UTOPIAE Summer, July 2018

Introduction to Bayesian Statistics[a] – Monday 2^{th} July, 2018

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

[a] Link to material: `http://www.maths.dur.ac.uk/~mffk55/teaching.html`

1. Preliminaries

1.1. **Notation.** Consider a random variable (r.v.) $x \in \mathbb{R}^{d_x}$ on a probability triple $(\Omega, \mathcal{F}, \pi)$, as a function $x : \Omega \rightarrow \mathbb{R}^{d_x}$, with $x := x(\omega)$, which induces a probability π_x such as, for any $A \subseteq \mathbb{R}^{d_x}$,

$$\pi_x(x \in A) = \pi(\{\omega \in \Omega \,|\, x(\omega) \in A\}).$$

- The expected value of a measurable function $h(x)$ w.r.t. π_x is

$$\mathrm{E}_{\pi_x}(h(x)) \quad = \int_{\mathbb{R}^{d_x}} h(x)\mathrm{d}\pi_x(x) \quad = \int_{\mathbb{R}^{d_x}} h(x)\pi_x(\mathrm{d}x) \quad = \begin{cases} \int_{\mathbb{R}^{d_x}} h(x)\pi_x(x)\mathrm{d}x & \text{if } x \text{ is continuous} \\ \sum_{\forall x \in \mathbb{R}^{d_x}} h(x)\pi_x(x) & \text{if } x \text{ is discrete} \end{cases}$$

- The probability of $A \subseteq \mathbb{R}^{d_x}$ is

$$\pi_x(x \in A) = \mathrm{E}_{\pi_x}(\mathbb{1}(x \in A)) = \int_{\mathbb{R}^{d_x}} \mathbb{1}(x \in A)\pi_x(\mathrm{d}x) = \int_A \pi_x(\mathrm{d}x) = \begin{cases} \int_A \pi_x(x)\mathrm{d}x & \text{if } x \text{ is continuous} \\ \sum_{\forall x \in A} h(x)\pi_x(x) & \text{if } x \text{ is discrete} \end{cases}$$

1.2. **The Bayes theorem (Sets).** Consider a probability triplet $(\Omega, \mathcal{F}, \pi)$, then

- If $B \subseteq \Omega$, and $A \subseteq \Omega$, then it is

$$\pi(A|B) = \frac{\pi(B|A)\pi(A)}{\pi(B)}$$

  given that $B \neq \varnothing$

- If , $B \subseteq \Omega$, and $\{A_1, ..., A_k\}$ is a partition of Ω, then

$$\pi(A_j|B) = \frac{\pi(B|A_j)\pi(A_j)}{\sum_{j=1}^{k} \pi(B|A_j)\pi(A_j)}, \; \forall j = 1, ..., k$$

  given that $B \neq \varnothing$

1.3. **The Bayes theorem (random variables).** Consider, r.v. $x \in \mathbb{R}^{d_x}$, and r.v. $y \in \mathbb{R}^{d_y}$, then r.v. $x|y$ admits a distribution $\pi_{x|y}(\cdot|y)$ s.t.

- for any $C \subseteq \mathbb{R}^{d_x}$

$$\pi_{x|y}(x \in C|y) = \begin{cases} \int_C \quad \frac{\pi_{y|x}(y|x)\pi_x(x)}{\int_{\mathbb{R}^{d_x}} \quad \pi_{y|x}(y|x)\pi_x(x)\mathrm{d}x} \quad \mathrm{d}x & \text{, if } x \text{ is continuous} \\ \sum_{x \in C} \quad \frac{\pi_{y|x}(y|x)\pi_x(x)}{\sum_{x \in \mathbb{R}^{d_x}} \pi_{y|x}(y|x)\pi_x(x)} & \text{, if } x \text{ is discrete} \end{cases}$$

- the PDF, or PMF of r.v. $x|y$ is
$$\pi_{x|y}(x|y) = \frac{\pi_{y|x}(y|x)\pi_x(x)}{\int_{\mathbb{R}^{d_x}} \pi_{y|x}(y|x)\pi_x(\mathrm{d}x)}$$
- for any tiny set $\mathrm{d}x \subseteq \mathbb{R}^{d_x}$
$$\pi_{x|y}(\mathrm{d}x|y) = \frac{\pi_{y|x}(y|x)\pi_x(\mathrm{d}x)}{\int_{\mathbb{R}^{d_x}} \pi_{y|x}(y|x)\pi_x(\mathrm{d}x)}$$

2. Schools of Statistics

The Frequentist school: It uses the 'Frequency interpretation of probability' which asserts that the probability $\pi(A)$ of an event A is the limiting relative frequency of occurrence of the event in an infinite sequence of trials. Recall that classical rules of inference are judged on their long-run behavior in repeated sampling.

The Subjective Bayesian school: (the good one...) It uses the 'Subjective interpretation of probability' which asserts that the probability $\pi(A) := \pi(A|\Omega)$ represents a degree of belief in a proposition A, based on all the available information Ω. All probabilities are subjective, or personalistic judjements; they represent the investigator's degrees of belief. Hence, statistical analyses based on the same data, but performed by different researchers may be different.

3. The Exchangeable model

We introduce a simple, but surprisingly quite general probabilistic model.

