Imperial College
London

23rd October 2018

Bayesian methods

`https://bitbucket.org/mfumagal/statistical_inference`

Matteo Fumagalli

Imperial College London

Intended Learning Outcomes

At the end of this module you will be able to:

- critically discuss advantages (and disadvantages) of Bayesian data analysis,
- illustrate Bayes' Theorem and concepts of prior and posterior distributions,
- implement simple Bayesian methods in `R`, including sampling and approximated techniques,
- apply Bayesian methods to solve problems in biology.

Plan of action

- Thinking (ideas, theorem and probabilities)
- Concepts (priors and inferences)
- Computation (asymptotic and sampling methods)
- Approximate bayesian computation

Imperial College London

Intended Learning Outcomes

At the end of this day you will be able to:

- appreciate the use of Bayesian statistics in life sciences,
- formulate and explain Bayes' theorem.
- describe a Normal-Normal model and implement it in `R` with or without Monte Carlo sampling,
- apply Bayesian statistics to estimate genotypes from DNA sequencing data.

Imperial College London

Meet Nessie

Figure 1: Nessie, the Loch Ness Monster. True or fake news?

Imperial College London

Likelihood for a monster to exist (!?)

- $D = \{0, 1\}$ be our data, whether I tell you I saw Nessie or not.
- $\theta = \{0, 1\}$ is the probability distribution for Nessie existing (or not).

Questions

- What are $p(D = 1|\theta = 1)$ and $p(D = 1|\theta = 0)$?
- What is a Maximum Likelihood Estimate of N?

Likelihood thinking...

Our inference on θ is driven solely by our observations, given by our likelihood function.

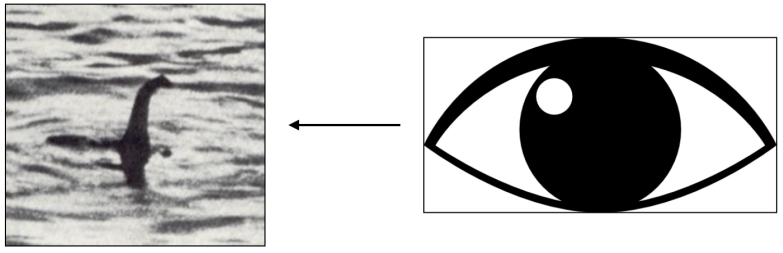


Figure 2: The eye: a "likelihood" organ.

Non-likelihood thinking...

In real life we take many decisions based not only on what we observe but also on some believes of ours.

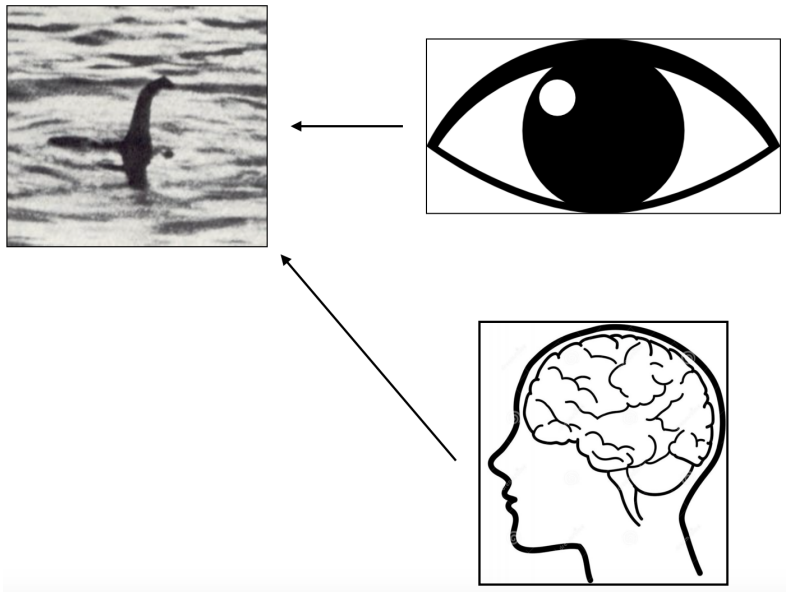


Figure 3: The brain: a "non-likelihood" organ.

Bayesian thinking

- with "eyes only" our intuition is that $p(\theta|D) \propto p(D|\theta)$
- with "the brain" our intuition is that $p(\theta|D) \propto p(D|\theta)p(\theta)$

Our "belief" expresses the probability $p(\theta)$ **unconditional** of the data.

Question

How can we define $p(\theta)$?

Imperial College London

Prior and posterior probability

The "belief" function $p(\theta)$ is called **prior probability** and the joint product of the likelihood $p(D|\theta)$ and the prior is proportional to the **posterior probability** $p(\theta|D)$.

The use of posterior probabilities for inferences is called Bayesian statistics.

Bayesian statistics

Bayesian statistics is an alternative to frequentist approaches but without a definite division as in many cases the approach taken is **eclectic**.

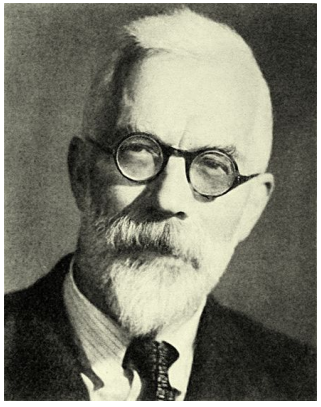


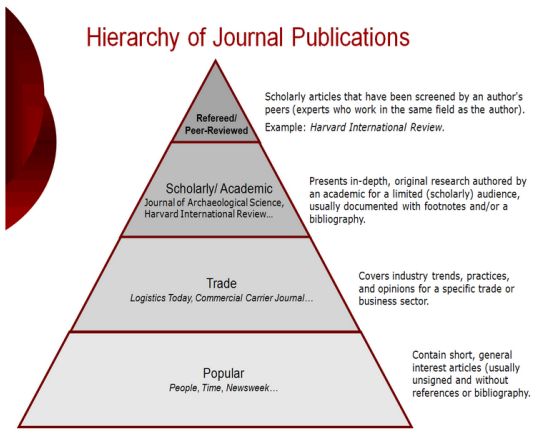
Figure 4: Ronald Fisher.

Imperial College London

Publishing a paper

Example:

You submitted a manuscript for publication to a peer-reviewed journal and you want to assess its probability of being accepted and published.



Which information do you need (and use) to make such inference?

Imperial College London

Measuring biodiversity

Example:

You are measuring the biodiversity of crabs on Scottish rock shores in four different locations over three years.

Table 1: Biodiversity levels.

Year	Loc. A	Loc. B	Loc. C	Loc. D
2016	45	54	47	52
2017	41	?	43	45
2018	32	38	37	35

Question

What is a reasonable value for the missing entry?

Statistical inference

- Frequentist (from data only)
- Likelihoodist (using a statistical model)
- Bayesian (using prior information)
- Empirical Bayesian (observed data contribute to the prior)

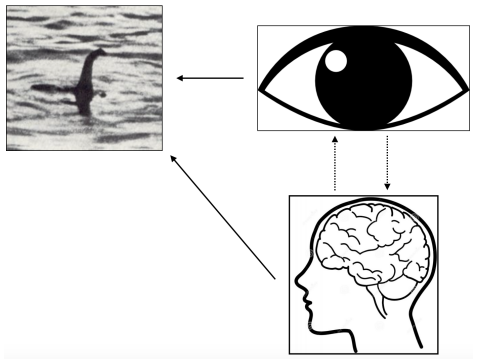


Figure 5: The brain **and** the eye: an Empirical Bayesian organ.

Statistical inference

If D is the data and θ is your unknown parameter, then

- the frequentist conditions on parameters and integrates over the data, $p(D|\theta)$,
- the Bayesian conditions on the data and integrates over the parameters, $p(\theta|D)$.

