Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

# Handout 3: The Bayesian paradigm & Subjective probability

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

**Aim**

To understand some foundation concepts of the Bayesian statistics: Subjective probability, an Axiomatic system, Bayesian paradigm.

**Reading list:**

- DeGroot, M. H. (1970, or 2005; Chapter 6). Optimal statistical decisions (Vol. 82). John Wiley & Sons.
- O'Hagan, A., & Forster, J. J. (2004; Paragraphs 4.1-1.16). Kendall's advanced theory of statistics, volume 2B: Bayesian inference (Vol. 2). Arnold.
- Robert, C. (2007; Sections 1.1, 1.2, 1.4). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

## 1 Some Schools of Statistics

Probability is a well defined mathematical quantity, however it has several interpretations; e.g. Frequetist, Subjective, etc. Probability is a fundamental concept in Statistics; different interpretations of Probability lead to different schools of statistics. Below we give more details about the main two (Frequentist and Bayesian) schools of statistics.

**The Frequentist school of statistics**

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

Frequentist statisticians presume that the observations have been generated from a (idealized) model which would presumably run along the lines "out of infinitely many worlds one is selected at random...". Often that model has quantities (parameters) whose values are unknown to You, but You are interested in learning. These parameters are presumed to be constants quantities in the sense that they are equal to an ideal/real value which You want to discover.

**The Subjective Bayesian school of statistics**

The Subjective Bayesian school of statistics uses the 'Subjective interpretation of probability' which asserts that the probability $\mathsf{P}(A)$ represents a degree of belief in a proposition $A$, based on all the available information $\Omega$. In Subjective Bayesian statistics all probabilities and distributions are subjective, or personalistic; they represent Your (investigator's) degrees of belief.

Subjective probability concerns Yours judgments about uncertain events or propositions. Eg., $\mathsf{P}(A)$ measures the strength of Your degree of belief that $A$ will occur. $\mathsf{P}(A) = 1$ describes that You are certain that $A$ will occur, and $\mathsf{P}(A) = 0$ describes that You are certain that $A$ will not occur. As $\mathsf{P}(A)$ increases from $0$ to $1$, it describes an increasing degree of belief in the occurrence of $A$. Different researchers may have a different degree of belief in the same proposition, and these different researchers can assign a different probabilities to that proposition based on their

own judgments. The only constraint is that a Your probabilities should not be inconsistent, and therefore they should obey the Kolmogorov axioms of probability.

**The Objective Bayesian school of statistics**

The Objective Bayesian school of statistics uses the 'Logical interpretation of probability' which asserts that the probability represents a degree of belief in a proposition, based on all the available information; but that this is not a subjective matter of an individual's personal degree of belief.

Probabilities are expressions of the plausibility of statements given a state of knowledge. According to this point of view, given the same information, everyone should assign the same probabilities: there is no room for personal belief. As with the Bayesian subjective point of view, probability is not connected to 'randomness'. Nor does it have any particular connection to frequencies, although of course frequencies can be reasoned about as well as any other quantity.

## 2 Subjective probability and its construction

Acceptance of the Bayesian method as the natural and proper approach to statistical inference has become almost synonymous with the adoption of a subjective interpretation of probability. For instance, a fully subjective interpretation of probability, allows the (Subjective) Bayesian analysis to avoid to produce controversial results, such as violation of the Likelihood Principle.

### 2.1 Axiomatic formulation of relative likelihood

Consider $(\Omega, \mathscr{F})$ where $\Omega$ is a sample space and $\mathscr{F}$ is a $\sigma$-algebra of events. Consider events $A, B \in \mathscr{F}$ as sets containing one or more elements from the set $\Omega$. We think of $A$, $B$, $\Omega$ as a more general propositions. The intersection $A \cap B$ corresponds to the logical conjunction '$A$ and $B$'; the union $A \cup B$ corresponds to the logical dis-junction '$A$ or $B$'; the complement $A^{\complement}$ corresponds to the logical negation 'not A'; and $A \supset B$ corresponds to the logical expression '$B$ implies $A$'. The empty set $\emptyset$ corresponds to a proposition that is certainly false, and the universal set $\Omega$ (aka the sampling space) to a proposition that is certainly true.