Definition 3.1. (Finite exchangeability). The random quantities $\{x_1, ..., x_n\}$ are finitely exchangeable under a probability P if the implied distribution satisfies

$$P(x_1 \in A_1, ..., x_n \in A_n) = P(x_{\mathfrak{p}(1)} \in A_{\mathfrak{p}(1)}, ..., x_{\mathfrak{p}(n)} \in A_{\mathfrak{p}(n)})$$

for all permutations $\mathfrak{p}$ defined on the set $\{1, ..., n\}$.

- In terms of the corresponding PDF/PMF, the condition reduces to

$$p(x_1, ..., x_n) = p(x_{\mathfrak{p}(1)}, ..., x_{\mathfrak{p}(n)}).$$

Definition 3.2. (Infinite exchangeability). The random quantities $x_1, x_2, ...$ are said to be judged infinitely exchangeable under a probability P if every finite sub-sequence is judged exchangeable in the sense of Definition 3.1.

3.1. De Finetti representation theorem by (the 0-1 special case).

Theorem 3.1. *(Representation theorem for 0-1 random quantities). If $x_1, x_2, ...$ is an infinitely exchangeable sequence of $0-1$ random quantities with probability P, there exists a distribution π such that the joint prob. mass function $p(x_1, ..., x_n)$ for $x_1, ..., x_n$ has the form*
$$p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^n \underbrace{\theta^{x_i}(1-\theta)^{1-x_i}}_{=f(x_i|\theta)} \pi(\mathrm{d}\theta)$$

where

$$\pi(\theta \leqslant t) = \lim_{n\to\infty} \Pr(\frac{1}{n}\sum_{i=1}^{n} x_i \leqslant t)$$

and $\theta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} x_i$ *(i.e.* $\frac{1}{n}\sum_{i=1}^{n} x_i \to \theta$*, a.s.) is the (strong-law) limiting relative frequency of* 1*s.*

Interpretation: The representation of exchangeable sequence of $0-1$ random quantities $x_1, ..., x_n$ means that

- the x_i are considered to be independent Bernoulli random quantities, conditional on the random quantity θ. I.e., $x_i \overset{\text{IID}}{\sim} \text{Br}(\theta)$.
- θ is itself assigned a probability distribution $\pi(\mathrm{d}\theta)$,
- by the SLLN, $\theta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} x_i$, a.s., and hence $\pi(\mathrm{d}\cdot)$ can be interpreted as beliefs about the limiting relative frequency of 1's.

Proposition 3.1. *If* $x_1, x_2, ...$ *is an infinitely exchangeable sequence of random quantities admitting a PDF/PMF representation as in Theorem 3.1, then*

$$p(x_{n+1}|x_{1:n}) = \int_{\Theta} f(x_{n+1}|\theta)\pi(\mathrm{d}\theta|x_{1:n})$$

where

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

4. The Bayesian paradigm

4.1. **Definitions.**

Definition 4.1. (Parametric statistical model) A parametric statistical model consists of the observation of a sequence of a random variables $x_{1:n} = (x_1, ..., x_n)$, distributed according to prob. distribution $f(\mathrm{d}\cdot|\theta)$, where only the parameter $\theta \in \Theta$ is unknown and belongs to a vector space Θ of finite dimension.

Definition 4.2. (Bayesian model) A Bayesian statistical model is made of a parametric statistical model, $f(x_{1:n}|\theta)$, and a prior distribution $\pi(\mathrm{d}\theta)$ on the unknown parameters θ. It is denoted as $(f(x_{1:n}|\theta), \pi(\mathrm{d}\theta))$ or as

$$\begin{cases} x_{1:n}|\theta & \sim f(\mathrm{d}\cdot|\theta),\ \theta \in \Theta \\ \theta & \sim \pi(\mathrm{d}\theta) \end{cases} \tag{4.1}$$

4.2. **Distributions involved.** Consider the Bayesian model (4.1)

Sampling distribution of $x_{1:n}$ given θ: A parametrised probability distr. $f(\mathrm{d}\cdot|\theta)$ from which the data are assumed to have been generated (based on the researcher's judgments)

$$x_{1:n}|\theta \sim f(\mathrm{d}\cdot|\theta),\ \theta \in \Theta$$

Prior distribution of θ: Prior distribution $\pi(\mathrm{d}\theta)$

$$\theta \sim \pi(\mathrm{d}\theta)$$

is specified by the researcher. It represents the believes and judgments of the researcher before the collection of the data.

The likelihood function of θ given the data $x_{1:n}$: denoted as $L(\theta; x_{1:n})$, defined as ,

$$L(\theta; x_{1:n}) = f(x_{1:n}|\theta)$$

It contains the information available from the observed data $x_{1:n}$.

The posterior distribution of θ given the data $x_{1:n}$: is denoted as $\pi(\mathrm{d}\theta|x_{1:n})$ and has PDF/PMF

$$\pi(\theta|x_{1:n}) = \frac{L(\theta; x_{1:n})\pi(\theta)}{\int_\Theta L(\theta; x_{1:n})\pi(\mathrm{d}\theta)} = \frac{f(x_{1:n}|\theta)\pi(\theta)}{\int_\Theta f(x_{1:n}|\theta)\pi(\mathrm{d}\theta)} \tag{4.2}$$

$$\pi(\theta|x_{1:n}) \propto f(x_{1:n}|\theta)\pi(\theta) \qquad \text{...up to a normilisation constant}$$

which is derived from the Bayes theorem. It is presented the believes of the researcher in the light of the data after the experiment, and it is the main ingredient of the parametric inference. It can be considered as a mechanism to update one's believes about the uncertain parameter θ, in the light of the observed data.

