n}\mathscr{I}(\theta^*)^{-1})$ centered at $\hat{\theta}_n$ (the MLE of (16)), with variance $\frac{1}{n}\mathscr{I}(\theta^*)^{-1}$ when θ^* is the true value, when the size of the number of the observables n is really big. Here, $\mathscr{I}(\theta)$ is the Fisher information, where

$$\mathscr{I}(\theta) = \mathrm{E}^{f(\mathrm{d}\cdot|\theta)}((\nabla_\theta \log f(x|\theta))^\intercal(\nabla_\theta \log f(x|\theta))) = -\mathrm{E}^{f(\mathrm{d}\cdot|\theta)}(\nabla^2_\theta \log f(x|\theta))$$

The following version of the theorem is due to Le Cam (1953). It equivalently states that the posterior PDF of (the linear transformation) $\vartheta = \sqrt{n}(\theta - \hat{\theta}_n)$ given the data

$$\pi(\vartheta|x_{1:n}) = \frac{L_n(\vartheta\frac{1}{\sqrt{n}} + \hat{\theta}_n)\pi(\vartheta\frac{1}{\sqrt{n}} + \hat{\theta}_n)}{\int L_n(\vartheta\frac{1}{\sqrt{n}} + \hat{\theta}_n)\pi(\vartheta\frac{1}{\sqrt{n}} + \hat{\theta}_n)\mathrm{d}\vartheta}$$

approaches the PDF of $\mathrm{N}(0, \mathscr{I}(\theta^*)^{-1})$ and $n \to \infty$. Remarkably (!!!), this limiting posterior distribution is independent of the prior distribution $\pi(\theta)$.

Theorem 5. *(Bernstain-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 $\theta \in \Theta$. Under the following assumptions:*

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 $\mathrm{E}^{f(x|\theta_0)}(K(x)) < \infty$ and each component of $\nabla^2_\theta \log(f(x|\theta))$ is bounded in absolute value by $K(x)$ uniformly in some neighborhood of θ^**

d4 $\mathscr{I}(\theta_0) = -\mathrm{E}^{f(\mathrm{d}x|\theta_0)}(\nabla_\theta^2 \log(f(x|\theta)))$ *is positive definite*

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

It is

$$\frac{L_n(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}})}{L_n(\hat{\theta}_n)}\pi(\hat{\theta}_n + \vartheta\frac{1}{\sqrt{n}}) \xrightarrow{a.s.} \exp(-\frac{1}{2}\vartheta^\intercal \mathscr{I}(\theta_0)\vartheta)\pi(\theta^*), \tag{4}$$

where $\hat{\theta}_n$ *is the strongly consistent sequence of roots of the likelihood equation (16) of Theorem 17. If, additionally,*

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

then

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

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

- **Existence of consistent roots**

  I gonna use Theorem 17 to prove that there exists a consistent sequence $\hat{\theta}_n$ of roots of (16), and hence I need to show that its conditions are satisfied. Let $S_\rho = \{\theta : |\theta - \theta^*| \le \rho\}$, with $\rho > 0$, be a neighborhood of θ^* on which (d3) is satisfied. So for $\Theta = S_\rho$ (in Theorem 17) I have:

  - Conditions (c1), (c2), (c5), of that theorem are automatic!
  - Condition (c4) follows from continuity of $f(x|\theta)$ at θ. !
  - Condition (c3), ok.... By Taylor's theorem, I expand $U(x,\theta) = \log(f(x|\theta)) - \log(f(x|\theta^*))$ around θ^* as

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

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

- **Asymptotic Normality**

  Let

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

  By Taylor's Theorem 14, we expand $\ell_n(\theta)$ around $\hat{\theta}_n$ as

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

  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:

$$\ell_n(\theta) = \ell_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^\intercal I_n(\theta)(\theta - \hat{\theta}_n) \Longleftrightarrow$$
$$\frac{L_n(\theta)}{L_n(\hat{\theta}_n)} = \exp(-(\theta - \hat{\theta}_n)^\intercal I_n(\theta)(\theta - \hat{\theta}_n)), \quad \text{a.s.}$$

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

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

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

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

by Lemma **??**. 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 17.