Bayesian *vs.* Likelihoodist:

- we derive "proper" probability distributions of our parameters rather than deriving a point estimate,
- a probability is assigned to a hypothesis rather than a hypothesis is tested,
- we "accept" the null hypothesis rather than "fail to reject" it,
- parsimony imposed in model choice rather than correcting for multiple tests.

History

Figure 6: Rev. Thomas Bayes

Figure 7: Pierre-Simon, marquis de Laplace

Why?

Why is Bayesian statistics becoming so commonly used?

- recent increased computing power
- good frequentist properties
- answers are easily interpretable by non-specialists
- already implemented in packages

Troubles with the p-value?

John K. Kruschke (2010)

...the fundamental fatal flaw of p-values is that they are ill defined, because any set of data has many different p-values.
...many people mistake the p-value for the probability that the null hypothesis is true.

Troubles with the prior?

John K. Kruschke (2010)

Some people may have the mistaken impression that the advantages of Bayesian methods are negated by the need to specify a prior distribution.

- It is inappropriate not to use a prior.
- Priors are explicitly specified and must be agreeable to a skeptical scientific audience.
- When different groups of scientists have differing priors, stemming from differing theories and empirical, then Bayesian methods provide rational means for comparing the conclusions from the different priors.

Why are WE using it?

Bayesian statistics is very used in many topics in life sciences:

- genetics (e.g. fine mapping of disease-susceptibility genes)
- ecology (e.g. agent-based models)
- evolution (e.g. inference of phylogenetic trees)
- bioinformatics (e.g. genome assembly)
- systems biology (e.g. gene networks)
- ...

and provides a rationale for other approaches (e.g. artificial neural networks).

Imperial College London

The likelihood approach

- Y is a random variable
- $f(y|\theta)$ is a probability distribution (called the *likelihood*) representing the sampling model for the observed data $(y_1, y_2, ..., y_n)$ given a vector of unknown parameters θ
- $\int f(y|\theta)d\theta$ is not necessarily $= 1$ or even finite
- It is possible to find the value of θ that maximises the likelihood function: we can calculate a *maximum likelihood estimate* (MLE) for θ, as: $\hat{\theta} = argmax_{\theta} f(y|\theta)$

Imperial College London

Bayes' Theorem

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{m(y)} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \tag{1}$$

- θ is not a fixed parameter but a random quantity with prior distribution $\pi(\theta)$
- $p(\theta|y)$ is the posterior probability distribution of θ
- $\int p(\theta|y)d\theta = 1$

Imperial College London

Bayes' Theorem in action!

Figure 8: The chytrid fungus *Batrachochytrium dendrobatidis* is the most significant threat to amphibians.

Let's prove Bayes' Theorem!

Imperial College London

Venn diagrams

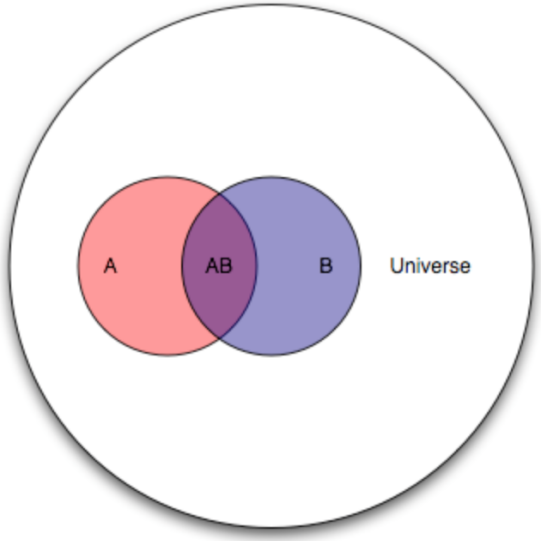


Figure 9: Sets U, A (samples with infection), B (samples with positive test result) and $A \cap B$ (or AB).

Given that the test is positive for a randomly selected sample, what is the probability that said sample is infected?

Normal-Normal model

If

$$f(y|\theta) = N(y|\theta, \sigma^2) \tag{2}$$

$$\pi(\theta) = N(\theta|\mu, \tau^2) \tag{3}$$

then

$$p(\theta|y) = N(\theta|\frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}) \tag{4}$$

Normal-Normal model

"Shrinking" factor B

$$B = \frac{\sigma^2}{\sigma^2 + \tau^2} \tag{5}$$

with $0 \leq B \geq 1$.
Then

$$E(\theta|y) = B\mu + (1 - B)y \tag{6}$$

$$Var(\theta|y) = (1 - B)\sigma^2 \equiv B\tau^2 \tag{7}$$

Normal-Normal model

"Shrinking" factor B

$$B = \frac{\sigma^2}{\sigma^2 + \tau^2} \tag{5}$$

with $0 \leq B \geq 1$.
Then

$$E(\theta|y) = B\mu + (1 - B)y \tag{6}$$

$$Var(\theta|y) = (1 - B)\sigma^2 \equiv B\tau^2 \tag{7}$$

Question:

- What if $\sigma^2 >> \tau^2$?
- What if $\sigma^2 << \tau^2$?

Save the frogs with a Normal-Normal model!

Infected frogs (observed and "believed" *a priori*)

Assume that $y = 6$, $\sigma = 1$, $\mu = 2$ and $\tau = 1$.

$$f(y = 6|\theta) = N(y = 6|\theta, 1) \tag{8}$$

$$\pi(\theta) = N(\theta|2, 1) \tag{9}$$

Save the frogs with a Normal-Normal model!

Infected frogs (observed and "believed" *a priori*)

Assume that $y = 6$, $\sigma = 1$, $\mu = 2$ and $\tau = 1$.

$$f(y = 6|\theta) = N(y = 6|\theta, 1) \tag{8}$$

$$\pi(\theta) = N(\theta|2, 1) \tag{9}$$

Exercise:

1. Open `R` and calculate and plot the prior, likelihood, and posterior distribution.
2. Calculate the *maximum a posteriori probability* (MAP).
3. What happens if we use a skewer (sharper) or wider prior?
4. What happens if we have more observations?

Monte Carlo sampling

Drawing random samples from the posterior distribution instead of calculating its parameters.

Figure 10: Monte Carlo and its famous casino.

Exercise:

1. Open `R` and plot the posterior distribution of the previous example using Monte Carlo sampling.
2. What happens if we use more or less samples?

Imperial College London

Practical 1

Reconstructing genomes from DNA sequencing data.
Follow instructions on jupyter notebook.

Imperial College London

Recap & Refresh

- Why going Bayesian? p-values are troublesome.
- Theorem: $p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}$
- Normal-Normal model: $p(\theta|y) = N(\theta|\frac{\sigma^2\mu+\tau^2 y}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$
- Monte Carlo sampling

Imperial College London

Intended Learning Outcomes

At the end of this day you will be able to:

- describe the pros and cons of using different priors (e.g. elicited, conjugate, ...);
- evaluate the interplay between prior and posterior distributions using R,
- calculate several quantities of interest from posterior distributions,
- apply Bayesian inference to estimate population variation from DNA data.

Prior distributions

How can we decide which prior distribution is more appropriate in our study?

- They are derived from past information or personal opinions from experts.
- They are typically distributed as commonly used distribution families.
- They can be limited to to bear little information.
- ...endless possibilities...

Elicited priors

- Define the collection of θ which are *possible*,
- assign some probability to each one of these cases,
- make sure that they sum up to 1.

Elicited **discrete** priors

Figure 11: How many kits do rabbits have in one litter?

Knowing that "rabbits can have anywhere from one to 14 babies in one litter with an average litter size of 6", let's build a prior distribution.

Elicited **continuous** priors

Figure 12: Bumpass Hell, hot springs and fumaroles at Lassen Volcanic National Park, California.