The predictive distribution of $y := x_{n+1}$ given the data $x_{1:n}$: is denoted as $p(\mathrm{d}y|x_{1:n})$ and has PDF/PMF

$$p(y|x_{1:n}) = \mathrm{E}_{\pi_{\theta|x_{1:n}}}(f(y|\theta)) = \int_\Theta f(y|\theta)\pi(\mathrm{d}\theta|x_{1:n}) \tag{4.3}$$

It presents the believes of the researcher about a future observation given the observed data.

5. Prior specification: The conjugate priors

Posterior PDF/PMF, predictive PDF/PMF, as well as their expected values, require the computation of integrals/series, which are not necessarily tractable. E.g., see (4.2) and (4.3). A modeling trick leading to standard posterior distribution, and hence have convenient computations, is the specification of conjugate priors.

Definition 5.1. (Conjugate prior family) If $\mathcal{F} = \{f(\cdot|\theta); \forall\theta \in \Theta\}$ is a class of parametric models (sampling distributions), and $\mathcal{P} = \{\pi(\theta)\}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is conjugate for F if

$$\pi(\theta|x_{1:n}) \in \mathcal{P}, \ \ \forall f(\cdot|\theta) \in \mathcal{F} \, \text{and} \, \pi(\theta){\in}\mathcal{P}$$

5.1. The exponential family of distributions.

It is possible to specify conjugate prior on a parameter if the sampling distribution $f(\mathrm{d}\cdot|\theta)$ is a member of the exponential distribution family.

Definition 5.2. (k-parameter exponential family) A distribution with PDF/PMF $f(x|\theta)$ labeled by $\theta \in \Theta$, is said to belong to the k-parameter exponential family if $f(x|\theta)$ is of the form

$$f(x|\theta) = \mathrm{Ef}_k(x|u, g, h, \phi, \theta, c) = u(x)g(\theta)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(x)) \tag{5.1}$$

for $x \in \mathcal{X}$ where $h := (h_1, ..., h_k)$, $\phi(\theta) = (\phi_1, ..., \phi_k)$ and given the functions u, h, ϕ, and constants $\{c_j\}$,

$$g(\theta)^{-1} = \begin{cases} \int_{\mathcal{X}} u(x) \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) h_j(x)) \mathrm{d}x < \infty & \text{, if } x \text{ is cont} \\ \sum_{x \in \mathcal{X}} u(x) \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) h_j(x)) < \infty & \text{, if } x \text{ is disc} \end{cases}$$

Theorem 5.1. *(Conjugate families for regular exponential families)*[1], [2] *Assume a sequence of observables $x := (x_1, ..., x_n)$ generated from a distribution in the exponential family such that*

$$x_i|\theta \overset{IID}{\sim} Ef_k(\cdot|u, g, h, \phi, \theta, c); \quad i = 1, ..., n$$

then

- *The likelihood function is*

$$\begin{aligned} f(x_{1:n}|\theta) &= \prod_{i=1}^{n} Ef_k(x_i|u, g, h, c, \phi, \theta, c), \\ &= \prod_{i=1}^{n} u(x_i) g(\theta)^n \exp(\sum_{j=1}^{k} c_j \phi_j(\theta)(\sum_{i=1}^{n} h_j(x_i))). \end{aligned}$$

- *The parametric sufficient statistic*[3] *is $t_n := t_n(x_{1:n}) = (n, \sum_{i=1}^{n} h_1(x_i), ..., \sum_{i=1}^{n} h_k(x_i))$*
- *The conjugate prior distribution of $\theta \in \Theta$ has the form*

$$\begin{aligned} \pi(\theta|\tau) &= \frac{1}{K(\tau)} g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j) \\ &\propto \quad g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j) \end{aligned}$$

  where τ is such that $K(\tau) = \int_\Theta g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j) \mathrm{d}\theta < \infty$.
- *The posterior distribution for $\theta \in \Theta$ has the form*

$$\begin{aligned} \pi(\theta|x_{1:n}, \tau) &= \frac{1}{K(\tau^*)} g(\theta)^{\tau_0^*} \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j^*) \ = \pi(\theta|\tau^*) \\ &\propto \quad g(\theta)^{\tau_0^*} \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j^*) \end{aligned}$$

  with $\tau^* = (\tau_0^*, \tau_1^*, ..., \tau_k^*)$, $\tau_0^* = \tau_0 + n$, and $\tau_j^* = \sum_{i=1}^{n} h_j(x_i) + \tau_j$ for $j = 1, ..., k$.

[1] Web-applet: `https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/`

[2] Web-applet: `https://georgios-stats-1.shinyapps.io/demo_mixturepriors/`

[3] What is a parametric sufficient statistic?

Definition 5.3. (Parametric sufficient statistic) A statistic $t_n := t_n(x_{1:n})$ is parametric sufficient for an exchangeable sequence of observables $x_1, ..., x_n$ if and only if for any $n \geqslant 1$

$$\pi(\mathrm{d}\theta|x_{1:n}) = \pi(\mathrm{d}\theta|t_n)$$

How to find it?