**Definition 1.** Let $A \precsim B$ denote the judgment that $A$ is not more likely to occur than $B$; $A \sim B$ denote the judgment that $A$ and $B$ are equally likely to occur; $A \prec B$ denote the judgment that $A$ is less likely to occur than $B$ given a common underlying (initial) information base.

Consider the following set of (reasonable) axioms:

**Axiom-LA1** For any $A, B$, only one of $A \prec B$, $B \prec A$, $A \sim B$ can occur.

**Axiom-LA2** If $A_1$, $A_2$, $B_1$, $B_2$ are events such that $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$, $A_1 \precsim B_1$, $A_2 \precsim B_2$ then $A_1 \cup A_2 \precsim B_1 \cup B_2$. Additionally, if either $A_1 \prec B_1$, $A_2 \prec B_2$ then $A_1 \cup A_2 \prec B_1 \cup B_2$.

**Axiom-LA3** For any $A$, $\emptyset \precsim A$. Furthermore, $\emptyset \prec \Omega$.

**Axiom-LA4** If $A_1 \supset A_2 \supset ...$ is a decreasing sequence of events with limit $\cap_{i=1}^{\infty} A_i$, and $B$ is some fixed event such that $A_i \succsim B$ for all $i = 1, 2...$ then $\cap_{i=1}^{\infty} A_i \succsim B$.

**Axiom-LA5** There exists a random variable $u \in [0, 1]$ such that if $A_1$ and $A_2$ are the events that $u$ falls in given sub-intervals of $[0, 1]$ with lengths $\ell_1$ and $\ell_2$ respectively, then $A_1 \precsim A_2$ if and only if $\ell_1 \leq \ell_2$.

(LA1) ensures that all events may be compared. (LA2) and (LA3) ensure that the comparisons are made in a logically consistent way, and simply reflect some obvious properties of any notion of 'not more probable than'. (LA4) is an assumption stronger than (LA2) and (LA3). (LA5), essentially defines the Uniform distribution, and allows the construction of the subjective probability suggesting how much more likely an event is.

### 2.2 Construction of probability

**Theorem 2.** *Let $\mu[a,b]$ denote the event that a random variable, from Axiom-LA5, lies in $[a,b]$. For any event $A \in \mathscr{F}$ satisfying the axioms LA1-LA5, there exists a unique number $a^\star \in [0,1]$ such that $A \sim \mu[0,a^\star]$*

**Definition 3.** Subjective probability (Your degree of believe) of an event $A \in \mathscr{F}$ satisfying the axioms LA1-LA5, we define the unique number $a^\star$ from Theorem 2. It is symbolized as $\mathsf{P}(A)$, and hence satisfies

$$A \sim \mu[0, \mathsf{P}(A)] \quad \text{, for all A}\in\mathscr{F}. \tag{1}$$

**Theorem 4.** *Let two events $A, B \in \mathscr{F}$ . Then $A \precsim B$* if and only if $\mathsf{P}(A) \leq \mathsf{P}(B)$

**Theorem 5.** *Given Axioms LA1-LA5, the quantity $\mathsf{P}(\cdot)$ (Definition 3) is the probability as it satisfies the usual probability axioms:*

**P1** $\mathsf{P}(A) \geq 0$ *and* $\mathsf{P}(\Omega) = 1$

**P2** *If $A \cap B = \emptyset$ then* $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B)$

**P3** *If $\{A_1, A_2, ...\}$ an infinite sequence of events such that $A_i \cap A_j = \emptyset$ for all $i, j$ then* $\mathsf{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i)$.

### 2.3 Extension to the conditional likelihoods & probability

Consider events $A, D \in \mathscr{F}$; $(A|D)$ corresponds to the conjunction '$A$ when $D$ is known'.