Knowing that "from past observations, the temperature has a range of (80.1, 110.4) with an average of 88.3 Celsius degrees", let's build a prior distribution.

Elicited **parametric** priors

θ belongs to a parametric distributional family $\pi(\theta|\nu)$.

Advantages:

- reduces the effort to the elicitee,
- overcomes the finite support problem,
- may lead to simplifications in the computation of the posterior.

Disadvantage:

Elicited **parametric** priors

θ belongs to a parametric distributional family $\pi(\theta|\nu)$.

Advantages:

- reduces the effort to the elicitee,
- overcomes the finite support problem,
- may lead to simplifications in the computation of the posterior.

Disadvantage:

- impossible to find a distribution that perfectly matches the elicitee's beliefs.

Elicited **parametric** priors

$$\pi(\theta) = \begin{cases} 0 & \text{for } \theta < 80.1 \text{ or } \theta > 110.4 \\ N(\mu, \sigma^2) & \text{for } 80.1 \leq \theta \leq 110.4 \end{cases} \tag{10}$$

with $\mu = 88.3$ and $\sigma^2 = 10$

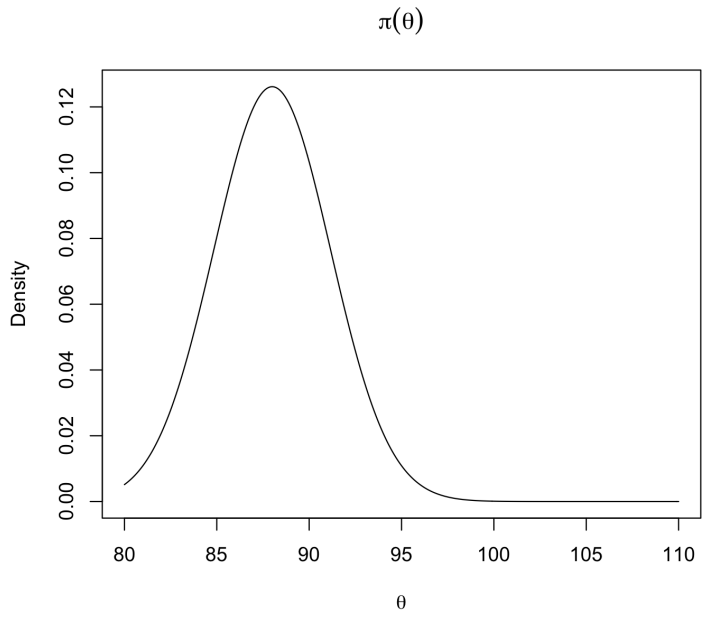


Figure 13: Elicited prior distribution of water temperature.

How to build elicited priors

- Focus on quantiles close to the middle of the distribution (e.g. the 50^{th}, 25^{th} and 75^{th}) rather than extreme quantiles (e.g. the 95^{th} and 5^{th}),
- assess the symmetry,
- priors can be updated and reassessed as new information is available,
- useful for experimental design and data exploration.

Imperial College London

Conjugate priors

$\pi(\theta)$ is member of a family which is *conjugate* with the likelihood $f(y|\theta)$ so that the posterior distribution $p(\theta|y)$ belongs to the same distributional family as the prior.

Conjugate priors

Example: Y is the count of distinct elephant herds arriving at the pool in a day during the migration season.

Figure 14: Elephants drinking at the pool. What's the arrival rate for distinct herds?

Poisson and elephants

Poisson distribution is an appropriate model for Y if:

1. Y is the number of times an event occurs in an interval and it can take values any positive integer values including 0;
2. the occurrence of one event does not affect the probability that a second event will occur (i.e. events occur independently);
3. the rate at which events occur is constant (it cannot be higher in some intervals and lower in other intervals);
4. two events cannot occur at exactly the same instant;
5. the probability of an event in an interval is proportional to the length of the interval.

Poisson distribution

If θ is the event rate (rate parameter), then the probability of observing y events in an interval is:

$$f(y|\theta) = \frac{e^{-\theta}\theta^{y}}{y!},\ y \in \{0, 1, 2, ...\},\ \theta > 0 \tag{11}$$

which is the probability mass function (pmf) for a Poisson distribution.

Let's plot the distribution for $\theta = 4$ using `R`.

Imperial College London

Poisson distribution

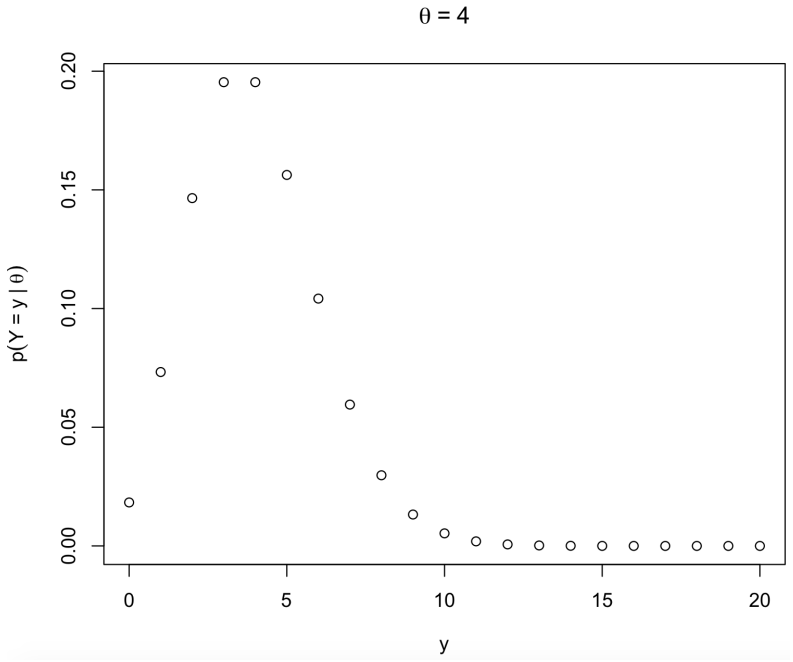


Figure 15: Poisson distribution for $\theta = 4$. This is the likelihood distribution for the number of herds per day with a rate of 4.

Conjugate prior for Poisson distribution

Prior: Gamma distribution

$$\pi(\theta) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \ \theta > 0, \ \alpha > 0, \ \beta > 0 \tag{12}$$