Theorem 5.2. *(Neyman factorization criterion). The summary statistic $t_n = t(x_{1:n})$ is parametric sufficient for an exchangeable sequence of observables $x_1, ..., x_n$ if and only if, for any $n \geqslant 1$ the joint PDF/PMF for $x_{1:n}$ given θ has the form*

(5.2) $$f(x_1, ..., x_n|\theta) = h(t_n, \theta) g(x_1, ..., x_n), \qquad \textit{for some functions } h \geqslant 0,\ g \geqslant 0.$$

In other words,

$$\pi(\theta|x_{1:n},\tau) = \pi(\theta|\underbrace{\tau + t_n(x_{1:n})}_{=\tau*})$$

where $t_n(x_{1:n}) = (n, \sum_{i=1}^{n} h_1(x_i), ..., \sum_{i=1}^{n} h_k(x_i))$*.*

6. Point estimation

6.1. **Parametric inference .** Given the Bayesian model (4.1), we present the construction of point estimators for parametric inference.

Definition. Loss function, $\ell(\theta,\delta)$, is any function $\ell : \Theta \times \mathcal{D} \to [0,+\infty)$ that is supposed to evaluate the penalty (or degree of suffering) associated to the 'decision' $\delta \in \mathcal{D}$ when the uncertain parameter gets value $\theta \in \Theta$.

Definition 6.1. Bayes point estimator of θ with respect to the loss function $\ell(\theta,\delta)$ and the posterior distribution $\pi(\mathrm{d}\theta|x_{1:n})$ is the quantity δ^{π} which minimizes $\int_{\Theta} \ell(\theta,\delta)\pi(\mathrm{d}\theta|x_{1:n})$; i.e.

$$\begin{aligned}\delta^{\pi}(x_{1:n}) &= \arg\min_{\forall\delta\in\mathcal{D}} \mathrm{E}_{\pi(\mathrm{d}\theta|x_{1:n})}(\ell(\theta,\delta)) \\ &= \arg\min_{\forall\delta\in\mathcal{D}} \int_{\Theta} \ell(\theta,\delta)\pi(\mathrm{d}\theta|x_{1:n})\end{aligned}$$

Definition 6.2. (Estimator error; Univariate case) If $\theta \in \Theta \subseteq \mathbb{R}$ with posterior distribution $\pi(\mathrm{d}\theta|x_{1:n})$, and Bayes point estimator Bayes point estimator δ, then the standard error of δ^{π} is defined as

$$\mathrm{se}(\delta^{\pi}|x_{1:n}) = \sqrt{\mathrm{MSE}_{\pi(\mathrm{d}\theta|x_{1:n})}(\delta^{\pi})}$$

where

$$\mathrm{MSE}_{\pi(\mathrm{d}\theta|x_{1:n})}(\delta^{\pi}) = \mathrm{E}_{\pi(\mathrm{d}\theta|x_{1:n})}(\theta - \delta^{\pi})^2$$

is the mean squared error of δ.

Remark 6.1. MSE can be decomposed as a posterior variance of $\theta \in \Theta \in \mathbb{R}$ and bias of δ as

$$\underbrace{\mathrm{E}_{\pi(\mathrm{d}\theta|x_{1:n})}(\theta-\delta)^2}_{=\text{post. MSE}} = \underbrace{\mathrm{Var}_{\pi(\mathrm{d}\theta|x_{1:n})}(\theta)}_{=\text{post. var.}} + \underbrace{(\mathrm{E}_{\pi(\mathrm{d}\theta|x_{1:n})}(\theta) - \delta)}_{=\text{bias}}{}^2$$

Proposition 6.1. *(Standard point estimators) The Bayes estimate* $\delta^{\pi}(x_{1:n})$ *of* θ *with respect to the*[a]*:*

- weighted quadratic loss function $\ell(\theta,\delta) = w(\theta)(\theta-\delta)^2$, $w(\theta) > 0$ is

$$\delta^{\pi}(x_{1:n}) = \frac{\mathrm{E}_{\pi(\theta|x_{1:n})}(w(\theta)\theta)}{\mathrm{E}_{\pi(\theta|x_{1:n})}(w(\theta))}. \tag{6.1}$$

– quadratic loss function $\ell(\theta,\delta)=(\theta-\delta)^2$, is

$$\delta^{\pi}(x_{1:n})=\mathrm{E}_{\pi(\theta|x_{1:n})}(\theta) \tag{6.2}$$

– linear loss function $\ell(\theta,\delta)=c_1(\delta-\theta)\mathbb{1}_{\theta\leqslant\delta}(\delta)+c_2(\theta-\delta)\mathbb{1}_{\theta>\delta}(\delta)$ is

$$\delta^{\pi}(x_{1:n})\quad\text{such that}\quad\pi(\theta<\delta^{\pi}(x_{1:n})|x_{1:n})=\frac{c_2}{c_1+c_2}. \tag{6.3}$$