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

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

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

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

for all $\theta \in \Theta$. Hence By Scheffe's Theorem 12 we get (6).

$\square$

Remark 6. Note that Bernstein-von Mises Theorem 5 and Theorem A imply 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

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

Remark 7. Bernstein-von Mises Theorem 5 says that the posterior PDFs of $\sqrt{n}\mathscr{I}(\theta^*)^{1/2}(\theta - \hat{\theta}_n)$ converges to the PDF of Normal $\mathrm{N}(0, I_k)$. Hence, this imply the CDF of $\sqrt{n}\mathscr{I}(\theta^*)^{1/2}(\theta - \hat{\theta}_n)$ converges to the PDF of Normal $\mathrm{N}(0, I_k)$ as well. Namely

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

and $\mathscr{I}(\cdot) = [\mathscr{I}(\cdot)^{1/2}][\mathscr{I}(\cdot)^{1/2}]^\intercal$.

Corollary. *If the conditions of Bernstein-von Mises Theorem 5 hold, and if $\mathscr{I}(\theta)$ is continuous at Θ, then*

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

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

Proof. From 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 17 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 $X_n Y_n \xrightarrow{D} Z$, i.e., $\sqrt{n}\mathscr{I}(\hat{\theta}_n)^{1/2}(\theta - \hat{\theta}_n) \xrightarrow{D} \mathrm{N}(0, I_k)$.
□

Asymptotic efficiency of Bayes Estimates

Consider the squared error loss $\ell(\theta, \delta) = (\theta - \delta)^\intercal(\theta - \delta)$ which implies the posterior expectation $\delta^\pi = \mathrm{E}^{\pi(\mathrm{d}|x_{1:n})}(\theta)$ 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 $\sqrt{n}(\delta^\pi - \hat{\theta}_n) \to 0$ meaning that δ^π and $\hat{\theta}_n$ are asymptotically equivalent; i.e. $\delta^\pi = \hat{\theta}_n$ a.s.. Hence

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

which means that the Bayes estimator is asymptotically efficient.

[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 $X_n Y_n \xrightarrow{D} cZ$

3 Implementation

Nowadays, Bayesian asymptotics can be mainly be used for theoretical purposes such as investigating the behavior of Bayesian procedures-methods-algorithms mainly addressing big dataset problems.

Below, we demonstrate a toy example for pedagogical purposes.

Example 8. Consider a Bayesian model

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

where $a > 0$, $b > 0$, and $n > 2$.

1. Find the asymptotic posterior distribution of θ as $n \to \infty$, by using Bernstein-von Mises Theorem.
2. What is the Bayes estimators under the square loss and under the 0-1 loss. How do they behave as $n \to \infty$.

Solution.

1. I will find the MLE by computing the roots of the equation (16). The likelihood is $f(x_{1:n}|\theta) = \prod_{i=1}^{n} \text{Bn}(x_i|\theta)$.Then

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

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

   Now, I will find the Fisher Information; it is

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

$$\mathscr{I}(\theta) = -\text{E}^{\text{Bn}(\theta)}(-\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}) = \frac{1}{\theta(1-\theta)} \tag{15}$$

   So according to Bernstein-von Mises Theorem 5, it is

$$\pi(\theta|x_{1:n}) \xrightarrow{\text{a.s.}} \text{N}(\theta|\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{=\hat{\theta}_n}, \underbrace{\frac{1}{n}\theta^*(1-\theta^*)}_{=\frac{1}{n}\mathscr{I}(\theta^*)^{-1}})$$

2. The exact posterior distr. is $\text{Be}(a + n\bar{x}, b + n - n\bar{x})$. 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 $\lim_{n\to\infty} \delta_2(x_{1:n}) = \bar{x}$.