**Definition 6.** Given a common underlying (initial) information base, $(A|D) \precsim (B|D)$ denotes the judgment that $A$ is not more likely to occur than $B$, when $D$ is known.

Consider the following additional (reasonable) axiom needed to :

**Axiom-LA6** For any $A, B, D \in \mathscr{F}$, $(A|D) \precsim (B|D)$, if and only if $A \cap D \precsim B \cap D$.

**Theorem 7.** *Given [LA1] and [LA6], for any $A, B, D \in \mathscr{F}$, exactly one of the following three relations can occur: $(A|D) \prec (B|D)$, $(A|D) \succ (B|D)$, or $(A|D) \sim (B|D)$.*

The following theorem relates the standard probability to the conditional likelihoods.

**Theorem 8.** *If relations $(\prec, \succ, \sim)$ satisfy assumptions [LA1]-[LA5], and [LA6] then quantity* $\mathsf{P}$ *in Definition 3 is the unique probability distribution which has the property: For any $A, B, D \in \mathscr{F}$ such that $\mathsf{P}(D) > 0$,*

$$(A|D) \lesssim (B|D) \qquad \textit{if and only if} \qquad \mathsf{P}(A|D) \leq \mathsf{P}(B|D)$$

*Remark* 9. All subjective probabilities are conditional. We recognize some initial information $\Omega$. $\mathsf{P}(A|\Omega)$ expresses Your degree of belief about $A$ based on the totality of his current information $\Omega$. If You observe the occurrence of another event $B$, Your probability for $A$ becomes $\mathsf{P}(A|B,\Omega)$ because $B$ has been added to Your information. As $\Omega$ is common, it is convenient to simplify the notation by suppressing $\Omega$, and writing $\mathsf{P}(A) := \mathsf{P}(A|\Omega)$ and $\mathsf{P}(A|B) := \mathsf{P}(A|B,\Omega)$. In Bayesian paradigm language: the background information $\Omega$ may reflect all one knows before the collection of the data; $\mathsf{P}(A)$ is your prior information; $B$ is the new information from an experiment; $\mathsf{P}(A|B)$ is Your degree of believe about $A$ after observing $B$; and the Bayesian theorem is the mechanism performing this update.

## 3 The Bayesian model

Assume that a specific experiment $e \in \mathcal{E}$ has been performed; $\mathcal{E}$ denotes the family of potential experiments. Assume there is available a sequence of observations (or data) $y = (y_1, ..., y_n)$, where $y_i \in \mathcal{Y}$ for $i = 1, ..., n$, generated as outcomes of the performed experiment $e \in \mathcal{E}$.

Let $F(\cdot|\theta)$ denote the sampling distribution with PDF/PMF $f(y|\theta)$, which models the data generating process of the performed experiment $e$. For simplicity, we suppress conditioning on $e$ from $F(\cdot|\theta, e)$ although we shouldn't. $F(\cdot|\theta)$ is a statement of the Your subjective views about the data generating process.

**Definition 10.** A **parametric statistical model** consists of a sequence of observations $y = (y_1, ..., y_n)$ of a random variable $y$, and the sampling distribution $F(\cdot|\theta)$ where only the parameter $\theta \in \Theta$ is unknown.

- Symbol. $y \sim F(\cdot|\theta)$ or $(y, F(\cdot|\theta))$.

**Definition 11.** The **likelihood** $L(y|\theta)$ of $y$ and $e$ given $\theta$ is defined as $L(y|\theta) = f(y|\theta)$, and contains all the information available from the observed data $y$ and the experiment performed.

**Definition 12. Priori distribution** $\Pi(\theta)$ of the unknown parameter $\theta \in \Theta$, with PDF/PMF $\pi(\theta)$, quantifies Your believes or judgments about parameter $\theta$ before You perform the experiment and get the observations.

- symbol. $\theta \sim \Pi(\theta)$.