– absolute loss function $\ell(\theta,\delta)=|\theta-\delta|$, is

$$\delta(x_{1:n})=\mathrm{median}_{\pi(\theta|x_{1:n})}(\theta). \tag{6.4}$$

Proof. For weighted quadratic loss function, it is

$$\begin{aligned}0=\frac{\mathrm{d}}{\mathrm{d}\delta}\int_{\Theta}\ell(\theta,\delta)\pi(\mathrm{d}\theta|x_{1:n})|_{\delta=\delta^{\pi}}= &\quad =\frac{\mathrm{d}}{\mathrm{d}\delta}\int_{\Theta}\Theta w(\theta)(\theta-\delta)^2\pi(\theta|x_{1:n})\mathrm{d}\theta|_{\delta=\delta^{\pi}}\\ =2\int_{\Theta}w(\theta)(\theta-\delta^{\pi})\pi(\theta|x_{1:n})\mathrm{d}\theta &\quad =2[\int_{\Theta}w(\theta)\theta\pi(\theta|x_{1:n})\mathrm{d}\theta]-2[\int_{\Theta}w(\theta)\pi(\theta|x_{1:n})\mathrm{d}\theta]\delta^{\pi}.\\ =2\mathrm{E}(w(\theta)\theta|x_{1:n})-2\mathrm{E}(w(\theta)|x_{1:n})\delta^{\pi} &\quad =2(\mathrm{E}(w(\theta)\theta|x_{1:n})-\mathrm{E}(w(\theta)|x_{1:n})\delta^{\pi})\end{aligned}$$

Also, $\frac{\mathrm{d}^2}{\mathrm{d}\delta^2}\int_{\Theta}\ell(\theta,\delta)\pi(\mathrm{d}\theta|x_{1:n})|_{\delta=\delta^{\pi}}=-2\mathrm{E}(w(\theta)|x_{1:n})<0$. This completes the proof. The rest are yours. □

[a] Web-applet: `https://georgios-stats-1.shinyapps.io/demo_PointEstimation/`

Remark 6.2. Equations (6.1) and (6.2) exhibit a duality between loss and prior distribution, in the sense that it is equivalent to estimate θ with loss $\ell(\theta,\delta)=w(\theta)(\theta-\delta)^2$ and prior $\pi(\mathrm{d}\theta)=\pi(\theta)\mathrm{d}\theta$ (under (6.1)), or with loss $\tilde{\ell}(\theta,\delta)=(\theta-\delta)^2$ the prior $\tilde{\pi}(\mathrm{d}\theta)\propto w(\theta)\pi(\theta)\mathrm{d}\theta$ (under 6.2).

6.2. **Predictive inference.** Given the Bayesian model (4.1), we present the construction of point estimators for predictive inference.

Definition 6.3. (Bayes predictive point estimate) Bayes predictive point estimate of $y=x_{n+1}\in\mathcal{X}$ with respect to the loss function $\ell(y,\delta)$ and the predictive distribution $p(\mathrm{d}y|x_{1:n})$ is the quantity $\delta\in\mathcal{D}=\mathcal{X}$ which minimizes $\int_{\mathcal{X}}\ell(y,\delta)p(\mathrm{d}y|x_{1:n})$; i.e.

$$\begin{aligned}\delta^{p}(x_{1:n})&=\arg\min_{\forall\delta}\mathrm{E}_{p(\mathrm{d}y|x_{1:n})}(\ell(y,\delta))\\&=\arg\min_{\forall\delta}\int_{\mathcal{X}}\ell(y,\delta)p(\mathrm{d}y|x_{1:n})\end{aligned}$$

Proposition 6.2. *(Standard point estimators) The Bayes estimate $\delta^p(x_{1:n})$ of $y = x_{n+1}$ with respect to the*[a]*:*

- *weighted quadratic loss function $\ell(y,\delta) = w(y)(y-\delta)^2$, $w(y) > 0$ is*

$$\delta^p(x_{1:n}) = \frac{\mathrm{E}_{p(\mathrm{d}y|x_{1:n})}(w(y)y)}{\mathrm{E}_{p(\mathrm{d}y|x_{1:n})}(w(y))}.$$

- *quadratic loss function $\ell(\theta,\delta) = (\theta-\delta)^2$, is*

$$\delta^p(x_{1:n}) = \mathrm{E}_{p(y\mathrm{d}|x_{1:n})}(y)$$

- *linear loss function $\ell(y,\delta) = c_1(\delta - y)\mathbb{1}_{y\leqslant\delta}(\delta) + c_2(y-\delta)\mathbb{1}_{y>\delta}(\delta)$ is*

$$\frac{c_2}{c_1+c_2}\text{-th quantile of the predictive distribution } p(\mathrm{d}y|x_{1:n}).$$

- *absolute loss function $\ell(y,\delta) = |y-\delta|$, is*

$$\delta^p(x_{1:n}) = \mathrm{median}_{p(\mathrm{d}y|x_{1:n})}(y).$$

[a] Web-applet: `https://georgios-stats-1.shinyapps.io/demo_PointEstimation/`

7. Credible sets

7.1. **Parametric credible sets.** Given the Bayesian model (4.1), we present the construction of the credible sets for parametric inference.

Definition 7.1. (Posterior Credible Set) Any set $C_a \subseteq \Theta$ such that [4]

$$\pi(\theta \in C_a|x_{1:n}) = \int_{C_a} \pi(\mathrm{d}\theta|x_{1:n}) \geqslant 1-a \tag{7.1}$$

is called '$100(1-a)\%$' posterior credible set for θ, with respect to the posterior distribution $\pi(\mathrm{d}\theta|x_{1:n})$.

