Stat 535 C - Statistical Computing & Monte Carlo Methods

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

- Summary of Previous Lecture.
- Maximum Likelihood.
- Bayesian Statistics.

3.1– Likelihood function

• **Parametric modelling**: The observations x are the realization of a random variable X of probability density function $f\left(x|\,\theta\right)$.

• 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.

• The likelihood function is

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

to emphasize that the observations are *fixed*.

3.2– Sufficient statistics

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

$$f(x|\theta) = h(x)\, g(T(x)|\theta).$$

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

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

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

3.3– Sufficiency principle

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

- Another way to think of it is that the inference on θ is only based on $T(x)$ and not on x: $T(x)$ is sufficient.

- Note that the sufficiency principle is also useful in practice. It is cheaper to store $T(x)$ rather than x.

3.4– Likelihood Principle

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

$$l_1(\theta|x) = c(x)\, l_2(\theta|x).$$

- A simpler (?) way to think of it: You can have two different probabilistic models for the data. However, if $l_1(\theta|x) \propto l_2(\theta|x)$ then this should lead to the same inference.

- Some standard classical statistics procedures do not satisfy this principle because they rely on quantity such as $\Pr(X > \alpha) = \int_\alpha^\infty f(x|\theta)\, dx$ whereas the likelihood principle does not bother about data you have not observed!

• The likelihood principle is fairly vague since it does not lead to the selection of a particular procedure.

• Maximum likelihood estimation is one way to implement the sufficiency and likelihood principles

$$\widehat{\theta} = \arg \sup_{\theta} l\left(\theta | x\right)$$

• Proof:

$$\arg \sup_{\theta} l\left(\theta | x\right) = \arg \sup_{\theta} h\left(x\right) g\left(T\left(x\right) | \theta\right) = \arg \sup_{\theta} g\left(T\left(x\right) | \theta\right).$$

$$l_1\left(\theta | x\right) = c\left(x\right) l_2\left(\theta | x\right) \Rightarrow \arg \sup_{\theta} l_1\left(\theta | x\right) = \arg \sup_{\theta} l_2\left(\theta | x\right)$$

4.2– Maximum Likelihood Estimation

• **Be careful**: Maximum likelihood estimation is just one way to implement the likelihood principle.

• Maximization can be difficult or several equivalent global maxima. However, consistent and efficient in most cases. (asymptotic properties).

• ML estimates can vary widely for small variations of the observations (for small sample sizes).
Example: If $X_i \sim \theta^{-1}\mathbf{1}_{[0,\theta]}(x_i)$ then for n data

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

• Tests require frequentists justifications.

- Many approaches have been proposed: penalized likelihood (e.g. Akaike Information Criterion) or stochastic complexity theory.

- Many of these approaches have a Bayesian flavor.

5.2– Bayesian Statistics

- A Bayesian model is made of a parametric statistical model $(\mathcal{X}, f(x|\theta))$ and a prior distribution on the parameters $(\Theta, \pi(\theta))$.

- The unknown parameters are now considered RANDOM.

- Many statisticians do not like this although they accept the probabilistic modeling on the observations.

- **Example**: Assume you want to measure the speed of light given some observations. Why should I put a prior on this physical constant? Because of the limited accuracy of the measurement, this constant will never be known exactly and thus it is justified to put say a (uniform) prior on this parameter reflecting this uncertainty.

5.2– Bayesian Statistics

• In the Bayesian approach, probability describes degrees of belief.

• In the frequentist interpretation, you should repeat an infinite number of times an experiment and the probabilities corresponds to the limiting frequencies.

• *Problem.* How do you attribute a probability to the following event "There will be a major earthquake in Tokyo on the 27th April 2013"?

• The selection of a prior has an obvious impact on the inference results! However, Bayesian statisticians are honest about it.

- Based on a Bayesian model, we can define

  - The *joint distribution* of (θ, X)

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

  - The *marginal distribution* of X

$$\pi\left(x\right) = \int \pi\left(\theta\right) f\left(\left. x \right| \theta\right) d\theta$$

  For a realization $X = x$, $\pi\left(x\right)$ is called *marginal likelihood* or *evidence.*

5.3– Ingredients of Bayesian Inference

• Given the prior $\pi(\theta)$ and the likelihood $l(\theta|x) = f(x|\theta)$ then Bayes's formula yields

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta}.$$

$\Rightarrow$ It represents all the information on θ than can be extracted from x.

• Note the integral appearing at the denominator of the Bayes' rule!

• The *predictive distribution* of Y when $Y \sim g(y|\theta, x)$

$$g(y|x) = \int g(y|\theta, x)\,\pi(\theta|x)\,d\theta.$$

This is to distinguish from prediction based on $g\left(y|\widehat{\theta}, x\right)$.

5.3– Ingredients of Bayesian Inference

• In case where $\theta = (\theta_1, ..., \theta_p)$ and one is only interested in the parameter θ_k. Then $\theta_{-k} = (\theta_1, ..., \theta_{k-1}, \theta_{k+1}, ..., \theta_p)$ are so-called nuisance parameters.

