Stat 535 C - CPSC 540

Statistical Computing & Monte Carlo Methods

Lecture 2 - Revised version

Arnaud Doucet

Email: arnaud@cs.ubc.ca

- Slides available on the Web before lectures:

`www.cs.ubc.ca/~arnaud/stat535.html`

- Textbook: C.P. Robert & G. Casella, *Monte Carlo Statistical Methods*, Springer, 2nd Edition.

- Additional lecture notes available on the Web.

- Textbooks which might also be of help:

  - A. Gelman, J.B. Carlin, H. Stern and D.B. Rubin, *Bayesian Data Analysis*, Chapman&Hall/CRC, 2nd edition.
  - C.P. Robert, *The Bayesian Choice*, Springer, 2nd edition.

2.1– Outline

- Preliminaries,
- The sufficiency principle.
- The likelihood principle.
- The conditionality principle.

3.1– Preliminaries

• Main objective of statistical theory: Derive from observations of a random phenomenon an inference about the probability distribution underlying this phenomenon.

• In this course, we only consider parametric modelling.
The observations x are the realization of a random variable X of probability density function $f\left(x \middle| \theta\right)$ where

- θ is *unknown* and belongs to a space Θ of finite dimension.
- the functional form $f\left(x \middle| \theta\right)$ is known.

• The function $f\left(x|\,\theta\right)$ considered as a function of θ for a fixed realization of the observation $X = x$ is called the likelihood function.

• Dependent on the authors one writes

$$l\left(\theta|\,x\right) = f\left(x|\,\theta\right)$$

or even

$$l\left(\theta\right) = f\left(x|\,\theta\right)$$

to emphasize that the observations are fixed. The second notation should be avoided in a Bayesian context.

• **Example**: Consider a radioactive material with unknown half-life $\theta = H$. For a given atom, the time before desintegration is an exponential distribution of parameter $\log 2/H$.

• Most of the time, statistical modelling only approximates the reality thus losing part of its richness but gaining in efficiency.

• **Example**: Price and salary variations are closely related.
We can assume the following model

$$\Delta P = a + b\Delta S + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

where the data are $(\Delta P, \Delta S)$ and $\theta = \left(a, b, \sigma^2\right)$.

• The reductive effect can be sought as it partly removes unimportant perturbations of the phenomenon.

• **Example**: Consider the problem of forest fires. Determining the probability p of fire as a function of ecological and meteorological factors could be useful. It could be model through say

$$p = \frac{\exp\left(\beta_1 h + \beta_2 t + \beta_3 x\right)}{1 + \exp\left(\beta_1 h + \beta_2 t + \beta_3 x\right)}$$

where $\theta = (\beta_1, \beta_2, \beta_3)$ and

h is the humidity rate

t the average temperature

x the degree of management

• Data modelled as Bernoulli r.v.s. of parameter p.

3.1– Preliminaries

• An alternative approach consists of incorporating as much as possible the complexity of a phenomenon, and thus aims at estimating the distribution underlying the phenomenon under minimal assumptions, generally using functional estimation (density, regression function, etc.).

• The parametric approach is (in my opinion!) more pragmatic. It takes into account that a finite number of observations can efficiently estimate only a finite number of parameters.

• In any case, model checking/assessment or model choice should be considered.

- When $X \sim f\left(x|\,\theta\right)$, a function T of X (also called a statistic) is said to be sufficient if the distribution of X conditional upon $T\left(X\right)$ is independent of θ.

- Example: Let $X = (X_1, ..., X_n)$ i.i.d. from $\mathcal{N}\left(\mu, \sigma^2\right)$ with $\theta = \left(\mu, \sigma^2\right)$ then

$$f\left(x|\,\theta\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right)$$

• In this case,

$$f\left(x|\,\theta\right) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} - \frac{\mu \sum_{i=1}^n x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right)$$

$f\left(x|\,\theta\right)$ only depends on x through $\left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i\right)$ so $T\left(x\right) = \left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i\right)$ is a set of sufficient statistics.