*Remark* 13. We say 'To account for the uncertainty of the unknown parameter $\theta \in \Theta$, we assign an a priori distribution $\Pi(\theta)$ with PDF/PMF $\pi(\theta)$'. It is specified in an entirely subjective manner based on Your judgments.

**Definition 14.** (Bayesian model) A Bayesian statistical model is made of a parametric statistical model, $y \sim F(\cdot|\theta)$, and a prior distribution (or prior model) $\theta \sim \Pi(\theta)$, which admits PDF/PMF $\pi(\theta)$, on the unknown parameters $\theta$. It is denoted as a hierarchical model:

$$\begin{cases} y|\theta & \sim F(y|\theta) \\ \theta & \sim \Pi(\theta) \end{cases} \tag{2}$$

*Remark* 15. Equation 2 implies that observations $y_{1:n}$ have been generated from a distribution $F(y|\theta)$ parameterized by $\theta$, where $\theta$ is a latent variable that follows a distribution $\Pi(\theta)$ prior to the generation of $y$.

**Definition 16.** The joint distribution $P(\theta, y)$ of $(\theta, y)$ can be defined from (2) as

$$\mathrm{d}P(\theta, y) = \mathrm{d}F(y|\theta)\mathrm{d}\Pi(\theta) \quad \text{ and has PDF/PMF } \quad p(\theta, y) = f(y|\theta)\pi(\theta)$$

**Definition 17.** The prior predictive distribution $F(y)$ of $y$ results by integrating out $F(y|\theta)$ with respect to $\Pi(\theta)$. We are interested in its PDF/PMF called marginal likelihood

$$f(y) = \int_\Theta f(y|\theta)\mathrm{d}\Pi(\theta) = \begin{cases} \int_\Theta f(y|\theta)\pi(\theta)\mathrm{d}\theta & \text{, if } \theta \text{ cont} \\ \sum_{\forall\theta\in\Theta} f(y|\theta)\pi(\theta) & \text{, if } \theta \text{ disc} \end{cases} \tag{3}$$

*Remark* 18. Prior predictive distribution $F(y)$ is how You modeled the unknown real data generation process $R(\cdot)$. It aims at representing the distribution according to which the observables $y$ occurred.

*Remark* 19. $f(y)$ is also called evidence of the Bayesian model as it can indicate whether the statistical model $F(\cdot|\theta)$ is useful. Unreasonably small $f(y)$ may indicate that it is unlikely that the Bayesian model could actually predict the accrued $y$ ; hence it is not representative to the real generating process $R(\cdot)$; hence it is not a useful model.

*Remark* 20. $f(y)$ is also called statistical evidence because:

**Definition 21.** The posterior distribution $\Pi(\theta|y)$ of $\theta \in \Theta$ given $y$ and $e$ is defined by the Bayes Theorem as

$$\mathrm{d}\Pi(\theta|y) = \frac{L(y|\theta)\mathrm{d}\Pi(\theta)}{\int_\Theta L(y|\theta)\mathrm{d}\Pi(\theta)} \quad \text{ and has PDF/PMF } \quad \pi(\theta|y) = \frac{L(y|\theta)\pi(\theta)}{\int_\Theta L(y|\theta)\mathrm{d}\Pi(\theta)} \tag{4}$$

*Note* 22. Posterior expectation of any integrate function $h(\cdot)$ defined on $\Theta$ is

$$\mathrm{E}_\Pi(h(\theta)|y) = \frac{\int_\Theta h(\theta)L(y|\theta)\mathrm{d}\Pi(\theta)}{\int_\Theta L(y|\theta)\mathrm{d}\Pi(\theta)}$$

*Note* 23. The posterior probability that $\theta \in A$ where $A \subseteq \Theta$ is

$$\Pi(\theta \in A|y) = \mathrm{E}_{\Pi}(1(\theta \in A)|y) = \frac{\int_A L(y|\theta)\mathrm{d}\Pi(\theta)}{\int_\Theta L(y|\theta)\mathrm{d}\Pi(\theta)}$$