$$E[G(\alpha, \beta)] = \alpha\beta \tag{13}$$

$$Var[G(\alpha, \beta)] = \alpha\beta^2 \tag{14}$$

Imperial College London

Gamma distribution

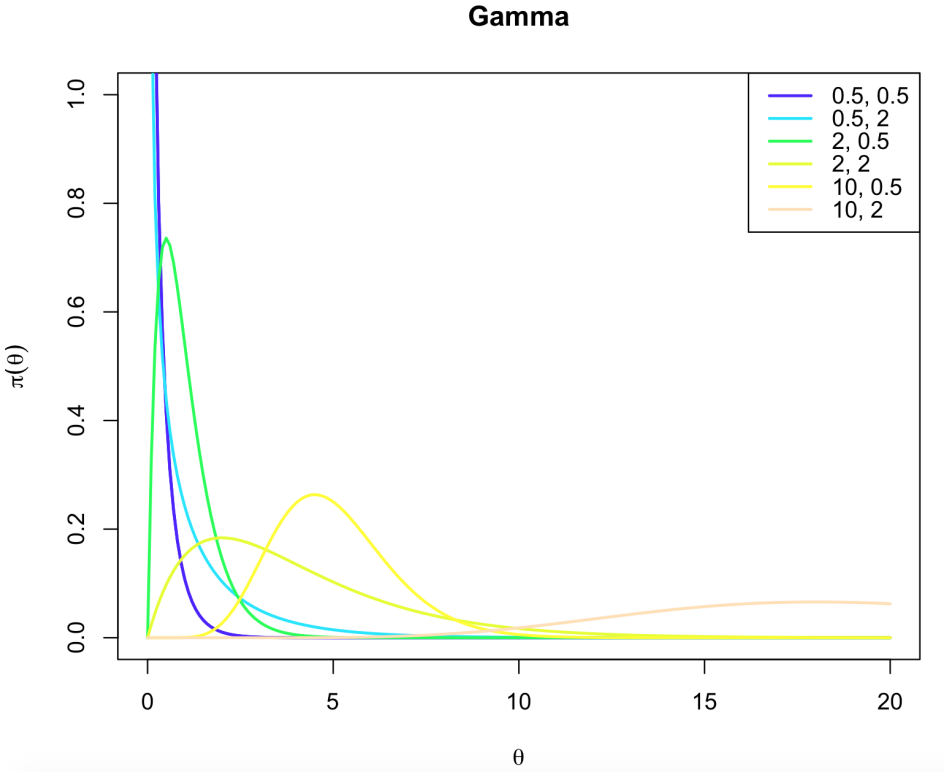


Figure 16: Gamma distribution for different values of shape and rate parameters.

Gamma + Poisson = ?

$$p(\theta|y) \approx f(y|\theta)\pi(\theta) \tag{15}$$

$$\approx ? \tag{16}$$

Gamma + Poisson = Gamma'

$$p(\theta|y) \approx f(y|\theta)\pi(\theta) \tag{17}$$
$$\approx (e^{-\theta}\theta^{y})(\theta^{\alpha-1}e^{-\theta/\beta}) \tag{18}$$
$$= \theta^{y+\alpha-1}e^{-\theta(1+1/\beta)} \tag{19}$$

$p(\theta|y) \sim G(\alpha', \beta')$ with $\alpha' = y + \alpha$ and $\beta' = (1 + 1/\beta)^{-1}$
Posterior is (another) Gamma distribution.

Conjugate priors allow for posterior distributions to emerge without numerical integration!

Elephants' arrivals

Example:

- we have some intuition that we expect to see 3 herds per day (prior)
- we observed 4 herds (data)

Exercise:
What is the posterior distribution of θ, the average number of herds (Gamma-Poisson model)? Calculate and plot the distribution in `R` (use Monte Carlo sampling too).

Hierarchical modelling

Hyperpriors define the density distribution of hyperparameters.

$$p(\theta|y) = \frac{\int f(y|\theta)\pi(\theta|\nu)h(\nu)d\nu}{\int\int f(y|\theta)\pi(\theta|\nu)h(\nu)d\nu d\theta} \tag{20}$$

Hierarchical modelling

Hyperpriors define the density distribution of hyperparameters.

$$p(\theta|y) = \frac{\int f(y|\theta)\pi(\theta|\nu)h(\nu)d\nu}{\int\int f(y|\theta)\pi(\theta|\nu)h(\nu)d\nu d\theta} \tag{20}$$

Empirical Bayesian

Estimated posterior $p(\theta|y,\hat{\nu})$ by replacing ν with an estimate $\hat{\nu}$ obtained by maximising the marginal distribution $m(y|\nu)$.

If $\nu \sim h(\nu|\lambda)$ with unknown parameters λ, then we have a third-stage prior $g(\lambda)$.

Non-informative priors

Can we use a Bayesian approach when no reliable prior information on θ is available?
Yes, a *noninformative prior* distribution for θ contains "no information" about θ and all the information in the posterior will (mostly) arise from the data.

Noninformative priors

Discrete case:
If $\vec{\Theta} = \{\theta_1, \theta_2, ..., \theta_n\}$, then

$$p(\theta_i) = \frac{1}{n},\ i = 1, 2, ..., n$$

with

$$\sum_{1}^{n} \frac{1}{n} = 1$$

Noninformative priors

Continuous and bounded case:
If $\vec{\Theta} = [a, b]$ with $-\infty < a < b < +\infty$, then

$$p(\theta) = \frac{1}{b-a},\ a < \theta < b$$

Noninformative priors

Continuous and unbounded case:
If $\vec{\Theta} = (-\infty, +\infty)$ then

$$p(\theta) = c, \text{ any } c > 0$$

is an *improper* distribution as

$$\int_{-\infty}^{+\infty} p(\theta) d\theta = +\infty$$

Bayesian inference is still possible under some circumstances.

Noninformative priors

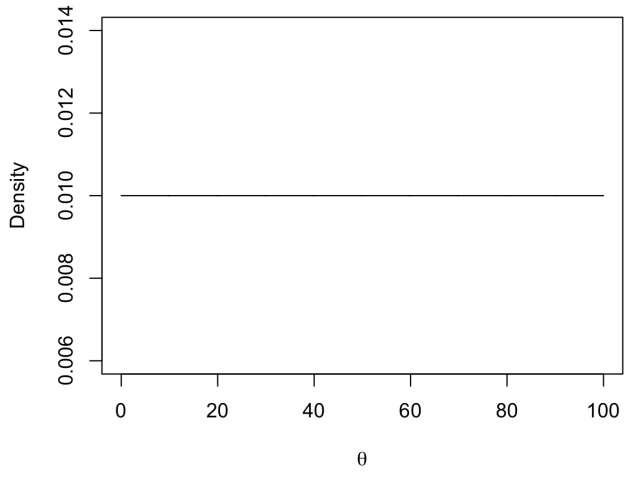


Figure 17: A uniform prior distribution for the arrival rate of elephant herds.

- Rule out scenarios that are impossible in real life.
- Lack a conjugate model (sampling methods are required).
- Non-informative priors are related to *reference* priors.

Bayesian inference

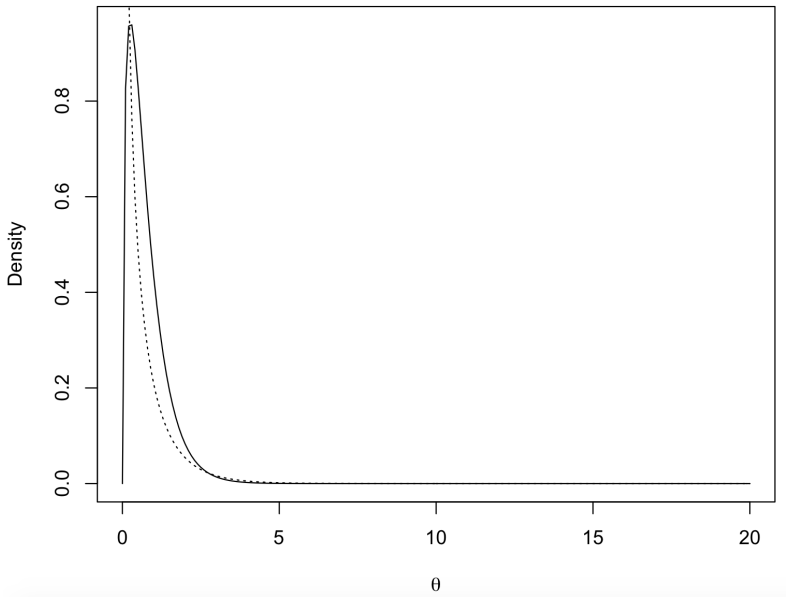
The posterior distribution of model parameters can be difficult to interpret.

We want to **summarise** the information enclosed in these distribution.

Imperial College London

Point estimation

Example: Arrival rate with $y = 1$ and prior $\sim G(0.5, 1)$.



Question:

What is the (i) mean, (ii) mode and (iii) median of this resulting posterior distribution?