• Note that $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $s^2 = \sum_{i=1}^n \left(x_i - \overline{x}\right)^2$ is also a set of sufficient statistics because

$$\sum_{i=1}^n x_i^2 = s^2 - n\overline{x}^2$$

so we can rewrite

$$f\left(x|\,\theta\right) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{\left(s^2 - n\overline{x}^2\right)}{2\sigma^2} - \frac{\mu\overline{x}}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right)$$

and $f\left(x|\,\theta\right)$ only depends on x through $\overline{x}$ and s^2.

4.1– Sufficiency principle

• Consider the independent binomial rvs $X_1 \sim \mathcal{B}(n_1, p)$, $X_2 \sim \mathcal{B}(n_2, p)$, $X_3 \sim \mathcal{B}(n_3, p)$ where n_1, n_2 and n_3 are known. Then

$$f(x_1, x_2, x_3 | p) = \begin{pmatrix} n_1 \\ \\ x_1 \end{pmatrix} \begin{pmatrix} n_2 \\ \\ x_2 \end{pmatrix} \begin{pmatrix} n_3 \\ \\ x_3 \end{pmatrix} p^{x_1+x_2+x_3} (1-p)^{n_1+n_2+n_3-x_1-x_2-x_3}$$

and the statistics

$$T_1(x_1, x_2, x_3) = x_1 + x_2 + x_3 \text{ or } T_2(x_1, x_2, x_3) = \frac{x_1 + x_2 + x_3}{n_1 + n_2 + n_3}$$

are sufficients because $f(x_1, x_2, x_3 | p)$ only depend on (x_1, x_2, x_3) through $T_1(x_1, x_2, x_3)$ or $T_2(x_1, x_2, x_3)$ but $\frac{x_1}{n_1} + \frac{x_2}{n_2} + \frac{x_3}{n_3}$ is not sufficient.

• Let $X = (X_1, ..., X_n)$ i.i.d. from $\mathcal{U}(0, \theta)$ of density $f(x_i|\theta) = \theta^{-1}\mathbf{1}_{[0,\theta]}(x_i)$. Then

$$l(\theta|x) = f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n}\mathbf{1}_{[\max\{x_i\},\infty)}(\theta).$$

$\Rightarrow$ The statistic $T(X) = \max\{X_i\}$ is sufficient.

• Let $X = (X_1, ..., X_n)$ i.i.d. from $\mathcal{P}(\theta)$ of distribution $f(x_i|\theta) = e^{-\theta}\frac{\theta^x}{x!}$. Then

$$l(\theta|x) = f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{e^{-n\theta}}{\prod_{i=1}^{n} x_i!}\theta^{\sum_{i=1}^{n} x_i}.$$

$\Rightarrow$ The statistics $T(X) = \sum_{i=1}^{n} X_i$ is sufficient.

- **Sufficiency principle**: Two observations x and y such that $T(x) = T(y)$ must lead to the same inference on θ.

- Consider the model $X_i \sim \mathcal{N}(\mu, 1)$ and we want to estimate μ based on n data. In this case the sufficient statistic is $T(x_{1:n}) = \sum_{i=1}^{n} x_i$.

- Consider the estimate $\widehat{\mu}_1 = \frac{1}{n} T(x_{1:n})$, then this estimate satisfies the sufficiency principle because if I have another dataset $x'_{1:n}$ such that $T(x_{1:n}) = T(x'_{1:n})$ then I obtain $\widehat{\mu}_2 = \frac{1}{n} T(x'_{1:n}) = \frac{1}{n} T(x_{1:n}) = \widehat{\mu}_1$.

- The estimate $\widehat{\mu}_1 = x_1$ does not satisfies the sufficiency principle for $n > 1$ because even if I have another dataset $x'_{1:n}$ such that $T(x_{1:n}) = T(x'_{1:n})$, then $\widehat{\mu}_2 = x'_1 \neq \widehat{\mu}_1$ if $x_1 \neq x_2$.

