Introduction to Bayesian methods in (ecology) and evolution

Matteo Fumagalli

February 14, 2017

Contents

Disclaimer

Do not distribute!

This document uses material protected by copyright. This document should be use only by students of MSc and MRes CMEE course at Silwood.

The main source of inspiration for this material was "Bayesian Methods for Data Analysis" by B.P. Carlin and T.A. Louis, CRC Press.

Please report ant typo or evident mistakes to m.fumagalli@imperial.ac.uk

Requirements for practicals: R and packages: coda, abc, grid, maps, spam, fields. Software: ms (Hudson's coalescent simulator).

1 Bayesian thinking

1.1 The eyes and the brain

Imagine I tell you that I have just spotted the Loch Ness monster in the lake here at Silwood Park campus (Figure 1). What does this information tell you on the existence or not of Nessie[1]?

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

In the classic frequentist, or likelihood, approach you make some inferences based on all the data that you have observed. The only data that you observe here is me telling you whether or not I have seen Nessie. In other words, your inference on whether Nessie exists (at Silwood!) or not will be solely based on such observations.

Let's denote T as the set of observations specifying whether I tell you that I have seen Nessie ($T = 1$) or not ($T = 0$). T is our sample space, the set of all possible outcomes of the experiment, and $T = \{0, 1\}$. We want to make some inferences on the probability that Nessie exists, or that it is true that I have seen it (her?). Let's denote this probability as N and assume for simplicity that $N = \{0, 1\}$.

We can define a likelihood function for $p(T|N)$. For instance, assuming that our observation is $T = 1$, we could set $p(T = 1|N = 0) = 0.01$ (what are the chances that I did not see Nessie, so it doesn't exist, but I tell I saw it?) and $p(T = 1|N = 1) = 0.90$ (what are the chances that I did see Nessie and tell you I saw it?). In this very trivial example we maximise this likelihood function for N and obtain a Maximum Likelihood Estimate (MLE) of $N = 1$. We can see how our inference on N is driven solely by our observations, and this given by our likelihood function. In a very informal notation, we can write $p(N|T) \approx p(T|N)$, where we stress the conditional of $p(N)$ on the observed data.

[1] I must acknowledge Daniel Wegmann who used a similar analogy (but with ET) during one of his lectures I had the fortune to attend. Any misinterpretation of this example is purely mine.

An analogy here can explain this concept further. Imagine that in the likelihood approach we use only one visual (or auditive) organ, i.e. our eyes (or ears) (Figure 2).

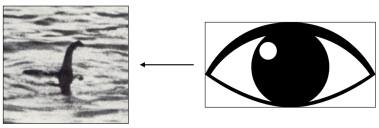

Figure 2: The eye: a "likelihood" organ.

However, in real life, we take many decisions based not solely on what we observe but also on some believes of ours. Typically we use another organ, the brain, to make some inferences on the probability of a particular event to occur (Figure 3). Note that in this cartoon the brain is "blind", in the sense that it does not observe the data (no arrow pointing to the eye) but its inferences on the event are based on its believes.

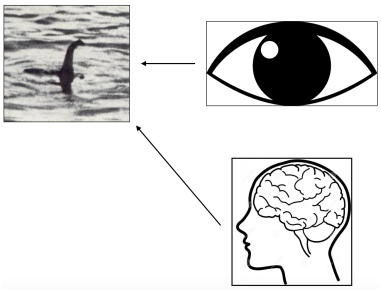

Figure 3: The brain: a "non-likelihood" organ.

Back to the Loch Ness monster case, we can clearly have some believes whether or not Nessie exists not only because I told you I have seen it in the campus. This "belief" expresses the probability of Nessie existing $p(N)$ unconditional of the data. Looking at Figure 3, our intuition is that the probability of $N = 1$ is somehow a joint product of the likelihood (the eyes) and the belief (the brain). Therefore, $p(N|T) \approx p(T|N)p(N)$.

How can we define $p(N)$? This depends on our blind "belief" function. Suppose you are a Sci-Fi fan you might be inclined to set a higher probability than the one a pragmatical sceptical person would set. As an illustration, if $p(T = 1|N = 0) = 0.001$ and $p(T = 1|N = 1) = 0.1$, then in the likelihood approach we have that $p(N = 1|T = 1) \approx p(T = 1|N = 1) \approx 0.1$. In the "Sci-Fi brain" approach, $p(N = 1|T = 1) \approx p(T = 1|N = 1)p(N = 1) \approx 0.1 * 0.2 = 2e - 2$. In the "sceptical brain" approach, $p(N = 1|T = 1) \approx p(T = 1|N = 1)p(N = 1) \approx 0.1 * 0.002 = 2e - 4$. Note that these are not probabilities. Nevertheless, we can deduct how the choice of a different "belief" function can point us to either different conclusions or confidence levels.