Definition 7.2. (Posterior highest probability density (HPD) Set)[5] The $100(1-a)\%$ highest probability density set for $\theta \in \Theta$ with respect to the posterior distribution $\pi(\mathrm{d}\theta|x_{1:n})$ is the subset C_a of Θ of the form

$$C_a = \{\theta \in \Theta : \pi(\theta|x_{1:n}) \geqslant k_a\}$$

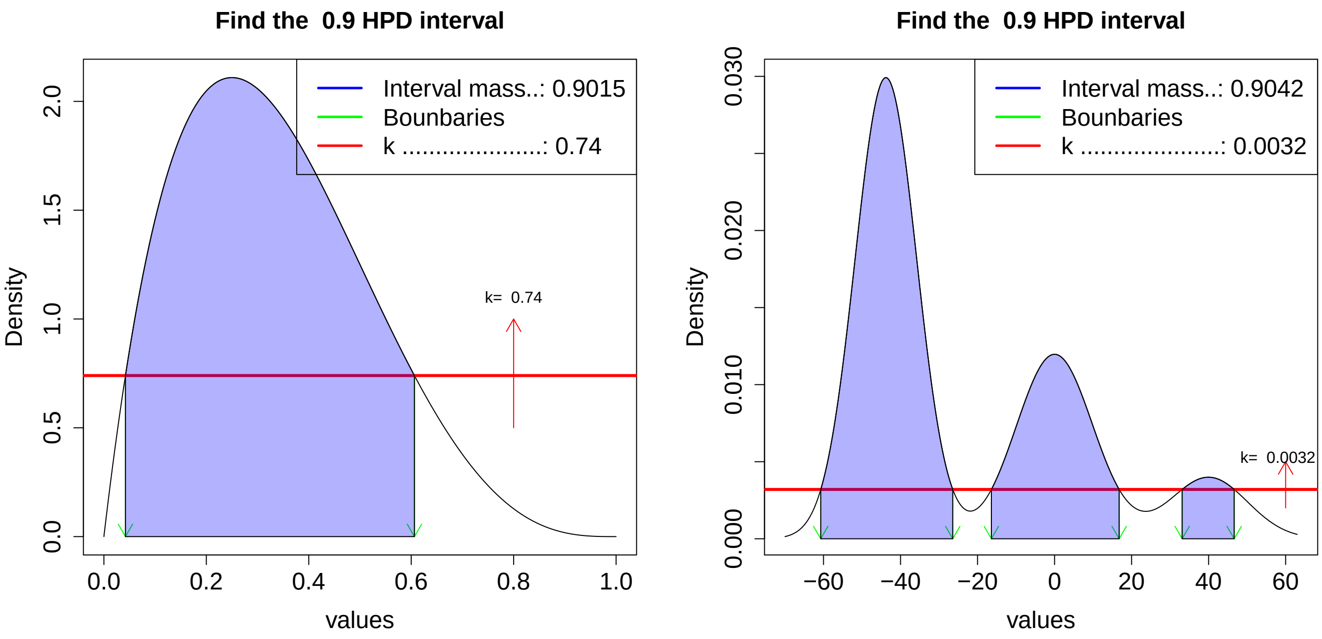
where k_a is the largest constant such that

$$\pi(\theta \in C_a|x_{1:n}) \geqslant 1-a$$

Remark 7.1. HPD sets are the credible sets with the smallest 'size'; (proof is omitted) .

[4] In (7.1), '=' is for continuous θ, and '$\geqslant$' is for discrete to consider step functions

[5] Web-applet: `https://georgios-stats-1.shinyapps.io/demo_CredibleSets`



(A) Uni-modal case

(B) Multi-modal case

FIGURE 7.1. Schematic of 1D HPD set Remark 7.1

When $\theta \in \Theta \subseteq \mathbb{R}$ is 1D, the following theorem facilitates the computations for the evaluation of the HPD credible interval.

Theorem 7.1. *Let θ be a random variable that admits measure $\pi(\mathrm{d}\theta|x_{1:n})$ with unimodal density $\pi(\theta|x_{1:n})$. If the interval $C_a = [L, U]$ satisfies*

(1) $\int_L^U \pi(\theta|x_{1:n})\mathrm{d}\theta = 1 - a$,

(2) $\pi(U|x_{1:n}) = \pi(L|x_{1:n}) > 0$, *and*

(3) $\theta_{mode} \in (L, U)$, *where* θ_{mode} is the mode of $\pi(\theta|x_{1:n})$,

then interval $C_a = [L, U]$ is the HPD interval of θ with respect to $\pi(\mathrm{d}\theta|x_{1:n})$.

7.2. **Predictive credible sets.** Given the Bayesian model (4.1), the construction of the predictive credible sets for a future observable $y = x_{n+1}$ is analogues to that of $\theta \in \Theta$. We just need to replace θ with y, and $\pi(\theta|x_{1:n})$ with the predictive distribution $p(y|x_{1:n})$. Try to modify definitions (7.1) and (7.2) accordingly by yourself.

8. Hypothesis tests

Given the (overall) Bayesian model

$$\begin{cases} x_i|\theta & \overset{\text{IID}}{\sim} f(\mathrm{d}\cdot|\theta), \qquad \theta \in \Theta, \qquad i = 1, ..., n \\ \theta & \sim \pi(\mathrm{d}\theta) \end{cases} \tag{8.1}$$

we present the construction of the Bayesian hypothesis test of the form

$$\begin{cases} \mathrm{H}_0 : & \theta \in \Theta_0 \\ \mathrm{H}_1 : & \theta \in \Theta_1 \end{cases} \tag{8.2}$$

where sub-sets $\{\Theta_0, \Theta_1\}$ partition Θ; such as $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \varnothing$.