*Remark* 24. The posterior distribution $\Pi(\theta|y)$ represents Your degree of believe about $\theta$ after You performed the experiment and saw the data $y = (y_1, ..., y_n)$ generated. Elaborating on (4), the Bayesian Theorem can be seen as a reasonable probabilistic mechanism to

- combine Your a priori information about $\theta$ (exclusively incorporated in $\Pi(\theta)$) and the experimental information about $\theta$ (exclusively incorporated in likelihood $L(y|\theta)$
- update Your degree of believe about $\theta$ from the a priori distribution $\pi(\theta)$ to the a posteriori distribution $\pi(\theta|y)$ in the light of new information from experiment.
- perform the inversion $(y|\theta) \mapsto (\theta|y)$ in a subjective probabilistic manner. Usually, the experiment, as described by the parametric model $F(y|\theta)$ , is a process there we observe the effect $y$ but we are interested in learning the unknown cause $\theta$ .

**Definition 25.** The predictive distribution of a sequence $z = (y_{n+1}, ..., y_{n+m})$ of $m$ future outcomes given a sequence of observations $y$ has PDF/PMF

$$f(z|y) = \mathrm{E}_{\Pi}(f(z|y,\theta)|y) = \int_\Theta f(z|y,\theta)\mathrm{d}\Pi(\theta|y)$$

It the case that $z$ and $y$ are conditionally independent given $\theta$; i.e. $f(z|y,\theta) = f(z|\theta)$ it is

$$g(z|y) = \mathrm{E}_{\Pi}(f(z|\theta)|y) = \int_\Theta f(z|\theta)\mathrm{d}\Pi(\theta|y)$$

*Remark* 26. Specification of the subjective probability should be the result of very careful weighing of all the available information, in order to avoid the case that the probability reflects a person's prejudices and without any scientific basis at all. For a given problem, usually the researcher specifies (i.) $F(y|\theta)$ and $\Pi(\theta)$ , or (ii.) $P(y,\theta)$ and derives the rest distributions, by probability calculations, since [1]

$$\mathrm{d}P(y,\theta) = \mathrm{d}F(y|\theta)\mathrm{d}\Pi(\theta) = \mathrm{d}\Pi(\theta|y)\mathrm{d}F(y)$$

**Definition 27.** If the likelihood of $y$ given $\theta$ is factorized as

$$L(y|\theta) = k(y|\theta)\rho(y),$$

then $k(y|\theta)$ is called kernel of the likelihood of $y$ given $\theta$, and $\rho(y)$ is called residue of this likelihood.

**Definition 28.** If a PDF/PMF $\pi(\theta)$ can be written as

$$\pi(\theta) = \frac{K(\theta)}{\int K(\theta)\mathrm{d}\theta},$$

then then $K(\theta)$ is called kernel of the density $\pi(\theta)$. We can write $\pi(\theta) \propto K(\theta)$ as the normalizing constant $\int K(\theta)\mathrm{d}\theta$ is implied by the specification of $K(\theta)$.

[1] Notation $\mathrm{d}P(y,\theta) = \mathrm{d}F(y|\theta)\mathrm{d}\Pi(\theta) = \mathrm{d}\Pi(\theta|y)\mathrm{d}F(y)$ is used as abbreviation to

$$\int h(y,\theta)\mathrm{d}P(y,\theta) = \int h(y,\theta)\mathrm{d}F(y|\theta)\mathrm{d}\Pi(\theta) = \int h(y,\theta)\mathrm{d}\Pi(\theta|y)\mathrm{d}F(y)$$

for any measurable function $h$ defined on $\mathcal{Y} \times \Theta$.

**Example 29.** [Bernoulli-Beta model] Let $e$ be a Bernoulli experiment with unknown probability of success $\theta \in [0,1]$. We specify a Bayesian model

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

where $y = (y_1, ..., y_n)$, $\theta \in [0,1]$, and $a > 0$, $b > 0$ are known (fixed) hyper-parameters.