In statistics, the "belief" function (e.g. $p(N)$) is called **prior probability** and the joint product of the likelihood and the prior is proportional to the **posterior probability** (e.g. $p(N|T)$). The use of posterior probabilities for inferences is called Bayesian statistics.

1.2 What, who, and why?

What

Bayesian statistics is an alternative to classical frequentist approaches, where maximum likelihood estimates (MLE) and hypothesis tested based on p-values are often used. However, there is no definite division between frequentists and Bayesians as, in many modern applications, the approach taken is eclectic. We now discuss further examples (additionally to the Loch Ness monster case, but more concrete) where a Bayesian approach seems the more appropriate strategy to adopt for statistics inference.

Imagine you have submitted a manuscript for publication to a peer-reviewed journal. You want to assess its probability of being accepted and published. This assessment may use, for instance, the information regarding the journal's acceptance rate (say 20%), the quality of the study and its relevance to the journal's scope[2]. Whatever the reason, your manuscript is accepted and now you want to re-assess the probability that your next manuscript will be accepted. What is the direct estimate of this probability? You had one success over one trial. Therefore, the probability is 100%. However, it looks clear that this estimate is somehow "wrong" as it is based on a small sample size and we know that the acceptance rate is anyway smaller than 100%. You can think of the journal's acceptance rate as our prior information. You are then tempted to set a number smaller than 100% and doing so you are behaving as a Bayesian statistician, as you are adjusting the direct estimate in light of the prior information. Bayesian statistics have the ability to incorporate prior information into an analysis.

Suppose you are conducting an experiment where you are measuring the biodiversity of some species on some rock shores in Scotland. Specifically, you are collecting the number of different species of crabs/mussels/fishes/algae (pick your favourite one) in four different locations across time, over three years (Table 1). Unfortunately, something happened in 2015 for Location B and you do not have data reported. What is a reasonable value for that entry, with no direct information? Would you think that 100 could be an answer? Probably 100 is too high since the numbers surrounding the entry may point towards a value of around 45. We could fit a model or take an average to impute the missing data. Now assume that you have access to some data for Location B in 2015. Specifically, you have partial data where you could retrieve biodiversity levels only for a fifth on Location B on 2015 You extrapolate to obtain a full

[2]In practise more variables influence the probability of your paper being accepted, namely your affiliation, the senior author's reputation, and whether the editor is a friend of your PI's. Obviously, this last sentence is ironic as we all know that the current single-blind peer-review system is perfectly unbiased.

Table 1: Biodiversity levels in Scottish rock shore.

Year	Loc. A	Loc. B	Loc. C	Loc. D
2014	45	54	47	52
2015	41	?	43	45
2016	32	38	37	35

estimate for that cell and retrieve 100. Are you willing now to impute missing data with 100, extrapolated from some partial coverage, while before you thought this number was much higher than expected? A more intuitive solution would be to take a sort of weighted average between this direct (but uncertain) measurement (100) and the indirect estimate you used (45) when there was no information available. Finally, imagine that you can direct observation for half of Location B in 2015. If so, then you would like to "weight" more your direct observation compared to the previous case where only a fifth of the area was available. Bayesian statistics formalises such integration between direct and indirect information.

Let's recapitulate and describe all inference approaches we have been discussing so far: frequentist, likelihoodist, Bayesian, and Empirical Bayesian.

- The *frequentist* is based on imagining repeated sampling from a particular model, which defines the probability of the observed data conditional on unknown parameters.
- The *likelihoodist* uses the sampling model as the frequentists but all inferences used the data only.
- The *Bayesian* requires a sampling model (the likelihood) and a prior distribution on all unknown parameters. The prior and the likelihood are used to compute the conditional distribution of the unknown parameters given the observed data.
- The *Empirical Bayesian* (EB) allows the observed data to contribute to defining the prior distribution (Figure 4).

To put in a different perspective, assuming D is the data and θ is your unknown parameter, the frequentist conditions on parameters and integrates over the data, $p(D|\theta)$. On the other hand, the Bayesian conditions on the data and integrates over the parameters, $p(\theta|D)$. Therefore, in Bayesian statistics we derive "proper" probability distributions of our parameters of interest, rather than deriving a point estimate. In other words, In Bayesian statistics a probability is assigned to a hypothesis, while under a frequentist inference, a hypothesis is tested without being assigned a probability. Unlike likelihoodist, Bayesian inferences can really "accept" the null hypothesis rather than "fail to reject" it. Bayesian procedures can also impose parsimony in model choice and avoid further testing for multiple comparisons.

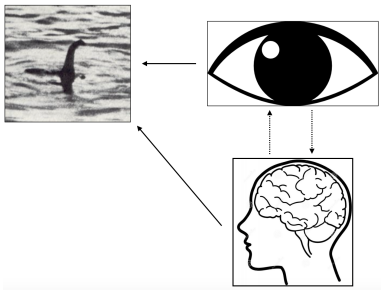

Figure 4: The brain and the eye: an Empirical Bayesian organ.

