Statistics for Applications

Chapter 8: Bayesian Statistics
The Bayesian approach (1)

- So far, we have studied the frequentist approach of statistics.

- The frequentist approach:
  - Observe data
  - These data were generated randomly (by Nature, by measurements, by designing a survey, etc...)
  - We made assumptions on the generating process (e.g., i.i.d., Gaussian data, smooth density, linear regression function, etc...)
  - The generating process was associated to some object of interest (e.g., a parameter, a density, etc...)
  - This object was unknown but fixed and we wanted to find it: we either estimated it or tested a hypothesis about this object, etc...
The Bayesian approach (2)

- Now, we still observe data, assumed to be randomly generated by some process. Under some assumptions (e.g., parametric distribution), this process is associated with some fixed object.

- We have a **prior belief** about it.

- Using the data, we want to update that belief and transform it into a **posterior belief**.
The Bayesian approach (3)

Example

- Let \( p \) be the proportion of woman in the population.

- Sample \( n \) people randomly with replacement in the population and denote by \( X_1, \ldots, X_n \) their gender (1 for woman, 0 otherwise).

- In the frequentist approach, we estimated \( p \) (using the MLE), we constructed some confidence interval for \( p \), we did hypothesis testing (e.g., \( H_0 : p = .5 \) v.s. \( H_1 : p \neq .5 \)).

- Before analyzing the data, we may believe that \( p \) is likely to be close to \( 1/2 \).

- The Bayesian approach is a tool to:
  1. include mathematically our prior belief in statistical procedures.
  2. update our prior belief using the data.
The Bayesian approach (4)

Example (continued)

- Our prior belief about $p$ can be quantified:

- E.g., we are 90% sure that $p$ is between .4 and .6, 95% that it is between .3 and .8, etc...

- Hence, we can model our prior belief using a distribution for $p$, as if $p$ was random.

- In reality, the true parameter is not random! However, the Bayesian approach is a way of modeling our belief about the parameter by doing as if it was random.

- E.g., $p \sim \mathcal{B}(a, a)$ (*Beta distribution*) for some $a > 0$.

- This distribution is called the *prior distribution*. 
The Bayesian approach (5)

Example (continued)

▶ In our statistical experiment, $X_1, \ldots, X_n$ are assumed to be i.i.d. Bernoulli r.v. with parameter $p$ conditionally on $p$.

▶ After observing the available sample $X_1, \ldots, X_n$, we can update our belief about $p$ by taking its distribution conditionally on the data.

▶ The distribution of $p$ conditionally on the data is called the \textit{posterior distribution}.

▶ Here, the posterior distribution is

\[
\mathcal{B} \left( a + \sum_{i=1}^{n} X_i, a + n - \sum_{i=1}^{n} X_i \right) .
\]
The Bayes rule and the posterior distribution (1)

- Consider a probability distribution on a parameter space $\Theta$ with some pdf $\pi(\cdot)$: the prior distribution.

- Let $X_1, \ldots, X_n$ be a sample of $n$ random variables.

- Denote by $p_n(\cdot|\theta)$ the joint pdf of $X_1, \ldots, X_n$ conditionally on $\theta$, where $\theta \sim \pi$.

- Usually, one assumes that $X_1, \ldots, X_n$ are i.i.d. conditionally on $\theta$.

- The conditional distribution of $\theta$ given $X_1, \ldots, X_n$ is called the posterior distribution. Denote by $\pi(\cdot|X_1, \ldots, X_n)$ its pdf.
Bayes’ formula states that:

\[ \pi(\theta|X_1, \ldots, X_n) \propto \pi(\theta)p_n(X_1, \ldots, X_n|\theta), \quad \forall \theta \in \Theta. \]

The constant does not depend on \( \theta \):

\[ \pi(\theta|X_1, \ldots, X_n) = \frac{\pi(\theta)p_n(X_1, \ldots, X_n|\theta)}{\int_{\Theta} p_n(X_1, \ldots, X_n|t) \, d\pi(t)}, \quad \forall \theta \in \Theta. \]
The Bayes rule and the posterior distribution (3)

In the previous example:

- \( \pi(p) \propto p^{a-1}(1-p)^{a-1}, p \in (0, 1). \)

- Given \( p, X_1, \ldots, X_n \) i.i.d. \( Ber(p) \), so
  \[
  p_n(X_1, \ldots, X_n|\theta) = p^{\sum_{i=1}^{n} X_i} (1-p)^{n-\sum_{i=1}^{n} X_i}.
  \]

- Hence,
  \[
  \pi(\theta|X_1, \ldots, X_n) \propto p^{a-1+\sum_{i=1}^{n} X_i} (1-p)^{a-1+n-\sum_{i=1}^{n} X_i}.
  \]

- The posterior distribution is
  \[
  \mathcal{B} \ a + \sum_{i=1}^{n} X_i, \ a + n - \sum_{i=1}^{n} X_i.
  \]
Non informative priors (1)

- Idea: In case of ignorance, or of lack of prior information, one may want to use a prior that is as little informative as possible.