1. State the sampling distribution of $y$, the likelihood of $y$ given $\theta$, the prior distribution of $\theta$, and the joint PMF/PDF of $(y, \theta)$.
2. Calculate the marginal PMF of $y$
3. Calculate the posterior PDF of $\theta$ given $y$ and recognize the distribution family.
4. Calculate the predictive PMF of a sequence of future outcomes $z = (z_1, .., z_m)$ given $y$.

**Hint:** Consider PDF/PMF:

$$f_{\text{Br}(\theta)}(y) = \theta^y (1-\theta)^{1-y} 1(y \in \{0,1\})\,; \qquad \pi_{\text{Be}(a,b)}(\theta) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} 1(\theta \in [0,1])$$

**Solution.**

1. The sampling distribution is a Bernoulli distribution with parameter $\theta$; i.e. $\text{Br}(\theta)$. The Likelihood function is

$$L(y|\theta) = f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}.$$

   The prior is the Beta distribution with parameters $a > 0$ and $b > 0$; i.e. $\theta \sim \text{Be}(a, b)$:

$$\pi(\theta) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} 1(\theta \in [0,1])$$

   The joint PMF/PDF of $(y_{1:n}, \theta)$ is

$$\begin{aligned} p(y, \theta) = f(y|\theta)\pi(\theta) &= \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta) \\ &= \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} 1_{\{0,1\}^n}(y) \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} 1_{[0,1]}(\theta) \\ &= \frac{1}{B(a,b)} \theta^{\sum_{i=1}^n y_i + a - 1}(1-\theta)^{n-\sum_{i=1}^n y_i + b - 1} 1_{\{0,1\}^n}(y) 1_{[0,1]}(\theta) \end{aligned}$$

2. The marginal likelihood of the observations $y_{1:n}$ is

$$\begin{aligned} f(y) = \int f(y|\theta)\pi(\theta) \text{d}\theta &= \frac{1}{B(a,b)} \int_0^1 \theta^{\sum_{i=1}^n y_i + a - 1}(1-\theta)^{n-\sum_{i=1}^n y_i + b - 1} \text{d}\theta \ 1_{\{0,1\}^n}(y) \\ = \frac{1}{B(a,b)} B\left(\sum_{i=1}^n y_i + a, n - \sum_{i=1}^n y_i + b\right) 1_{\{0,1\}^n}(y) \end{aligned}$$

3. The posterior of $\theta$ given the observations $y_{1:n}$ has PDF

$$\begin{aligned} \pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int_\Theta f(y|\theta)\text{d}\Pi(\theta)} &\propto f(y|\theta)\pi(\theta) = \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta|a,b) \\ &= \theta^{\sum_{i=1}^n y_i + a - 1}(1-\theta)^{n-\sum_{i=1}^n y_i + b - 1} \propto \text{Be}(\theta|a^*, b^*) \end{aligned}$$

where $a^* = \sum_{i=1}^n y_i + a$ , $b^* = n - \sum_{i=1}^n y_i + b$. Hence the posterior distribution of $\theta$ is a Beta distribution with parameters $a^* = \sum_{i=1}^n y_i + a$ , and $b^* = n - \sum_{i=1}^n y_i + b$.