Who

Let's now address perhaps the question you are all asking yourself: why is it called Bayesian? It is called after Thomas Bayes (1701–1761)[3], an English statistician, philosopher and Presbyterian minister (Figure 5). Thomas Bayes never published his famous accomplishment in statistics. His notes were edited and published after his death. He studied logic and theology and at the age of 33 he became a minister in a village in Kent. Only in his later years Thomas Bayes took a deep interest in probability. The general interpretation of statisti-

Figure 5: Thomas Bayes.

cal inference called Bayesian was in reality pioneered by Pierre-Simon Laplace (Figure 6). In fact, some argue that Bayes intended his results in a very limited way than modern Bayesians would intend them. In the special case Thomas Bayes presented, the prior and posterior distributions were Beta distributions and the data came from Bernoulli trials. Interestingly, early Bayesian inference was called "inverse probability", because it infers backwards from observations to parameters.

[3]These notes are all taken from Wikipedia.

Figure 6: Pierre-Simon, marquis de Laplace.

Why

The recent explosion of interest in Bayesian methods for data analysis is mainly because:

- of the increased computing power over the last years,
- they have good frequentist properties,
- their answers are more easily interpretable by non-specialists,
- they are already implemented in packages.

We will be able to appreciate these points later on after discussing some features and properties of Bayesian statistics.

According to an open letter by John K. Kruschke (Revised 14 November 2010),
Statistical methods have been evolving rapidly, and many people think it's time to adopt modern Bayesian data analysis as standard procedure in our scientific practice and in our educational curriculum. Three reasons:
1. Scientific disciplines from astronomy to zoology are moving to Bayesian data analysis. We should be leaders of the move, not followers.
2. Modern Bayesian methods provide richer information, with greater flexibility and broader applicability than 20th century methods. Bayesian methods are intellectually coherent and intuitive. Bayesian analyses are readily computed with modern software and hardware.
3. Null-hypothesis significance testing (NHST), with its reliance on p values, has many problems. There is little reason to persist with NHST now that Bayesian methods are accessible to everyone.

Furthermore, he writes that
** In NHST, the data collector must pretend to plan the sample size in advance and pretend not to let preliminary looks at the data influence the final sample size. Bayesian design, on the contrary, has no such pretenses because inference is not based on p values.*
** In contingency table analysis, the traditional chi-square test suffers if expected values of cell frequencies are less than 5. There is no such issue in Bayesian analysis, which handles small or large frequencies seamlessly. * In NHST, the power of an experiment, i.e., the probability of rejecting the null hypothesis, is based on a single alternative hypothesis. And the probability of replicating a significant outcome is "virtually unknowable" according to recent research. But in Bayesian analysis, both power and replication probability can be computed in straight forward manner, with the uncertainty of the hypothesis directly represented.*

Regarding the use of prior information, he states that
Some people may have the mistaken impression that the advantages of Bayesian methods are negated by the need to specify a prior distribution. In fact, the use of a prior is both appropriate for rational inference and advantageous in practical applications.
** It is inappropriate not to use a prior. Consider the well known example of random disease screening. A person is selected at random to be tested for a rare disease. The test result is positive. What is the probability that the person actually has the disease? It turns out, even if the test is highly accurate, the posterior probability of actually having the disease is surprisingly small. Why? Because the prior probability of the disease was so small. Thus, incorporating the prior is crucial for coming to the right conclusion.*
** Priors are explicitly specified and must be agreeable to a skeptical scientific audience. Priors are not capricious and cannot be covertly manipulated to predetermine a conclusion. If skeptics disagree with the specification of the prior, then the robustness of the conclusion can be explicitly examined by considering other reasonable priors. In most applications, with moderately large data sets and reasonably informed priors, the conclusions are quite robust.*
** Priors are useful for cumulative scientific knowledge and for leveraging inference from small-sample research. As an empirical domain matures, more and more data accumulate regarding particular procedures and outcomes. The accumulated results can inform the priors of subsequent research, yielding greater precision and firmer conclusions.*
** When different groups of scientists have differing priors, stemming from differing theories and empirical emphases, then Bayesian methods provide rational means for comparing the conclusions from the different priors.*

On the use of p-value he states that
Although there are many difficulties in using p values, the fundamental fatal flaw of p values is that they are ill defined, because any set of data has many different p values. [...] The literature is full of articles pointing out the many conceptual misunderstandings held by practitioners of NHST. For example, many people

mistake the p value for the probability that the null hypothesis is true. Even if those misunderstandings could be eradicated, such that everyone clearly understood what p values really are, the p values would still be ill defined. Every fixed set of data would still have many different p values.