• Bayesian inference tells us that all the information on θ_k that can be extracted from x is the *marginal* posterior distribution.

$$\pi\left(\theta_k \middle| x\right) = \int \cdots \int \pi\left(\theta \middle| x\right) d\theta_{-k}.$$

• Once more, computing $\pi\left(\theta_k \middle| x\right)$ requires computing a (possibly high dimensional) integral.

• Nuisance parameters are often handled using profile likelihood technique in a maximum likelihood framework.

5.3– Ingredients of Bayesian Inference

- Bayesian statistics do satisfy automatically the sufficiency principle, and the likelihood principle.
- *Sufficiency principle*: If $f\left(x|\,\theta\right) = h\left(x\right) g\left(T\left(x\right)|\,\theta\right)$ then

$$\begin{aligned}\pi\left(\theta|\,x\right) &= \frac{h\left(x\right) g\left(T\left(x\right)|\,\theta\right)\pi\left(\theta\right)}{h\left(x\right)\int g\left(T\left(x\right)|\,\theta\right)\pi\left(\theta\right) d\theta} = \frac{g\left(T\left(x\right)|\,\theta\right)\pi\left(\theta\right)}{\int g\left(T\left(x\right)|\,\theta\right)\pi\left(\theta\right) d\theta} \\ &= \pi\left(\theta|\,T\left(x\right)\right).\end{aligned}$$

- *Likelihood principle*: Assume we have $f_1\left(x|\,\theta\right) = c\left(x\right) f_2\left(x|\,\theta\right)$ then

$$\begin{aligned}\pi\left(\theta|\,x\right) &= \frac{f_1\left(x|\,\theta\right)\pi\left(\theta\right)}{\int f_1\left(x|\,\theta\right)\pi\left(\theta\right) d\theta} = \frac{c\left(x\right) f_2\left(x|\,\theta\right)\pi\left(\theta\right)}{\int c\left(x\right) f_2\left(x|\,\theta\right)\pi\left(\theta\right) d\theta} \\ &= \frac{f_2\left(x|\,\theta\right)\pi\left(\theta\right)}{\int f_2\left(x|\,\theta\right)\pi\left(\theta\right) d\theta}.\end{aligned}$$

5.4– Simple Examples

- For events A and B, the Bayes rule is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P\left(B|\overline{A}\right)P\left(\overline{A}\right)} = \frac{P(B|A)P(A)}{P(B)}$$

- Be careful to subtle exchanging of $P(A|B)$ for $P(B|A)$.

- **Prosecutor's Fallacy**. A zealous prosecutor has collected an evidence and has an expert testify that the probability of finding this evidence if the accused were innocent is one-in-a-million. The prosecutor concludes that the probability of the accused being innocent is one-in-a-million. This is WRONG.

5.4– Simple Examples

- Assume no other evidence is available and the population is of 10 million people.
- Defining $A =$ "The accused is guilty" then $P(A) = 10^{-7}$.
- Defining $B =$"Finding this evidence" then $P(B|A) = 1$ & $P\left(B|\overline{A}\right) = 10^{-6}$.

- Bayes formula yields

$$\frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + P\left(B|\overline{A}\right)P\left(\overline{A}\right)} = \frac{10^{-7}}{10^{-7} + 10^{-6} \times (1 - 10^{-7})} \approx 0.1.$$

- Real-life Example: Sally Clark was condemned in UK (The RSS pointed out the mistake). Her convinction was eventually quashed (on other grounds).

5.4– Simple Examples

- Coming back from a trip, you feel sick and your GP thinks you might have contracted a rare disease (0.01% of the population has the disease).

- A test is available but not perfect.

  If a tested patient has the disease, 100% of the time the test will be positive.

  If a tested patient does not have the diseases, 95% of the the time the test will be negative (5% false positive).

- Your test is positive, should you really care?

• Let A be the event that the patient has the disease and B be the event that the test returns a positive result

$$P\left(A \middle| B \right) = \frac{1 \times 0.0001}{1 \times 0.0001 + 0.05 \times 0.9999} \simeq 0.002$$

• Such a test would be a complete waste of money for you or the National Health System.

• A similar question was asked to 60 students and staff at Harvard Medical School: 18% got the right answer, the modal response was 95%!

5.5– What do we gain from information?

- Bayesian inference involves passing from a prior $\pi(\theta)$ to a posterior $\pi(\theta|x)$ .We might expect that because the posterior incorporates the information from the data, it will be less variable than the prior.

- We have the following identities

$$\begin{aligned} E[\theta] &= E[E[\theta|X]], \\ var[\theta] &= E[var[\theta|X]] + var[E[\theta|X]]. \end{aligned}$$

- It means that, *on average (over the realizations of the data X)* we expect the conditional expectation $E[\theta|X]$ to be equal to $E[\theta]$ and *the posterior variance to be on average smaller than the prior variance* by an amount that depend on the variations in posterior means over the distribution of possible data.

5.6– Variance Decomposition Identity

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

$$var(\theta) = E(var(\theta|X)) + var(E(\theta|X)).$$

Proof:

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

5.7– Be careful

- Such results appear attractive but one should be careful.

- Here there is an underlying assumption that the observations are indeed distributed according to $\pi(x) = \int \pi(\theta) f(x|\theta)\, d\theta$.