• The Sufficiently principle is generally accepted by most statisticians because of the Rao-Blackwell theorem.

• **Rao-Blackwell theorem**. Let $\delta(X)$ be an unbiased estimate of θ and $\delta_{RB}(X) = \mathbb{E}[\delta(X) | T(X)]$ then $\delta_{RB}(X)$ is unbiased and

$$var[\delta_{RB}(X)] \leq var[\delta(X)]$$

Proof: $$\begin{aligned} var[\delta(X)] &= \mathbb{E}[var[\delta(X)| T(X)]] + var[\mathbb{E}[\delta(X)| T(X)]] \\ &= \mathbb{E}[var[\delta(X)| T(X)]] + var[\delta_{RB}(X)]. \end{aligned}$$

4.2– Variance Decomposition Identity

If (X, Y) are two scalar random variables then we have

$$var(X) = E(var(X|Y)) + var(E(X|Y)).$$

Proof:

$$\begin{aligned}
var(X) &= E\left(X^2\right) - E(X)^2 \\
&= E\left(E\left(X^2\middle|Y\right)\right) - (E(E(X|Y)))^2 \\
&= E\left(E\left(X^2\middle|Y\right)\right) - E\left((E(X|Y))^2\right) \\
&\quad + E\left((E(X|Y))^2\right) - (E(E(X|Y)))^2 \\
&= E(var(X|Y)) + var(E(X|Y)).
\end{aligned}$$

5.1– The Likelihood Principle

• **Likelihood Principle**. The information brought by an observation x about θ is entirely contained in the likelihood function $l\left(\theta \mid x\right) = f\left(x \mid \theta\right)$. Moreover, two likelihood functions contain the same information about θ if they are proportional to each other; i.e.

$$l_1\left(\theta \mid x\right) = c\left(x\right) l_2\left(\theta \mid x\right)$$

• The maximum likelihood procedure does satisfy the likelihood principle because

$$\arg\max_{\theta}\, l_1\left(\theta \mid x\right) = \arg\max_{\theta}\, l_2\left(\theta \mid x\right)$$

if $l_1\left(\theta \mid x\right) = c\left(x\right) l_2\left(\theta \mid x\right)$.

• Classical approaches do not necessarily satisfy the likelihood principle.

• **Testing Fairness**. Suppose we want to test θ, the unknown probability of heads for possibly biased coin. Suppose

$$H_0 : \theta = \frac{1}{2} \quad \text{v.s.} \quad H_1 : \theta > \frac{1}{2}.$$

• **Scenario 1**: Number of flips $n = 12$ predetermined and number of heads $X \sim \mathcal{B}(n, \theta)$; that is if we collect $x = 9$ heads

$$P_\theta (X = x) = f(x | \theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} = \begin{pmatrix} 12 \\ 9 \end{pmatrix} \theta^9 (1-\theta)^3 = 220.\theta^9 (1-\theta)^3 .$$

For a frequentist, the p-value of the test is $P_\theta (X \geq 9 | H_0) = 0.073$ and H_0 is not rejected at level $\alpha = 0.05$.

• **Scenario 2**: Number of tails $\alpha = 3$ is predetermined, i.e. the flipping is continued until 3 tails are observed. Then $X \sim \mathcal{NB}(3, 1-\theta)$ and assuming we collected $x = 9$ heads then

$$P_\theta (X = x) = f(x|\theta) = \begin{pmatrix} \alpha + x - 1 \\ \\ \alpha - 1 \end{pmatrix} (1-\theta)^\alpha [1 - (1-\theta)]^x = 55.\theta^9 (1-\theta)^3 .$$

For a frequentist, the p-value of the test is $P_\theta(X \geq 9 | H_0) = 0.0327$ and H_0 is rejected at level $\alpha = 0.05$.

• The likelihood principle is here violated because in both cases

$$f(x|\theta) \propto \theta^9 (1-\theta)^3 .$$

5.2– Stopping rule Principle

• A direct implication of the likelihood principle is the stopping rule principe in sequential analysis.