Bayesian statistics is very used in many topics in life sciences, such as:

- genetics (e.g. fine-mapping of disease-susceptibility genes)
- ecology (e.g. agent-based models)
- evolution (e.g. inference of phylogenetic trees
- ...

1.2.1 Case study: overview

We will apply Bayesian statistics to a real case of scientific investigation. The ultimate goal of these exercises is to understand the evolutionary origin of polar bears (Figure 7). We are also interested in their genetic relationship with brown bears (Figure 8), their sister species.

Figure 7: Polar bears. What's their evolutionary history?

For this purpose, we collected genetic data for several samples across the whole high Arctic region (Figure 9) and we want to infer something about their shared history. Specifically, we are interested in when they diverge from each other and whether there has been exchange of genetic material after speciation.

In this experimental setting, once we collected our biological material we sequenced the genome of each sample. In these exercises we will use Bayesian methods to:

Figure 8: Brown bears. What's their genetic relationship with polar bears?

1. reconstruct genomes from sequencing data,
2. estimate frequencies of mutations and genome diversity,
3. infer speciation time and test for hybridisation.

These steps are illustrated in Figure 10.

Figure 9: Location of samples of polar and brown bears collected in the high Arctic.

2 Bayesian concepts

In the previous section we learned the rationale behind Bayesian statistical inference. We will now formalise Bayesian equations and provide case studies where such approach has been successfully used in ecological and/or evolutionary studies.

2.1 Bayes' Theorem

Suppose we have a random variable Y. Then $f(\vec{y}|\vec{\theta})$ is a probability distribution representing the sampling model for the observed data $\vec{y} = (y_1, y_2, ..., y_n)$ given a vector of unknown parameters $\vec{\theta})$. The distribution $f(\vec{y}|\vec{\theta})$ is often called the *likelihood* and sometimes written as $L(\vec{\theta}; \vec{y})$. We know that $L(\vec{\theta}; \vec{y})$ is not a probability distribution for $\vec{\theta}$ given $\vec{y}$. Therefore $\int L(\vec{\theta}; \vec{y}) d\vec{\theta}$ is not necessarily 1 or even finite. Nevertheless it is possible to find the value of $\vec{\theta}$ that maximises the likelihood function. In other words we can calculate a *maximum likelihood estimate* (MLE) for $\vec{\theta}$, as: $\hat{\vec{\theta}} = argmax_{\vec{\theta}} L(\vec{\theta}; \vec{y})$

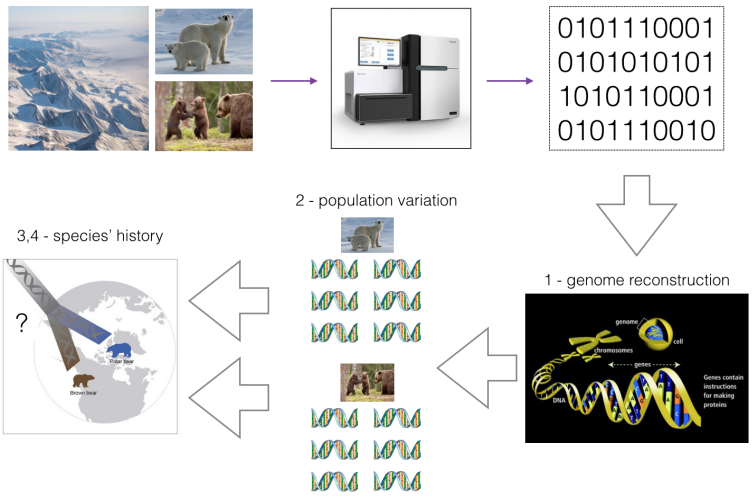

Figure 10: What is the evolutionary relationship between polar bears and brown bears? Insights from genomic data.

In Bayesian statistics, $\vec{\theta}$ is not a fixed (although unknown) parameter but a random quantity. This is done by adopting a probability distribution, called *prior distribution*, for $\vec{\theta}$ that contains any information we have about it not related to the data $\vec{y}$. Under this approach, inference on $\vec{\theta}$ is then based on its *posterior distribution* given by:

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

This formula is known as *Bayes' Theorem*. The posterior probability is simply the product of the likelihood and the prior, renormalised so that integrates to 1, and thus it is a valid probability distribution.

Let's prove this Theorem using a concrete example. We will also visualise probabilities using Venn diagrams. The greatest loss of vertebrate biodiversity we observed in the past 30 years is due to a chytrids fungus which is responsible for the extinction of over a hundred species of amphibians (Figure 11).

One of the easiest ways to understand probabilities is to think of them in terms of Venn Diagrams [4]. We have a Universe U with all the possible outcomes of an experiment and we are interested in some subset of them, namely some event. In our example we are interested in detecting which samples of frogs are infected or not by the fungus. Therefore we take samples and check whether they are infected or not. If we take as our Universe all the samples collected in a particular area, then there it is infected not. We can then split our Universe U in two events: the event "samples with infection" (designated as A), and

[4] https://oscarbonilla.com/2009/05/visualizing-bayes-theorem/

Figure 11: The chytrid fungus (*Batrachochytrium dendrobatidis*) is the most significant threat to amphibian populations.