Appendix

A An inventory of definitions

The following are definitions and statements used in proofs in Sections 1 and 2.

Definition 9. (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. 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 10. (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 Θsuch that $\theta_n \to \theta$, we have

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

Proposition 11. *If $\pi_n(\cdot)$ and $\pi(\cdot)$ are the PDFs of $x_n$and x correspondingly, then*

$$\sup_{\forall B\subset\Theta} |P(x_n \in B) - P(x \in B)| = \int \frac{1}{2}|\pi_n(t) - \pi(t)|\mathrm{d}t$$

Theorem 12. *(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*

$$\lim_{n\to\infty} \int_{\mathcal{X}} |f_n(x) - g(x)|\mathrm{d}x = 0$$

(that is a point-wise convergence of densities)

Theorem If random variables x_n has density $f_n(x)$ and random variable x has density $g(x)$, and if $\lim_{n\to\infty} \int_{\mathcal{X}} |f_n(x) - g(x)|\mathrm{d}x = 0$ then

$$\sup_{\forall A\subset\mathcal{X}} |\mathsf{P}(x_n \in A) - \mathsf{P}(x \in A)|\mathrm{d}x \to 0, \qquad \text{as } n \to \infty$$

that is called convergence in Total Variation.

Theorem 13. *(A Strong low of large numbers (SLLN)) Let $\{x_i\}_{i=1}^n$ be a sequence of IID random quantities, with $\mathrm{E}(x_i) = \mu < \infty$, 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 14. *(Taylor's theorem) If $f : \mathbb{R}^d \to \mathbb{R}$, and if $\nabla^2 f(x) = \nabla(\nabla f(x))^\intercal$ is continuous in the ball $\{x \in \mathcal{X} : |x - x_0| < r\}$, then for $|t| < r$, it is*

$$f(x_0 + t) = f(x_0) + \nabla f(x_0)t + t^\intercal \cdot \int_0^1 \int_0^1 u\nabla^2 f(x_0 + uvt)\mathrm{d}u\mathrm{d}v \cdot t$$

Lemma 15. *(Shannon-Kolmogorov Information Inequality) Let $f_0(x)$ and $f_1(x)$ be densities with respect to Lesbeque measure $\mathrm{d}x$. Then*

$$KL(f_0, f_1) = \mathrm{E}^{f_0(\mathrm{d}x)}(\log \frac{f_0(x)}{f_1(x)}) = \int_{\mathcal{X}} \log \frac{f_0(x)}{f_1(x)} f_0(x)\mathrm{d}x \geq 0,$$

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

Lemma 16. *(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 $\int K(x)\mathrm{d}x < \infty$ and if $\int g(x, \theta)\mathrm{d}x$ exists on $\mathcal{S}$, then*

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

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:

$$\hat{\theta}_n : \ \nabla_\theta \log f(x_{1:n}|\theta)|_{\theta=\hat{\theta}_n} = \sum_{i=1}^{n} \nabla_\theta \log f(x_i|\theta)|_{\theta=\hat{\theta}_n} = 0 \tag{16}$$

The following theorem states (more or less) that the MLE $\hat{\theta}_n$ in (16) is consistent.

Theorem 17. *(Strong consistency of MLE) Let x_1, x_2,... be IID random variables with density $f(x|\theta)$ (with respect to measure $\mathrm{d}x$), $\theta \in \Theta$, 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 $x \in \mathcal{X}$*

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

$$\hat{\theta}_n \xrightarrow{a.s.} \theta^* \tag{17}$$

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

Theorem 18. *(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[3] (d1)-(d5) stated in Theorem 5 (check in the next Theorem) are satisfied, then there exists a strongly consistent sequence $\hat{\theta}_n$ of roots of the likelihood equation (16) such that*

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

[3] The conditions of the theorem are the (d1)-(d5) in Theorem 5.