Point estimation

- The mode is the easiest to calculate as we can work directly with the numerator.
- If the prior distribution is flat then the *posterior mode* will be equal to the maximum likelihood estimate.
- If the posterior distribution is symmetric, then the mean and the median are equivalent.
- For symmetric unimodal distributions, all these three features are equivalent.
- For asymmetric distributions, the median is often the best choice as it is less affected by outliers and it is an intermediate to the mode and the mean.

Point estimation

If we want to obtain a measure of accuracy of a point estimate $\hat{\theta}(\vec{y})$, we can calculate the *posterior variance*:

$$E_{\theta|\vec{y}}(\theta - \hat{\theta})^2 \tag{21}$$

In the multivariate case the posterior mode is $\hat{\vec{\theta}}(\vec{y}) = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$.

Credible intervals

A $100 \times (1-\alpha)$ credible set for θ is a subset C of Θ such that:

$$1 - \alpha \leq P(C|y) = \int_C p(\theta|y) d\theta \tag{22}$$

"The probability that θ lies in C given the observed data y is at least $(1-\alpha)$"
e.g. $\alpha = 0.05$

Credible intervals

In continuous settings we can calculate the *highest posterior density*, or **HPD**, credible set, defined as:

$$C = \{\theta \in \Theta : p(\theta|y) \geq k(\alpha)\} \tag{23}$$

where $k(\alpha)$ is the largest constant satisfying $P(C|y) \geq (1 - \alpha)$.

Example: $p(\theta|y) \sim G(2, 1)$ and $k(\alpha) = 0.1$.

Credible intervals

How can we summarise our results?

- the posterior mean
- several posterior percentiles (e.g. 0.025, 0.25, 0.50, 0.75, 0.975)
- a credible interval
- posterior probabilities $p(\theta > c|y)$ where c is a notable point (e.g. 0, 1, depending on the problem)
- a plot of the distribution to check whether it is unimodal, multimodal, skewed, ...

Imperial College London

Hypothesis testing

In the **frequentist** approach,

1. one formulates a null hypothesis H_0 and an alternative hypothesis H_a,
2. an appropriate test statistic is chosen $T(Y)$,
3. one computes the *observed significance*, or *p-value*, of the test as the chance that $T(Y)$ is "more extreme" that $T(y_{obs})$, where the "extremeness" is towards the alternate hypothesis,
4. if the p-value is less than some threshold, typically in the form of a pre-specified Type I error rate, H_0 is rejected, otherwise it is not.

Hypothesis testing

Limits of frequentist approach:

1. only when two hypotheses are nested (e.g. H_0 is a simplification of H_a and involves setting one parameter of H_a to some known constant value)
2. evidence *against* the null hypothesis (e.g. a large p-value does not mean that the two models are equivalent, but only that we lack evidence of the contrary; we don't "accept the null hypothesis" but "fail to reject it")
3. no direct interpretation as weight of evidence (but only as a long-term probability; p-values are not the probability that H_0 is true!)

Hypothesis testing

In the **Bayesian** approach,

1. one can test as many models as desired, $M_i, i = 1, ..., m$,
2. one calculates the posterior probability that each model is correct
3. one compares each pair of posterior probabilities.

Hypothesis testing

Suppose we have two models M_1 and M_2 for data Y and the two models have parameters θ_1 and θ_2.

With prior densities $\pi_i(\theta_i)$ and $i = 1, 2$, the marginal distributions of Y are:

$$p(y|M_i) = \int f(y|\theta_i, M_i)\pi_i(\theta_i)d\theta_i \tag{24}$$

We can calculate the posterior probabilities $P(M_1|y)$ and $P(M_2|y) = 1 - P(M_1|y)$ for the two models.

Bayes factors

A Bayes factor (BF) is used to summarise these results, and it is equal to the ration of posterior odds of M_1 to the prior odds of M_1:

$$BF = \frac{P(M_1|y)/P(M_2|y)}{P(M_1)/P(M_2)} = \frac{p(y|M_1)}{p(y|M_2)} \tag{25}$$

If the two models are *a priori* equally probable then:

$$BF = p(M_1|y)/p(M_2|y) \tag{26}$$

which is the posterior odds of M_1.

Imperial College London

Bayes factors

Interpretation

BF captures the change in the odds in favour of model 1 (vs. 2) as we move from the prior to the posterior.

BF	**Strength of evidence**
1 to 3	not worth more than a bare mention
3 to 20	positive
20 to 150	strong
> 150	very strong

Imperial College London

Practical 2: estimate population variation

We now have sequenced our (bears') genomes and, using the method in Practical 1, assigned each individual genotype.
What is the frequency of a certain allele at the **population** level?
Follow instructions in the jupyter notebook.

Recap & Refresh

- How to build elicited priors
- Conjugate priors: normal-normal, poisson-gamma, beta-binomial
- Point estimates (mean, mode, median) and HPD credible intervals
- Model testing

Imperial College London

Intended Learning Outcomes

At the end of this day you will be able to:

- describe the use of asymptotic methods,
- illustrate the utility of direct and indirect sampling methods,
- evaluate the feasibility of Markov Chain Monte Carlo sampling,
- implement simple indirect sampling methods in `R`.

Bayesian computation

The calculation of posterior distributions often involves the evaluation of complex high-dimensional integrals.

When a conjugate prior is not available or appropriate we can evaluate the posterior distribution with:

1. asymptotic methods for approximating the posterior density;
2. numerical integration.

Asymptotic methods

Bayesian Central Limit Theorem

When there are many data points $p(\theta|x)$ will be approximately normally distributed.

For large data points, the posterior can be approximated by a normal distribution with mean equal to the posterior mode and (co)variance (matrix) equal to minus the inverse of the second derivative matrix of the log posterior evaluated at the mode.

Asymptotic methods

Example:

Recalling the beta-binomial model with flat prior, $p(\theta|x) \propto \theta^x(1-\theta)^{n-x}$.

The approximation is given by:

1. take the log:

Asymptotic methods

Example:

Recalling the beta-binomial model with flat prior,
$p(\theta|x) \propto \theta^x(1-\theta)^{n-x}$.

The approximation is given by:

1. take the log: $l(\theta) = x \log \theta + (n-x) \log(1-\theta)$
2. take the derivative of $l(\theta)$ and set it to zero, obtaining

Imperial College London

Asymptotic methods

Example:

Recalling the beta-binomial model with flat prior, $p(\theta|x) \propto \theta^x(1-\theta)^{n-x}$.

The approximation is given by:

1. take the log: $l(\theta) = x \log \theta + (n-x) \log(1-\theta)$
2. take the derivative of $l(\theta)$ and set it to zero, obtaining $\hat{\theta}^{\pi} = \frac{x}{n}$
3. take the second derivative evaluated at $\hat{\theta}$, obtaining

Asymptotic methods

Example:

Recalling the beta-binomial model with flat prior, $p(\theta|x) \propto \theta^x(1-\theta)^{n-x}$.

The approximation is given by:

1. take the log: $l(\theta) = x \log \theta + (n-x) \log(1-\theta)$
2. take the derivative of $l(\theta)$ and set it to zero, obtaining $\hat{\theta}^{\pi} = \frac{x}{n}$
3. take the second derivative evaluated at $\hat{\theta}$, obtaining $-\frac{n}{\hat{\theta}} - \frac{n}{1-\hat{\theta}}$
4. take the minus inverse, obtaining

Imperial College London

Asymptotic methods

Example:

Recalling the beta-binomial model with flat prior, $p(\theta|x) \propto \theta^x(1-\theta)^{n-x}$.