"samples with no infection" (complement of A or A'). If so, we could build a corresponding diagram (Figure 12).

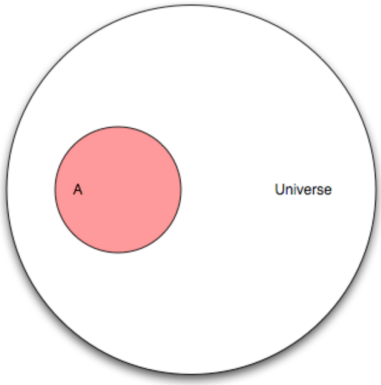

Figure 12: Sets U and A.

What is the probability that a randomly chosen samples is infected? It is the number of elements in A divided by the number of elements of U. We denote the number of elements of A as $|A|$, called the cardinality of A. We can define the probability of A, $P(A)$, as:

$$P(A) = \frac{|A|}{|U|} \tag{2}$$

with $0 \leq P(A) \leq 1$.

Let's add another event. Let's say that we use a molecular screening test which takes a biological samples and quantifies the presence of the fungus on the skin. This test will be "positive" for some samples, and "negative" for some other samples. If the event B is the collection of "samples for which the test is positive", then we can create another diagram (Figure 13).

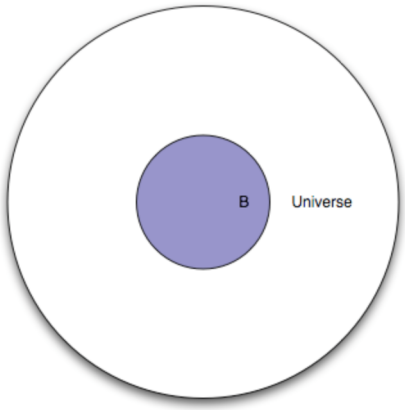

Figure 13: Sets U and B.

What is the probability that the test will be "positive" for a randomly selected samples? It will be:

$$P(B) = \frac{|B|}{|U|} \tag{3}$$

with $0 \leq P(B) \leq 1$.

So far we have treated the two events in isolation. What happens if we put them together? We can compute the probability of both events occurring (where AB is a shorthand for $A \cap B$) in the same way (Figure 14):

$$P(A \cap B) = \frac{|A \cap B|}{|U|} \tag{4}$$

with $0 \leq P(A \cap B) \leq 1$.

We are dealing with an entire Universe (all samples), the event A (samples with infection), and the event B (samples for whom the test is positive). There is also an overlap now. The event $A \cap B$ represents "samples with infection and with a positive test result". There is also the event $B - AB$ or "samples without infection and with a positive test result", and the event $A - AB$ or "samples with infection and with a negative test result".

The question we'd like to answer now is "given that the test is positive for a randomly selected sample, what is the probability that said samples is infected?". In terms of our Venn diagram, that translates to "given that we are in region B, what is the probability that we are in region $A \cap B$?". We can also say that "if we make region B our new Universe, what is the probability of A?". The notation for this probability is $P(A|B)$, "the probability of A given B". This probability is equal to:

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|U|}{|B|/|U|} = \frac{P(A \cap B)}{P(B)}$$

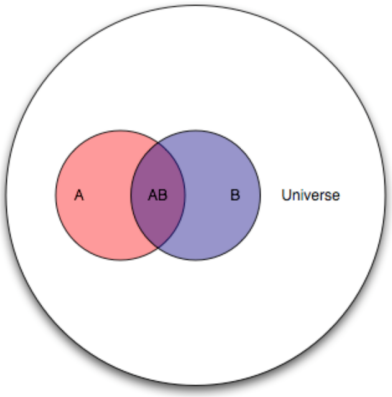

Figure 14: Sets U, A, B and $A \cap B$.

If we ask the converse question "given that a randomly selected samples is infected (event A), what is the probability that the test is positive for that sample (event $A \cap B$)?". This is:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

If we put together these last two equations (by $P(B \cap A)$) we obtain:

$$P(A)P(B|A) = P(B)P(A|B)$$

which shows that:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{5}$$

which is the Bayes' Theorem.

Now suppose that:

- 1% of samples collected are infected,
- 80% of samples with infection will get positive test 9.6% of samples with infection will also get positive tests.

A sample had a positive test. What is the probability that the sample is actually infected? The answer is (i) $P(A) = 0.01$, (ii) $P(B|A) = 0.8$, (iii) $P(B) = 0.8P(A)+0.096(1-P(A)) = 0.008+0.09504 = 0.10304$, which follows $P(A|B) = (0.8 * 0.01)/0.10304 = 0.0776$.