4. The predictive distribution of $z_{1:m}$ from parametric model $\text{Br}(\theta)$ given $y_{1:n}$ has PMF

$$\begin{aligned} g(z|y) &= \int_\Theta f(z|\theta)\pi(\theta|y)\mathrm{d}\theta \ = \int_\Theta \prod_{i=1}^m f(z_i|\theta)\pi(\theta|y)\mathrm{d}\theta \\ &= \int_0^1 \left[\theta^{\sum_{i=1}^m z_i}(1-\theta)^{m-\sum_{i=1}^m z_i} 1\left(z \in \{0,1\}^m\right)\right]\left[\frac{1}{B(a^*,b^*)}\theta^{a^*-1}(1-\theta)^{b^*-1}\right]\mathrm{d}\theta \\ &= \frac{1}{B(a^*,b^*)}\int_0^1 \theta^{\sum_{i=1}^m z_i + a^* - 1}(1-\theta)^{m-\sum_{i=1}^m z_i + b^* - 1}\mathrm{d}\theta \ 1\left(z \in \{0,1\}^m\right) \\ &= \frac{1}{B(a^*,b^*)} B\left(\sum_{i=1}^m z_i + a^*, m - \sum_{i=1}^m z_i + b^*\right) 1\left(z \in \{0,1\}^m\right) \end{aligned}$$

## 4 Sequential processing of data via Bayes theorem

*Note* 30. Bayesian paradigm enjoys a coherence property according to which updating the prior one observation at a time, or all observations together does not matter, it leads to the same posterior inference.

*Note* 31. Let $y = (y_1, y_2)$ be a partition of the observables.

- Consider the Learning Procedure 1 for $\theta$ where the Prior $\Pi(\theta)$ is updated to a posterior $\Pi(\theta|y_1, y_2)$ in the light of the full data $(y_1, y_2)$ observed at once; namely

$$\text{Learning Procedure 1: } \Pi(\theta) \overset{\substack{(y_1,y_2)\\ \longmapsto \\ f(y_1,y_2|\theta)}}{} \Pi(\theta|y_1, y_2)$$

  where

$$\Pi(\theta|y_1, y_2) \quad \text{with pdf } \pi(\theta|y_1, y_2) = \frac{f(y_1, y_2|\theta)\pi(\theta)}{f(y_1, y_2)}; \text{ where } f(y) = \int_\Theta f(y|\theta)\mathrm{d}\Pi(\theta)$$

- Consider Learning Procedure 2 for $\theta$ where at first stage the prior $\Pi(\theta)$ is updated to a posterior $\Pi(\theta|y_1)$ in the light of data $y_1$ and at second stage $\Pi(\theta|y_1)$ is updated to $\Pi(\theta|y_1, y_2)$ in the light of new data $y_2$. Here, $\Pi(\theta|y_1)$ is the distribution of $\theta$ posterior to observing $y_1$and prior to observing $y_2$! Similarly the likelihood should be conditional on all the observables incorporated so far as $f(y_2|y_1, \theta)$. Learning Procedure 2 for $\theta$ is

$$\text{Learning Procedure 2: } \Pi(\theta) \overset{\substack{y_1\\ \longmapsto \\ f(y_1|\theta)}}{} \Pi'(\theta|y_1) \overset{\substack{y_2\\ \longmapsto \\ f(y_2|y_2,\theta)}}{} \Pi'(\theta|y_1, y_2)$$

  with pdf/pmfs

$$\begin{aligned} \pi(\theta|y_1) &= \frac{f(y_1|\theta)\pi(\theta)}{f(y_1)}; \text{ where } f(y_1) = \int_\Theta f(y_1|\theta)\mathrm{d}\Pi(\theta) \\ \pi'(\theta|y_1, y_2) &= \frac{f(y_2|y_1,\theta)\pi(\theta|y_1)}{f(y_2|y_1)} = \frac{f(y_2|y_1,\theta)\frac{f(y_1|\theta)\pi(\theta)}{f(y_1)}}{f(y_2|y_1)} = \frac{f(y_2|y_1,\theta)f(y_1|\theta)\pi(\theta)}{f(y_2|y_1)f(y_1)} \\ &= \frac{f(y_1, y_2|,\theta)\pi(\theta)}{f(y_1, y_2)} = \pi(\theta|y_1, y_2) \end{aligned}$$

- We observe the two Learning Scenarios are equivalent in the sense that they lead to the same posterior $\Pi(\theta|y_1, y_2)$ at the end.

*Note* 32. By induction, this result is extended to the case where several data are collected sequentially as $y_1, y_2, y_3, y_4, ...$

## 5 Proper & improper priors

*Note* 33. Priors do not necessarily need to be probability distributions but they need to lead to posterior probability distributions.