The approximation is given by:

1. take the log: $l(\theta) = x \log \theta + (n-x) \log(1-\theta)$
2. take the derivative of $l(\theta)$ and set it to zero, obtaining $\hat{\theta}^\pi = \frac{x}{n}$
3. take the second derivative evaluated at $\hat{\theta}$, obtaining $-\frac{n}{\hat{\theta}} - \frac{n}{1-\hat{\theta}}$
4. take the minus inverse, obtaining $\frac{\hat{\theta}(1-\hat{\theta})}{n}$
5. $p(\theta|x) \sim N(\hat{\theta}^\pi, \frac{\hat{\theta}(1-\hat{\theta})}{n})$

Asymptotic methods

If $p(\theta|x) \propto \theta^x(1-\theta)^{n-x}$ then for large n we have
$p(\theta|x) \sim N(\hat{\theta}^\pi, \frac{\hat{\theta}(1-\hat{\theta})}{n})$.

Exercise:
$k = 20$
$n = 100$
$\pi(\theta) = G(1,1)$
Compare the exact and approximated posterior (e.g. use qqplot).
What happens if $n = 10$?

Asymptotic methods

Model approximations or *first order approximations*: the estimate θ by the mode and the error goes to 0 at a rate proportional to $1/n$.

The estimates of moments and quantiles may be poor if the posterior differs from normality.

The *Laplace's Method* provides a second order approximation to the posterior mean, with an error that decreases at a rate $1/n^2$.

Asymptotic methods

Advantages:

- they replace numerical integration with numerical differentiation,
- they are deterministic (without elements of stochasticity).
- they reduce the computational complexity if any study of robustness (how sensitive are our conclusions to changes in the prior/likelihood?).

Asymptotic methods

Disadvantages:

- they require that the posterior is unimodal,
- the require that the size of the data is large (how large is "large enough"?),
- for high high-dimensional parameters the calculation of Hessian matrices (second derivatives) are hard.

Noniterative Monte Carlo methods

If $\theta \sim h(\theta)$ with $h(\theta)$ being a posterior distribution, we wish to estimate γ, the posterior mean of $c(\theta)$, where
$\gamma \equiv E[c(\theta)] = \int c(\theta)h(\theta)d\theta$.

If $\theta_1, \theta_2, ..., \theta_N$ are independent and identically distributed (iid) as $h(\theta)$, then:

Noniterative Monte Carlo methods

If $\theta \sim h(\theta)$ with $h(\theta)$ being a posterior distribution, we wish to estimate γ, the posterior mean of $c(\theta)$, where $\gamma \equiv E[c(\theta)] = \int c(\theta)h(\theta)d\theta$.

If $\theta_1, \theta_2, ..., \theta_N$ are independent and identically distributed (iid) as $h(\theta)$, then:

$$\hat{\gamma} = \frac{1}{N}\sum_{i=1}^{N} c(\theta_i) \tag{27}$$

which converges to $E[c(\theta)]$ with probability 1 as $N \to \infty$.

The computation of **posterior expectations** requires only a sample of size N from the posterior distribution.

Noniterative Monte Carlo methods

The variance of $\hat{\gamma}$ can be estimated from the sample variance of the $c(\theta_i)$ values.

$$\hat{se}(\hat{\gamma}) = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} [c(\theta_i) - \hat{\gamma}]^2} \tag{28}$$

The Central Limit Theorem implies that $\hat{\gamma} \pm 2\hat{se}(\hat{\gamma})$ provides the approximated 95% confidence interval.
N can be chosen as large as necessary to provide a desirable confidence interval.

Noniterative Monte Carlo methods

In the univariate case, a histogram of the sampled θ_i estimates the posterior itself.

An estimate of $p \equiv P\{a < c(\theta) < b\}$ is given by

$$\hat{p} = \frac{\text{number of } c(\theta_i) \in (a, b)}{N} \tag{29}$$

In contrast to asymptotic methods, accuracy improves with N, the Monte Carlo sample size (which we can choose and have control upon) rather than n the size of the data set (which can may not be able to control).

Noniterative Monte Carlo methods

What happens if we can't directly sample from this distribution? There are methods for **indirect** sampling of the posterior distribution: (i) importance sampling, (ii) rejection sampling, (iii) weighted bootstrap.

Rejection sampling

If we identify an *envelope function* $g(\theta)$ and a constant $M > 0$ such that $L(\theta)\pi(\theta) < Mg(\theta)$ for all θ, then:

1. Generate $\theta_i \sim g(\theta)$,
2. Generate $U \sim Uniform(0, 1)$,
3. If $MUg(\theta_i) < L(\theta_i)\pi(\theta_i)$ accept θ_i otherwise reject θ_i.

If we repeat this procedure until N samples are obtained, the members of this sample will be random variables from $h(\theta)$.

It is hard to sample from the true posterior but it is easier to sample from the envelope function.

Exercise: approximate a Beta distribution using a uniform envelope function.

Markov chain Monte Carlo methods

- All previous methods are non-iterative as they draw a sample of fixed size N.
- There is no notion of "convergence" but rather we require N to be sufficiently large.
- For many problems with high dimensionality it may be difficult to find an importance sampling density or an envelope function.

In these cases it is now standard practice to use *Markov chain Monte Carlo* (MCMC) methods.

Markov process and chain

1. A mathematical object following a stochastic (or random) process, typically defined as a collection of random variables.
2. The next value of the process depends only on the current value, but it is independent of the previous values.
3. A Markov chain is a Markov process that has a particular type of state space, which dictates the possible values that a stochastic process can take.

Markov chain Monte Carlo

Stationary distribution

The probability distribution to which the process converges for large values of steps, or iterations.

The stationary distribution of an MCMC is the desired posterior distribution.

Markov chain Monte Carlo

- The basic idea is to construct a Markov chain on the state space Θ whose stationary distribution is the target posterior density $p(\theta|Y)$.
- We perform a random walk on the state space, so that the fraction of time we spend in each state θ is proportional to $p(\theta|Y)$.
- By drawing correlated samples $\theta_0, \theta_1, \theta_2, \ldots,$ from the chain, we can perform Monte Carlo integration with respect to $p(\theta|Y)$.

Markov chain Monte Carlo

An assessment of *convergence* of the Markov chain to its stationary distribution is required.

The majority of Bayesian MCMC computation is based on two algorithms: the *Gibbs sampler* and the *Metropolis-Hastings (M-H)* algorithm.

Imperial College London

Gibbs sampler

Suppose our model has k parameters $\theta = (\theta_1, \theta_2, ..., \theta_k)$.

We assume that we can sample from the full conditional distributions.

The collection of full conditional distributions uniquely determines the joint posterior distribution $p(\theta, y)$ and therefore all marginal posterior distributions $p(\theta_i, \vec{y})$, for $i = 1, ..., k$.

...

Gibbs sampler

...

Given an arbitrary set of starting $\{\theta_2^{(0)}, ..., \theta_k^{(0)}\}$, the algorithm, for $(t = 1, ..., T)$, is:

- Draw $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_k^{(t-1)}, y)$
- Draw $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_k^{(t-1)}, y)$
- ...
- Draw $\theta_k^{(t)}$ from $p(\theta_k|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{k-1}^{(t)}, y)$

$(\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_k^{(t)})$ converges to a draw from the true joint posterior distribution $p(\theta_1, \theta_2, ..., \theta_k|y)$.

For $t > t_0$ then $\{\theta^{(t)}, t = t_0 + 1, ..., T\}$ is a correlated sample from the true posterior.

Gibbs sampler

1. The parameter space must be fully *connected*, without "holes".
2. When θ and ν are highly correlated the chain will have a "slow mixing".
3. To ensure that all the full conditional distributions are available, the prior distribution of each parameter can be chosen to be conjugate with the corresponding likelihood.

Gibbs sampler

A histogram of $\{\theta_i^{(t)}, t = t_0 + 1, ..., T\}$ provides an estimator of the marginal posterior distribution for θ_i.