2.1.1 A normal/normal model

We now illustrate a simple Bayesian model to appreciate the interplay between likelihood and prior. Let's use again the example of infected frogs and imagine

we want to inspect a particular location and assess the rate of infection. In particular assuming we monitor 20 frogs in a given location and we want to make some inference on the number of infected frogs θ.

We consider the case where both the prior and the likelihood are normal (or Gaussian) distributions:

$$f(y|\theta) = N(y|\theta, \sigma^2)$$
$$\pi(\theta) = N(\theta|\mu, \tau^2)$$

where μ and τ are known *hyperparameters* in addition to the unknown parameter θ.

As we will see later, the posterior distribution $p(\theta|y)$ is also a normal distribution:

$$p(\theta|y) = N(\theta|\frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}) \tag{6}$$

If we write

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

then

$$E(\theta|y) = B\mu + (1 - B)y$$
$$Var(\theta|y) = (1 - B)\sigma^2 \equiv B\tau^2$$

B is called *shrinking factor* because it gives the proportion to how much the posterior mean is "shrunk back" from the classical frequentist estimate y towards the prior mean μ. Note that $0 \leq B \geq 1$.

The posterior mean is a weighted average of the prior mean μ and the direct estimate Y. The weight on the prior mean B depends on the relative variability of the prior distribution and the likelihood. If $\sigma^2 >> \tau^2$ then $B \approx 1$ and our prior knowledge is more precise than the data information. On the contrary, if $\sigma^2 << \tau^2$ then $B \approx 0$ and our prior knowledge is imprecise and the final estimate will move very little towards the prior mean.

We can appreciate this trade-off between data information and prior distribution with the following example. Assume that we have a single observation $y = 6$ and our likelihood function has $\sigma = 1$. Furthermore, our prior distribution has $\mu = 2$ and $\tau = 1$. More formally we can write:

$$f(y = 6|\theta) = N(y = 6|\theta, 1)$$
$$\pi(\theta) = N(\theta|2, 1)$$

We now use the `R` package and refer to the normal/normal model described above. Let's calculate and plot the prior, likelihood, and posterior distribution.

```
# prior
mu <- 2
tau <- 1
x <- seq(-4,10,0.01)
```

```
plot(x=x, dnorm(x=x, mean=mu, sd=tau), ylim=c(0,0.6),
type="l", lty=1, ylab="Density", xlab=expression(theta), main="")
legend(x="topleft", legend=c(expression(pi(theta)),
expression(f(y~"|"~theta)), expression(p(theta~"|"~y))), lty=1:3)

# likelihood
y <- 6
sigma <- 1
points(x=x, y=dnorm(x=y, mean=x, sd=sigma), type="l", lty=2)

# posterior
B <- sigma^2/(sigma^2+tau^2)
postMean <- B*mu + (1-B)*y
postVar <- B*tau^2
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)),
    type="l", lty=3)
```

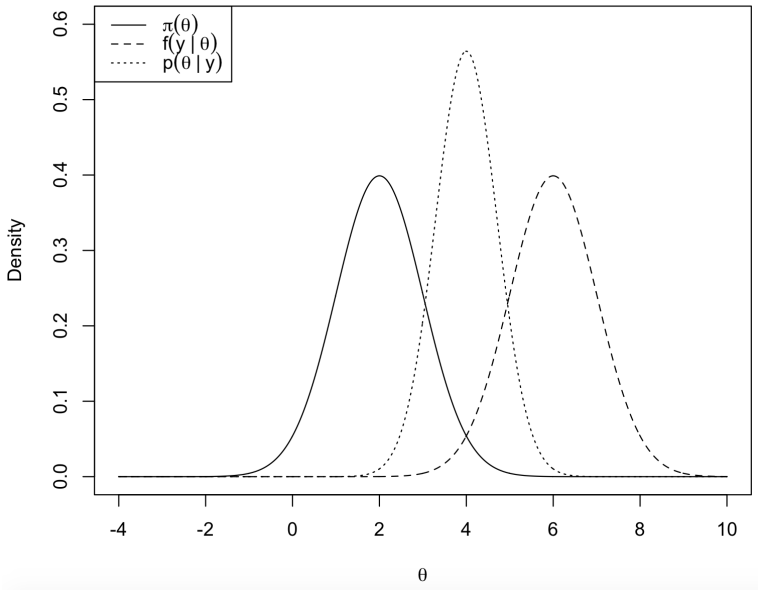

Figure 15: Prior, likelihood, and posterior distribution for the normal/normal model (see example in the text).

As you can see in Figure 15, the prior distribution is centred around 2 (μ), as expected. How about the likelihood function? It is centred around 6 which is the only observation we have. The posterior distribution is centred exactly in the middle because here $B = 0.5$ and therefore prior and data are equally weighted. The *maximum a posteriori probability* (MAP) estimate is 4, as it is equal to the mode of the posterior distribution. You can easily retrieve the MAP for this example in R with `x[which.max(dnorm(x=x, mean=postMean,`

`sd=sqrt(postVar))]`.