The overall prior distr. $\pi(\mathrm{d}\theta)$ can be factorized based on $\{\Theta_0, \Theta_1\}$ as

$$\pi(\mathrm{d}\theta) = \pi_0 \times \pi_0(\mathrm{d}\theta) + \pi_1 \times \pi_1(\mathrm{d}\theta) \;=\begin{cases} \pi_0 \times \pi_0(\theta) & ,\text{if}\, \theta \in \Theta_0 \\ \pi_1 \times \pi_1(\theta) & ,\text{if}\, \theta \in \Theta_1 \end{cases}, \tag{8.3}$$

$$\pi_0 = \underbrace{\int_{\Theta_0} \pi(\mathrm{d}\theta),}_{=\pi(\theta \in \Theta_0)} \qquad \pi_0(\theta) = \underbrace{\frac{\pi(\theta)\mathbb{1}_{\Theta_0}(\theta)}{\int_{\Theta_0} \pi(\mathrm{d}\theta)}}_{=\pi(\theta|\theta\in\Theta_0)}, \qquad \pi_1 = \underbrace{\int_{\Theta_1} \pi(\mathrm{d}\theta),}_{=\pi(\theta \in \Theta_1)} \qquad \pi_1(\theta) = \underbrace{\frac{\pi(\theta)\mathbb{1}_{\Theta_1}(\theta)}{\int_{\Theta_1} \pi(\mathrm{d}\theta)}}_{=\pi(\theta|\theta\in\Theta_1)}.$$

where $\theta \in \Theta = \Theta_0 \cup \Theta_1$.

Notice that, π_0, and π_1 describe the prior probabilities on H_0 and H_1 respectively, while $\pi_0(\theta)$ and $\pi_1(\theta)$ describe how the prior mass is spread out over the hypotheses H_0 and H_1 respectively.

Then note that formally, hypothesis test (8.2), can be consider as comparing Bayesian (sub-)models

$$\begin{cases} \mathrm{H}_0 : & \theta \in \Theta_0 \\ \mathrm{H}_1 : & \theta \in \Theta_1 \end{cases} \Longrightarrow \begin{cases} \mathrm{H}_0 : & (\, f(x_{1:n}|\theta)\;,\; \pi_0(\mathrm{d}\theta)\,) \\ \mathrm{H}_1 : & (\, f(x_{1:n}|\theta)\;,\; \pi_1(\mathrm{d}\theta)\,) \end{cases} \Longrightarrow \begin{cases} \mathrm{H}_0 : & x_{1:n} \sim p_0(x_{1:n}) = \int_{\Theta_0} \prod_{i=1}^n f(x_i|\theta)\pi_0(\mathrm{d}\theta) \\ \mathrm{H}_1 : & x_{1:n} \sim p_1(x_{1:n}) = \int_{\Theta_1} \prod_{i=1}^n f(x_i|\theta)\pi_1(\mathrm{d}\theta) \end{cases} \tag{8.4}$$

8.1. **General approach.** Bayes hypothesis test (8.4) can be addressed as a Bayesian parametric point estimation of the indicator function $\mathbb{1}_{\Theta_1}(\theta)$

$$\mathbb{1}_{\Theta_1}(\theta) = \begin{cases} 0 & , \theta \in \Theta_0 \\ 1 & , \theta \in \Theta_1 \end{cases} \tag{8.5}$$

under a loss function $\ell(\theta, \delta)$, with $\theta \in \Theta$, $\delta \in \mathcal{D} = \{0, 1\}$. Here, δ^π is the estimator of $\mathbb{1}_{\Theta_1}(\theta)$ in (8.5), where $\mathbb{1}_{\Theta_1}(\theta) = 1$ implies H_1, while $\mathbb{1}_{\Theta_1}(\theta) = 0$ implies H_0.

Theorem 8.1. *The Bayes estimator δ^π of $\mathbb{1}_{\Theta_1}(\theta)$ in (8.5), which is associated with the prior $\pi(\mathrm{d}\theta)$ and the $c_I - c_{II}$ loss function*

$$\ell(\theta, \delta) = \begin{cases} 0 & , \textit{if}\; \theta \in \Theta_0,\; \delta = 0 \\ 0 & , \textit{if}\; \theta \notin \Theta_0,\; \delta = 1 \\ c_{II} & , \textit{if}\; \theta \notin \Theta_0,\; \delta = 0 \\ c_I & , \textit{if}\; \theta \in \Theta_0,\; \delta = 1 \end{cases} \tag{8.6}$$

where $c_I > 0$ and $c_{II} > 0$ are specified by the researcher is

$$\delta(x_{1:n}) = \begin{cases} 0 & , \textit{if } \pi(\theta \in \Theta_0 | x_{1:n}) > \frac{c_{II}}{c_{II}+c_I} \\ 1 & , \textit{otherwise} \end{cases} \tag{8.7}$$

where $\pi(\theta \in \Theta_0|x_{1:n}) = \int_{\Theta_0} \pi(\mathrm{d}\theta|x_{1:n})$, and $\{\Theta_0, \Theta_1\}$ is a partition of Θ.

8.2. **Bayes Factor approach.** Equivalently, Bayesian hypothesis test is performed with a metric called Bayes Factor.