The posterior mean can be estimated as the posterior mean:

$$\hat{E}(\theta_i|y) = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \theta_i^{(t)} \tag{30}$$

The time $0 <= t <= t_0$ is called the *burn-in* period.

Imperial College London

Metropolis algorithm

Given:

- $p(\theta|\vec{y}) \propto h(\theta) \equiv f(y|\theta)\pi(\theta)$
- a *candidate*, or *proposal*, symmetric density $q(\theta^*|\theta^{(t-1)})$ which satisfies $q(\theta^*|\theta^{(t-1)}) = q(\theta^{(t-1)}|\theta^*)$,
- a starting value $\theta^{(0)}$ at iteration $t = 0$,

for $(t = 1, ..., T)$ the algorithm repeats:

1. Draw $\theta^* = q(\cdot|\theta^{(t-1)})$
2. Calculate $r = h(\theta^*)/h(\theta^{(t-1)})$
3. If $r \geq 1$, set $\theta^{(t)} = \theta^*$, otherwise set $\theta^{(t)} = \theta^*$ with probability r or set $\theta^{(t)} = \theta^{(t-1)}$ with probability $1 - r$.

$\theta^{(t)}$ converges in distribution to a draw from the true posterior density $p(\theta|y)$.

Metropolis algorithm

A usual candidate density is:

$$q(\theta^*|\theta^{(t-1)}) = N(\theta^*|\theta^{(t-1)}, \tilde{\Sigma}) \tag{31}$$

random walk Metropolis: symmetric and "self-correcting" distribution.

$\tilde{\Sigma}$, the posterior variance, can be empirically estimated from a preliminary run.

Metropolis-Hastings algorithm

When $q(\theta * |\theta^{(t-1)}) \neq q(\theta^{(t-1)}|\theta*)$ the acceptance rate r is:

$$r = \frac{h(\theta^*)q(\theta^{(t-1)}|\theta^*)}{h(\theta^{(t-1)})q(\theta^*|\theta^{(t-1)})} \tag{32}$$

A draw $\theta^{(t)}$ converges in distribution to a draw from the true posterior density as $t \to \infty$.

Hastings independence chain

If we set $q(\theta^*|\theta^{(t-1)}) = q(\theta^*)$ then the proposal ignores the current value of the variable.

The acceptance rate is:

$$r = \frac{h(\theta^*)/q(\theta^*)}{h(\theta^{(t-1)})/q(\theta^{(t-1)})} \tag{33}$$

MCMC algorithms

- *Langevin-Hastings* algorithm introduces a systematic drift in the candidate density.
- *Slice sampler* algorithm uses auxiliary variables to expand the parameter space.
- *Hybrid* forms combined multiple algorithm in a single problem.
- *Adaptive* algorithms use the early output from a chain to refine the sampling as it progresses.

Convergence

Diagnostic strategy:

- run parallel chains with starting points from a wide distribution;
- visually inspect these chains;
- for each graph calculate the scale reduction factor;
- investigate crosscorrelations among parameters.

Figure 18: Three chains mixing for increasing t.

Recap & Refresh

If there is no closed form for the posterior distribution:

- asymptotic methods (e.g. Normal approximation)
- direct sampling (Monte Carlo)
- non-iterative sampling (e.g. rejection algorithm)
- iterative sampling: MCMC (Gibbs vs. Metropolis algorithms)

Imperial College London

Intended Learning Outcomes

At the end of this day you will be able to:

- appreciate the applicability of ABC,
- describe the rejection algorithm,
- critically discuss the choice of summary statistics,
- implement ABC methods in `R`.

Posterior probability distribution

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)}$$

which can be difficult as the marginal likelihood

$$p(x) = \int p(x|\theta)\pi(\theta)d\theta$$

might involve a high dimensional integral difficult (or impossible) to solve.

Sampling from the posterior

- If the likelihood can be evaluated up to a normalising constant, Monte Carlo methods can be used to sample from the posterior.
- If the likelihood function becomes difficult to define and compute, it is easier to *simulate* data samples from the model given the value of a parameter.

Rejection algorithm

If data points are **discrete** and of low dimensionality, given observation y, repeat the following until N points have been accepted:

1. Draw $\theta_i \sim \pi(\theta)$
2. Simulate $x_i \sim p(x|\theta_i)$
3. Reject θ_i if $x_i \neq y$

These are sampled from $p(\theta|x)$.

Rejection algorithm (elephants are back!)

Example:

- We observe 4 herds arriving.
- The likelihood is Poisson-distributed and the prior is Gamma-shaped $G(3, 1)$.
- The posterior distribution is Gamma distributed with shape parameter $3 + 4 = 7$ and scale/rate 0.5.

Let's assume that we can't evaluate the likelihood but we know how to *simulate* y given a certain value of our parameter θ.
Exercise:
Calculate the posterior distribution of θ in R.

Rejection algorithm

If data points are **continuous** and of low dimensionality, given observation y, repeat the following until N points have been accepted:

1. Draw $\theta_i \sim \pi(\theta)$
2. Simulate $x_i \sim p(x|\theta_i)$
3. Reject θ_i if $\rho(x_i, y) > \epsilon$

where $\rho(\cdot)$ is a function measuring the distance between simulated and observed points.

Rejection algorithm (hot water is back!)

Example (water temperature):

- θ is continuous with prior distribution $U(80.1, 110.3)$.
- we have a single observation $y = 91.3514$.
- as $\rho(\cdot)$ we use the Euclidean distance:

$$\rho(x_i, y) = \sqrt{(x_i - y)^2} \tag{34}$$

We can't evaluate the likelihood function but we can simulate observations that are distributed according to it.
Exercise:
Calculate the posterior distribution of θ using R.

Rejection algorithm v2

Alternatively, ϵ is the proportion of accepted simulations (ranked by distance with observations). In this case one sets the number of simulations to be performed (not the number of accepted simulations).

Rejection algorithm v2

Exercise (water temperature):

1. $Y = \{91.34, 89.21, 88.98\}$
2. θ has prior $N(\mu = 90, \sigma^2 = 20)$ for $80 \geq \theta \leq 110$
3. the simulating function is `simulate <- function(param) rnorm(n=1, mean=param, sd=sqrt(10))`
4. the distance function is $\rho(x_i, Y) = \frac{\sum_{j \in Y} \sqrt{(x_i - j)^2}}{|Y|}$
5. $N = 10,000$ and $\epsilon = 0.05$

Tasks:

1. plot the sampled prior distribution
2. plot the distribution of ranked distances with indication of 5% threshold
3. plot the posterior distribution
4. calculate notable quantiles and HPD 95%

Rejection algorithm with high dimensionality

If data points are of **high dimensionality**, given observation y, repeat the following until N points have been accepted:

1. Draw $\theta_i \sim \pi(\theta)$
2. Simulate $x_i \sim p(x|\theta_i)$
3. Reject θ_i if $\rho(S(x_i), S(y)) > \epsilon$

with $S(y)$ being summary statistics.