You may also observed that the posterior distribution is more skewed than the prior and the likelihood, despite these two have the same variance. The posterior variance is smaller than the variance of either the prior or the likelihood. The precision, defined as the reciprocal of the variance, is the sum of the precisions in the prior and likelihood. The combined strength of prior and likelihood tends to increase the precision, or reduce the variance, in our inference of θ. In this example, the precision is $1 + 1 = 2$, hence the variance is $1/2$. Therefore the posterior roughly covers $4 \pm 3(\sqrt{1/2}) \approx (1.88, 6.12)$

What happens if we use a skewer (sharper) or wider prior? In other words, what is the shape of the posterior distribution if the variance of the prior is smaller (stronger belief) or larger (weaker belief)? Recalling the previous example, now we assume that our prior distribution has $\mu = 2$ but $\tau = 0.5$ (before $\tau = 1$). Let's calculate and plot prior, posterior and MAP in this case.

```
# skewer prior
mu <- 2
tau <- 0.5
x <- seq(-4,10,0.01)
plot(x=x, dnorm(x=x, mean=mu, sd=tau), ylim=c(0,1),
type="l", lty=1, ylab="Density", xlab=expression(theta), main="")
legend(x="topleft", legend=c(expression(pi(theta)),
expression(f(y~"|"~theta)), expression(p(theta~"|"~y))), lty=1:3)

# likelihood (obviously it does not change)
y <- 6
sigma <- 1
points(x=x, y=dnorm(x=y, mean=x, sd=sigma), type="l", lty=2)

# posterior with skewer prior
B <- sigma^2/(sigma^2+tau^2)
postMean <- B*mu + (1-B)*y
postVar <- B*tau^2
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)),
    type="l", lty=3)

# MAP with skewer prior
map <-  x[which.max(dnorm(x=x, mean=postMean, sd=sqrt(postVar)))]
cat("MAP:", map, "(",
    map-3*sqrt(postVar),",",map+3*sqrt(postVar),")\n")
```

You can find that $\theta_{MAP} = 2.8$ with a range of $(1.46, 4.14)$ (Figure 16). You can appreciate how the posterior mean is now shifted towards the prior mean. Note that $B = 0.8$, putting more weight on the prior contribution.

Recalling the previous example, now we assume that our prior distribution has $\mu = 2$ but $\tau = 2$ (before $\tau \leq 1$). Let's calculate and plot prior, posterior and MAP in this case.

```
# wider prior
mu <- 2
tau <- 2
```

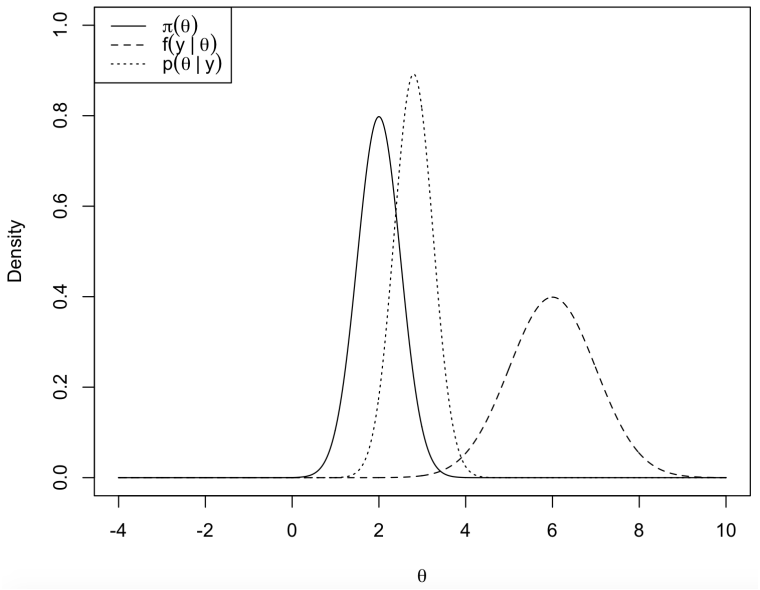

Figure 16: Prior, likelihood, and posterior distribution for the normal/normal model with a skewer prior (see example in the text).

```
x <- seq(-4,10,0.01)
plot(x=x, dnorm(x=x, mean=mu, sd=tau), ylim=c(0,0.5),
type="l", lty=1, ylab="Density", xlab=expression(theta), main="")
legend(x="topleft", legend=c(expression(pi(theta)),
expression(f(y~"|"~theta)), expression(p(theta~"|"~y))), lty=1:3)

# likelihood (obviously it does not change)
y <- 6
sigma <- 1
points(x=x, y=dnorm(x=y, mean=x, sd=sigma), type="l", lty=2)

# posterior with wider prior
B <- sigma^2/(sigma^2+tau^2)
postMean <- B*mu + (1-B)*y
postVar <- B*tau^2
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)),
    type="l", lty=3)

# MAP with wider prior
map <-  x[which.max(dnorm(x=x, mean=postMean, sd=sqrt(postVar)))]
cat("MAP:", map, "(",
    map-3*sqrt(postVar),",",map+3*sqrt(postVar),")\n")
```