- Good candidate: $\pi(\theta) \propto 1$, i.e., constant pdf on $\Theta$.

- If $\Theta$ is bounded, this is the uniform prior on $\Theta$.

- If $\Theta$ is unbounded, this does not define a proper pdf on $\Theta$!

- An *improper prior* on $\Theta$ is a measurable, nonnegative function $\pi(\cdot)$ defined on $\Theta$ that is not integrable.

- In general, one can still define a posterior distribution using an improper prior, using Bayes’ formula.
Non informative priors (2)

Examples:

- If $p \sim U(0, 1)$ and given $p$, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Ber(p)$:

  $\pi(p|X_1, \ldots, X_n) \propto p^{\sum_{i=1}^n X_i} (1 - p)^{n - \sum_{i=1}^n X_i}$,

  i.e., the posterior distribution is

  $\mathcal{B} \left( 1 + \sum_{i=1}^n X_i, 1 + n - \sum_{i=1}^n X_i \right)$.

- If $\pi(\theta) = 1, \forall \theta \in \mathbb{R}$ and given $\theta$, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$:

  $\pi(\theta|X_1, \ldots, X_n) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^n (X_i - \theta)^2\right)$,

  i.e., the posterior distribution is

  $\mathcal{N}\left(\bar{X}_n, \frac{1}{n}\right)$. 
Non informative priors (3)

- **Jeffreys prior:**
  \[
  \pi_J(\theta) \propto \sqrt{\det I(\theta)},
  \]
  where \(I(\theta)\) is the Fisher information matrix of the statistical model associated with \(X_1, \ldots, X_n\) in the frequentist approach (provided it exists).

- In the previous examples:
  - Ex. 1: \(\pi_J(p) \propto \frac{1}{\sqrt{p(1-p)}}\), \(p \in (0, 1)\): the prior is \(B(1/2, 1/2)\).
  - Ex. 2: \(\pi_J(\theta) \propto 1\), \(\theta \in \mathbb{R}\) is an improper prior.
Jeffreys prior satisfies a reparametrization invariance principle: If $\eta$ is a reparametrization of $\theta$ (i.e., $\eta = \phi(\theta)$ for some one-to-one map $\phi$), then the pdf $\tilde{\pi}(\cdot)$ of $\eta$ satisfies:

$$\tilde{\pi}(\eta) \propto \sqrt{\det \tilde{I}(\eta)},$$

where $\tilde{I}(\eta)$ is the Fisher information of the statistical model parametrized by $\eta$ instead of $\theta$. 

Non informative priors (4)
For $\alpha \in (0, 1)$, a Bayesian confidence region with level $\alpha$ is a random subset $\mathcal{R}$ of the parameter space $\Theta$, which depends on the sample $X_1, \ldots, X_n$, such that:

$$\Pr[\theta \in \mathcal{R} | X_1, \ldots, X_n] = 1 - \alpha.$$ 

Note that $\mathcal{R}$ depends on the prior $\pi(\cdot)$.

"Bayesian confidence region" and "confidence interval" are two distinct notions.
Bayesian estimation (1)

- The Bayesian framework can also be used to estimate the true underlying parameter (hence, in a frequentist approach).

- In this case, the prior distribution does not reflect a prior belief: It is just an artificial tool used in order to define a new class of estimators.

- **Back to the frequentist approach:** The sample $X_1, \ldots, X_n$ is associated with a statistical model $(E, \{P_\theta\}_{\theta \in \Theta})$.

- Define a distribution (that can be improper) with pdf $\pi$ on the parameter space $\Theta$.

- Compute the posterior pdf $\pi(\cdot|X_1, \ldots, X_n)$ associated with $\pi$, seen as a prior distribution.
Bayesian estimation (2)

- **Bayes estimator:**

  \[
  \hat{\theta}^{(\pi)} = \int_{\Theta} \theta \, d\pi(\theta|X_1, \ldots, X_n): 
  \]

  This is the *posterior mean*.

- The Bayesian estimator depends on the choice of the prior distribution \(\pi\) (hence the superscript \(\pi\)).
Bayesian estimation (3)

In the previous examples:

- Ex. 1 with prior $\mathcal{B}(a, a)$ ($a > 0$):
  
  $$
  \hat{p}^{(\pi)} = \frac{a + \sum_{i=1}^{n} X_i}{2a + n} = \frac{a/n + X_n}{2a/n + 1}.
  $$

  In particular, for $a = 1/2$ (Jeffreys prior),
  
  $$
  \hat{p}^{(\pi, J)} = \frac{1/(2n) + X_n}{1/n + 1}.
  $$

- Ex. 2: $\hat{\theta}^{(\pi, J)} = X_n$.

- In each of these examples, the Bayes estimator is consistent and asymptotically normal.

- In general, the asymptotic properties of the Bayes estimator do not depend on the choice of the prior.