Definition 8.1. The Bayes factor $\mathrm{B}_{01}(x_{1:n})$ is the ratio of the posterior probabilities of H_0 and H_1 over the ratio of the prior probabilities of H_0 and H_1.

$$\mathrm{B}_{01}(x_{1:n}) = \frac{\pi(\theta \in \Theta_0|x_{1:n})/\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1|x_{1:n})/\pi(\theta \in \Theta_1)} \tag{8.8}$$

where

$$\pi(\theta \in \Theta_0) = \int_{\Theta_0} \pi(\mathrm{d}\theta); \quad \pi(\theta \in \Theta_0|x_{1:n}) = \int_{\Theta_0} \pi(\mathrm{d}\theta|x_{1:n}); \quad \pi(\theta \in \Theta_1) = \int_{\Theta_1} \pi(\mathrm{d}\theta); \quad \pi(\theta \in \Theta_1|x_{1:n}) = \int_{\Theta_1} \pi(\mathrm{d}\theta|x_{1:n}).$$

Remark 8.1. Bayes factor $\mathrm{B}_{01}(x_{1:n})$

- can be interpreted as the 'odds in favour of H_0 against H_1 that are given by the data' $x_{1:n}$.
- evaluates the modification of the odds of Θ_0 against Θ_1 due to the observations $x_{1:n}$.
- is the ratio of the likelihoods weighted by the conditional priors $\pi_0(\mathrm{d}\theta)$ and $\pi_1(\mathrm{d}\theta)$; hence also depends on both data and priors

Proposition 8.1. *Consider a hypothesis test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ as described in (8.4), and given a Bayesian model (8.1). Then*

$$B_{01}(x_{1:n}) = \frac{\pi(\theta \in \Theta_0|x_{1:n})/\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1|x_{1:n})/\pi(\theta \in \Theta_1)} = \begin{cases} \frac{\int_{\Theta_0} f(x_{1:n}|\theta)\pi_0(\mathrm{d}\theta)}{\int_{\Theta_1} f(x_{1:n}|\theta)\pi_1(\mathrm{d}\theta)} & ; H_0 : \theta \in \Theta_0 \quad \textit{versus} \quad H_1 : \theta \in \Theta_1 \\ \frac{f(x_{1:n}|\theta_0)}{\int_{\Theta_1} f(x_{1:n}|\theta)\pi_1(\mathrm{d}\theta)} & ; H_0 : \theta \in \{\theta_0\} \quad \textit{versus} \quad H_1 : \theta \in \Theta_1 \\ \frac{f(x_{1:n}|\theta_0)}{f(x_{1:n}|\theta_1)} & ; H_0 : \theta \in \{\theta_0\} \quad \textit{versus} \quad H_1 : \theta \in \{\theta_1\} \end{cases} \tag{8.9}$$

π_0, π_1, $\pi_0(\mathrm{d}\theta)$ and $\pi_1(\mathrm{d}\theta)$ as described above.

Proposition 8.2. *Consider a hypothesis test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ as described in (8.4) with the $c_I - c_{II}$ loss function (8.6), and given a Bayesian model (8.1). The hypothesis H_0 is accepted if*

$$B_{01}(x_{1:n}) > \frac{c_{II}}{c_I}\frac{\pi_1}{\pi_0} \tag{8.10}$$

and, the hypothesis H_1 is accepted otherwise.

Proof. Just rearrange (8.7) □

9. Homework

(1) Prove:

   (a) Proposition 3.1

   (b) Theorem 5.1

   (c) Point estimators (6.3) and (6.4), in Proposition 6.1

   (d) Special cases of Bayes factor in the bracket of (8.9), in Proposition 8.1

(2) Consider observations generated from Normal sampling distribution $\mathrm{N}(\mu, \sigma^2)$ with unknown mean μ and known variance σ^2. Let $\mathrm{d}\pi(\mu)$ denote the prior of μ.

   (a) Find the conjugate prior for μ.

   (b) Construct the HPD posterior $(1-a)100\%$ interval for μ.

   (c) Construct the predictive distribution of a future $y = x_{n+1}$. [Hint: It is a non central t distribution]

Appendix A. Distributions

Bernoulli distribution.

Symbolized $x \sim \mathrm{Br}(\theta)$; with $\theta \in [0,1]$

PMF $\mathrm{Br}(x|\theta) = \theta^x (1-\theta)^{1-x} \mathbb{1}_{\{0,1\}}(x)$; with $\theta \in [0,1]$

Uniform distribution.

Symbolized $x \sim \mathrm{Un}(a,b)$; with and $b > a$, and $a \in \mathbb{R}$, and $b \in \mathbb{R}$

PDF $\mathrm{Un}(x|a,b) = \frac{1}{b-a} \mathbb{1}_{(a,b)}(x)$; with and $b > a$, and $a \in \mathbb{R}$, and $b \in \mathbb{R}$

Beta distribution.

Symbolized $x \sim \mathrm{Be}(a,b)$; with $a > 0$, and $b > 0$

PDF $\mathrm{Be}(x|a,b) = \frac{1}{\mathrm{B}(a,b)} x^{a-1}(1-x)^{b-1} \mathbb{1}_{(0,1)}(x)$; with $a > 0$, and $b > 0$

Appendix B. Web applications

The software for the web applications is available from GitHub at

`https://github.com/georgios-stats/UTOPIAE-Bayes.git`The applications can run live from:

Conjugate priors:

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

Point estimators:

`https://georgios-stats-1.shinyapps.io/demo_PointEstimation/`

Credible intervals:

`https://georgios-stats-1.shinyapps.io/demo_CredibleSets`