Imperial College London

Rejection algorithm

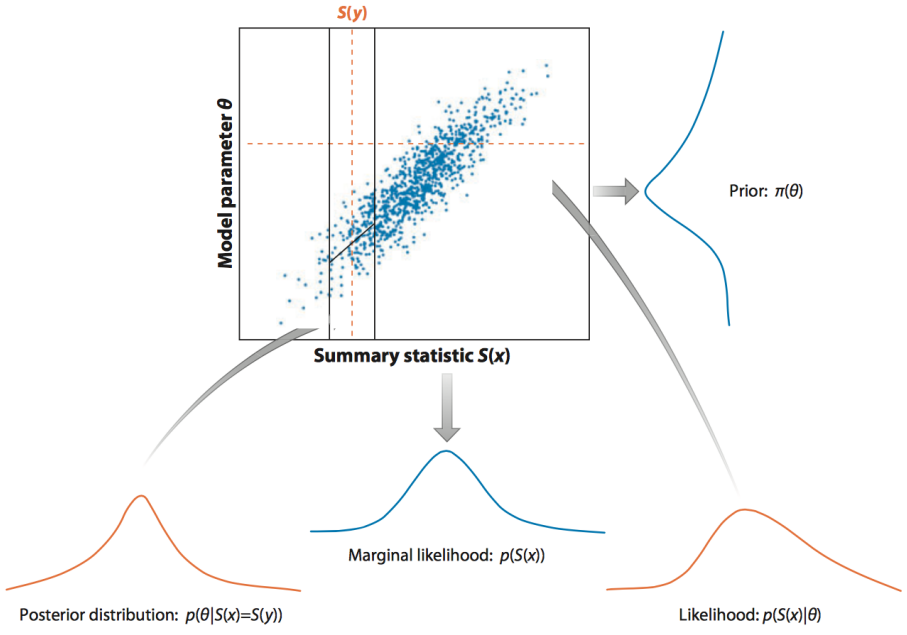


Figure 19: From Beaumont 2010 Annu Rev Ecol Evol Syst. Rejection- and regression-based approximate Bayesian computation (ABC).

Summary statistics

- The choice of summary statistics is a mapping from a high dimension to a low dimension.
- Some information is lost, but with enough summary statistics much of the information is kept.
- The aim for the summary statistics is to satisfy the Bayes' sufficiency:

$$p(\theta|x) = p(\theta|S(x))$$

Issues?

Summary statistics

- The choice of summary statistics is a mapping from a high dimension to a low dimension.
- Some information is lost, but with enough summary statistics much of the information is kept.
- The aim for the summary statistics is to satisfy the Bayes' sufficiency:

$$p(\theta|x) = p(\theta|S(x))$$

Issues?

Solutions:

1. use a wider acceptance tolerance
2. perform a better sampling from the prior

Regression-based ABC

1. Given observation y repeat the following until M points have been generated: A. Draw $\theta_i \sim \pi(\theta)$; B. Simulate $x_i \sim p(x|\theta_i)$
2. Calculate $S_j(x)$ for all j and k_j
3. $\rho(S(x), S(y)) : \sqrt{\sum_{j=1}^{s} (\frac{S_j(x)}{k_j} - \frac{S_j(y)}{k_j})^2}$
4. Choose ϵ such that the proportion of accepted points $P_\epsilon = \frac{N}{M}$
5. Weight the simulated points $S(x_i)$ using $K_\epsilon(\rho(S(x_i), S(y)))$

$$K_\epsilon(t) = \begin{cases} \epsilon^{-1}(1 - (t/\epsilon)^2) & \text{for } t \leq \epsilon \\ 0 & \text{for } t > \epsilon \end{cases}$$

6. Apply weighted linear regression to the N points that have nonzero weight to obtain an estimate of $\hat{E}(\theta|S(x))$
7. Adjust $\theta_i^* = \theta_i - \hat{E}(\theta|S(x)) + \hat{E}(\theta|S(y))$
8. The θ_i^* with weights $K_\epsilon(\rho(S(x_i), S(y)))$ are random draws from an

MCMC-ABC

Initialise by sampling $\theta^{(0)} \sim \pi(\theta)$.

At iteration $t \geq 1$,

1. Simulate $\theta' \sim K(\theta|\theta^{(t-1)})$ where $K(\cdot)$ is a proposal distribution that depends on the current value of θ
2. Simulate $x \sim p(x|\theta')$.
3. If $\rho(S(x), S(y)) < \epsilon$ (rejection step),
   - $u \sim U(0,1)$,
   - if $u \leq \pi(\theta')/\pi(\theta^{(t-1)}) \times K(\theta^{(t-1)}|\theta')/K(\theta'|\theta^{(t-1)})$, update $\theta(t) = \theta'$;
   - otherwise $\theta(t) = \theta^{(t-1)}$;
4. otherwise $\theta(t) = \theta^{(t-1)}$.

Model assessment

Model choice

Given a series of model $\mu_1, \mu_2, ..., \mu_N$ with prior probabilities $\sum_i \pi(\mu_i) = 1$, it is of interest to calculate Bayes factors between two models i and j:

$$\frac{p(\mu_i|x)}{p(\mu_j|x)} \div \frac{p(\mu_i)}{p(\mu_j)} \tag{35}$$

Choice of summary statistics

The more the merrier?

Figure 20: Choosing summary statistics: the issue of pulling a short blanket.

Choice of summary statistics

1. One could calculate the ratio of posterior density with or without a particular summary statistic. Departures greater than a threshold are suggestive that the excluded summary statistic is important.
2. Different summary statistics can be weighted differently according to their correlation with some model parameters.
3. The number of summary statistics can also be reduced via multivariate dimensional scaling summary statistics should be scaled in order to have equal mean and variance, if normally distributed.
4. Even if there is no need of a strong theory relating summary statistics to model parameters, it is suitable to have some expectations.

Model validation

Validation is the assessment of goodness-of-fit of the model and comparing alternative models, to distinguish errors due to the approximation from errors caused by the choice of the model.

1. The distributions of simulated summary statistics are visualised and compared to the corresponding target statistic. If the target is outside, then this could be a problem in the model.
2. The observations are compared with the posterior predictive distribution. This can be done by simulating data with parameters drawn randomly from the current posterior distribution.

Imperial College London

Applications of ABC in biology

Population genetics, agent-based models, protein interaction networks, speciation rates under a neutral ecological model, extinction rates from phylogenetic data, epidemiology, ...

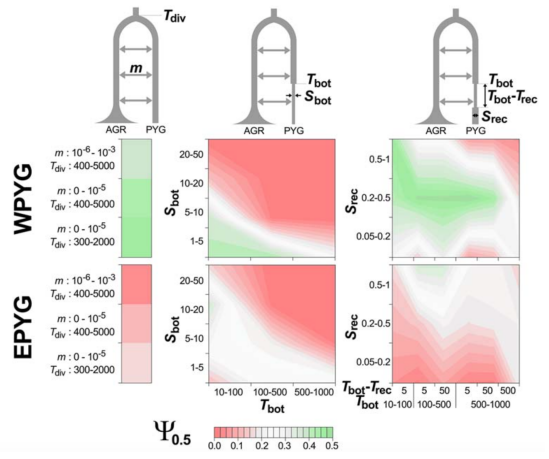


Figure 21: Patin et al. (2009). Different models simulating the demographic regime of the African groups and the mean proportion of small distances ($\Omega_{0.5}$) obtained in comparisons with simulated statistics.

To ABC or not to ABC?

- When a likelihood function is known and can be efficiently evaluated, then there is not advantage to use ABC.
- When the likelihood function is known but difficult to evaluate in practise, the ABC is a valid option.
- Many scenarios in evolutionary biology or ecology can be generated by simulations.
- ABC can be useful for initial exploratory phase.
- Be careful with the choice of your priors!

Imperial College London

Intended Learning Outcomes

At the end of this module you are now able to:

- critically discuss advantages (and disadvantages) of Bayesian data analysis,
- illustrate Bayes' Theorem and concepts of prior and posterior distributions,
- implement simple Bayesian methods in `R`, including sampling and approximated techniques,
- apply Bayesian methods to solve problems in (ecology) and evolution.

Imperial College London

Final remarks

- Available anytime for questions (email or my office is in Munro N1.6, Silwood Park) about practicals, coursework or use of Bayesian methods in your projects
- The bitbucket repository will go offline after the final exam, make sure you have a local copy

Questions?

Imperial College London

Thank you

- Submit your feedback, each constructive comment is seriously taken in consideration
- Possibility of master projects (statistical genetics or deep learning) + PhD position in genomics + neural networks with UoReding