**Definition 34.** The prior $\Pi(\theta)$ with pdf/pmf $\pi(\theta) > 0$ for $\theta \in \Theta$, is called proper prior if

$$\int_{\Theta} \pi(\theta)\mathrm{d}\theta < \infty \text{ when } \theta \text{ is continuous;} \qquad \text{and} \qquad \sum_{\forall \theta \in \Theta} \pi(\theta) < \infty, \text{ when } \theta \text{ is discrete}$$

and hence it is a probability distribution ; and improper prior if

$$\int_{\Theta} \pi(\theta)\mathrm{d}\theta = \infty \text{ when } \theta \text{ is continuous;} \qquad \text{and} \qquad \sum_{\forall \theta \in \Theta} \pi(\theta) = \infty, \text{ when } \theta \text{ is discrete}$$

and hence it is a not a probability distribution.

*Note* 35. An improper prior $\Pi(\theta)$ can only be used for inference if it leads to a well defined posterior probability distribution (aka proper posterior); namely if the 'Properness condition'

$$\int_{\Theta} f(y|\theta)\pi(\theta)\mathrm{d}\theta < \infty \tag{5}$$

is satisfied for the observable sequence $y$ at hand. If (5) is not satisfied, posterior quantities like mean, median, variance have no meaning.

*Note* 36. Improper priors are often used in the Objective Bayes framework or in the Subjective Bayes framework (as a last resort) in order to express ignorance or invariance about certain characteristics of the population (more details in later lectures.)

**Proposition 37.** *If the sampling distribution $F(\cdot|\theta)$ is discrete and the prior $\Pi(\theta)$ is proper, then the posterior $\Pi(\theta|y)$ is always proper.*

*Proof.* Provided as an Exercise 28 in the Exercise sheet. □

**Proposition 38.** *If the sampling distribution $F(\cdot|\theta)$ is continuous and the prior $\Pi(\theta)$ is proper, then the posterior $\Pi(\theta|y)$ is almost always proper.*

*Proof.* Provided as an Exercise 29 in the Exercise sheet. □

**Example 39.** Consider a sequence of observables $y = (y_1, ..., y_n)$ generated as $y_i|\mu \overset{\text{iid}}{\sim} \mathrm{N}(\mu, 1)$, for $i = 1, ..., n$, and assign $\mu$ a prior density such as $\pi(\mu) \propto 1(\mu \in \mathbb{R})$. Find if prior $\pi(\mu)$ can be used as a prior for Bayesian inference.

**Solution.** This is an improper prior because $\int_{\mathbb{R}} \pi(\mu)\mathrm{d}\mu = \int_{\mathbb{R}} 1\mathrm{d}\mu = +\infty$. Because $\mu$ is continuous and $\pi(\mu)$ improper prior, I need to check if the Properness Condition (Note 35) is satisfied and hence the Bayesian model can lead to valid Bayesian inference. The Properness Condition (Note 35) is satisfied, i.e.

$$\int_{\mathbb{R}} f(y|\mu)\pi(\mu)\mathrm{d}\mu = \int_{\mathbb{R}} \prod_{i=1}^{n} \mathrm{N}(y_i|\mu, 1)1(\mu \in \mathbb{R})\mathrm{d}\mu \propto 2^{-\frac{n}{2}}(\pi)^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2}(\sum y_i^2) + \frac{1}{2}(\sum y_i)^2\right) < \infty$$

Hence I can use Bayesian model $\begin{cases} y_i|\mu \overset{\text{iid}}{\sim} \mathrm{N}(\mu, 1), \ i = 1, ..., n \\ \mu \sim 1(\mu \in \mathbb{R})\mathrm{d}\mu \end{cases}$ with improper prior to perform Bayesian inference.

## 6 Practice

**Question 40.** *Feel free to work on the Exercises 24, 30, 33, in the Exercise sheet.*