• Consider a sequence of experiments that leads at time i to the observation $X_i \sim f\left(x_i\right|\theta)$ and we stops collecting data if at time n we have $(X_1, ..., X_n) \in A_n$; e.g. $A_n = \{X_1, ..., X_n : X_n > B\}$.
In this case

$$l\left(\theta\right| x_1, ..., x_n) \propto \prod_{i=1}^{n} f\left(x_i\right|\theta) \mathbf{1}_{A_n}\left(x_1, ..., x_n\right).$$

• **Stopping rule principle**: If a sequence of experiments is directed by a stopping rule which indicates when the experiments should stop, inference about θ must depend on the stopping rule only through the sample.

5.3– More p-values

• Consider the case where $X_i \sim \mathcal{N}(\theta, 1)$ and the hypothesis to be tested is $H_0 : \theta = 0$.

• The classical Neyman-Pearson test procedure at level 5% is to reject the hypothesis if $\frac{1}{n} \left|\sum_{i=1}^n x_i\right| > \frac{1.96}{\sqrt{n}}$ on the basis that

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \theta\right| \geq \frac{1.96}{\sqrt{n}}\middle| H_0\right) = \Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq \frac{1.96}{\sqrt{n}}\middle| H_0\right) = 0.05$$

• That is the decision is based on the event $\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq 1.96$ rather than on the observations themselves (conditioning by this value is impossible using frequentist theory).

• The frequency argument is that in 5% of the cases when H_0 is true, it rejects wrongly the null hypothesis.

• The stopping rule principle is definitely incompatible with frequentist modelling.

• Consider $X_i \sim \mathcal{N}(\theta, 1)$ and the hypothesis to be tested is $H_0 : \theta = 0$ and we stop collecting data at the first time n such that

$$\frac{1}{n} \left| \sum_{i=1}^{n} x_i \right| > \frac{1.96}{\sqrt{n}}.$$

• The resulting sample will always reject $H_0 : \theta = 0$ at the level 5%.

• Consider X_1, X_2 i.i.d. $\mathcal{N}(\theta, 1)$. The likelihood function is

$$l(\theta|\, x_1, x_2) = f(x_1, x_2|\, \theta) \propto \exp\left(-\left(\frac{x_1 + x_2}{2} - \theta\right)^2\right).$$

Now consider the alternative distribution

$$g(x_1, x_2|\, \theta) = \pi^{-3/2} \frac{\exp\left(-\left(\frac{x_1+x_2}{2} - \theta\right)^2\right)}{1 + (x_1 - x_2)^2} \propto l(\theta|\, x_1, x_2).$$

• If computing p-values, then one will obtain different results for $f(x_1, x_2|\, \theta)$ and $g(x_1, x_2|\, \theta)$ because of they have different tails and the likelihood principle will be violated.

• The likelihood principle does not bother about data you have not observed!

6.1– The Conditionality Principle

- Consider estimating θ in the model on basis of 2 observations, X_1 and X_2.

$$P_\theta\left(X = \theta - 1\right) = P_\theta\left(X = \theta + 1\right)$$

- The procedure suggested is

$$\delta\left(X\right) = \begin{cases} \frac{X_1+X_2}{2}, & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2 \end{cases}.$$

- For a frequentist, this procedure has confidence of 75%;

i.e. $P\left(\delta\left(X\right) = \theta\right) = 0.75$.

- The conditionalist would report 100% confidence if observed data are different or 50% if the observations coincide.

6.1– The Conditionality Principle

• The conditional perspective concerns reporting data specific measures of accuracy.

• In contrast to the frequentist, performance of statistical procedures are judged looking at the observed data.

• **Conditionality Principle.** If two experiments on θ are available and if one of these experiments is selected with proba. p, independently of θ, then the resulting inference should only depend on the selected experiment.

• **Theorem** (Birnbaum, 1962): The likelihood principle is equivalent to the conjunction of the Sufficiency and the Conditionality Principles.