You can find that $\theta_{MAP} = 5.2$ with a range of $(2.52, 7.88)$ (Figure 17). You can appreciate how the posterior mean is now shifted away from the prior mean. Note that $B = 0.2$, putting less weight on the prior contribution. In other words,

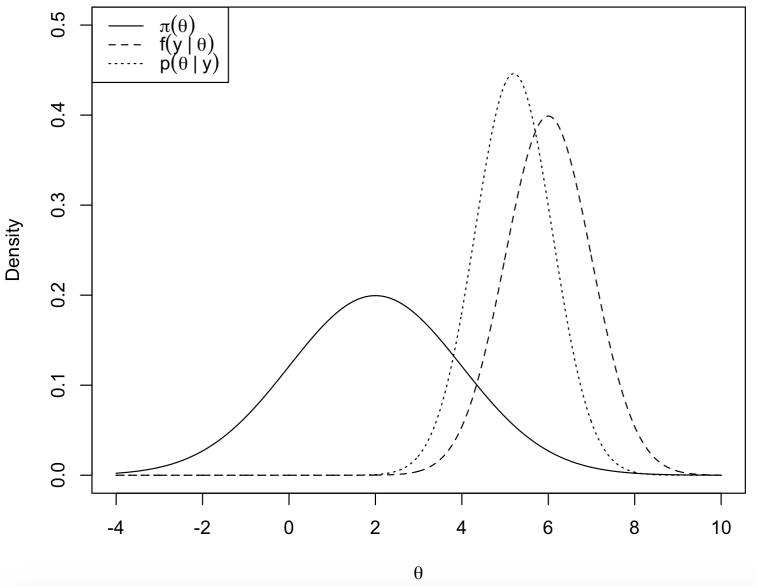

Figure 17: Prior, likelihood, and posterior distribution for the normal/normal model with a wider prior (see example in the text).

the data dominates the posterior more than the prior.

Let's look at a last case for this example. Imagine that we have more observations. For instance, we have have that $\vec{y} = 6, 5, 5.5, 6.5, 6$. Given a sample of n independent observations, then:

$$f(\vec{y}|\vec{\theta}) = \prod_{i=1}^{n} f(y_i|\vec{\theta}) \tag{7}$$

However, we can also use a transformation if we find a statistic $S(\vec{y})$ that is sufficient so that $p(\vec{\theta}|\vec{y}$.

In this example, $S(\vec{y}) = \bar{y}$, where $\bar{y}$ is the mean of $\vec{y}$. The likelihood function has the form $f(\bar{y}|/theta) = N(\theta, \sigma^2/n)$ and the posterior distributions is:

$$p(\theta|\bar{y}) = N(\theta|\frac{(\sigma^2/n)\mu + \tau^2\bar{y}}{(\sigma^2/n) + \tau^2}, \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n) + \tau^2} \tag{8}$$

Suppose we keep $\mu = 2$, $\sigma = \tau = 1$ and set $\bar{y} = 5.8$ and $n = 5$. Let's have a look at the resulting posterior distribution, using R.

```
# more observations (obviously the prior does not change)
mu <- 2
tau <- 1
x <- seq(-4,10,0.01)
plot(x=x, dnorm(x=x, mean=mu, sd=tau), ylim=c(0,1),
type="l", lty=1, ylab="Density", xlab=expression(theta), main="")
```

```
legend(x="topleft", legend=c(expression(pi(theta)),
expression(f(y~"|"~theta)), expression(p(theta~"|"~y))), lty=1:3)

# likelihood with more observations
y <- c(6, 5, 5.5, 6.5, 6)
n <- length(y)
sigma <- 1
points(x=x, y=dnorm(x=x, mean=mean(y), sd=sigma), type="l", lty=2)

# posterior with more observations
postMean <- ( (sigma^2/n)*mu + tau^2*mean(y) ) / ( (sigma^2/n)*mu
    + tau^2 )
postVar <- ( (sigma^2/n)*tau^2 ) / ( (sigma^2/n) + tau^2 )
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)),
    type="l", lty=3)

# MAP with more observations
map <-  x[which.max(dnorm(x=x, mean=postMean, sd=sqrt(postVar)))]
cat("MAP:", map, "(",
    map-3*sqrt(postVar),",",map+3*sqrt(postVar),")\n")
```

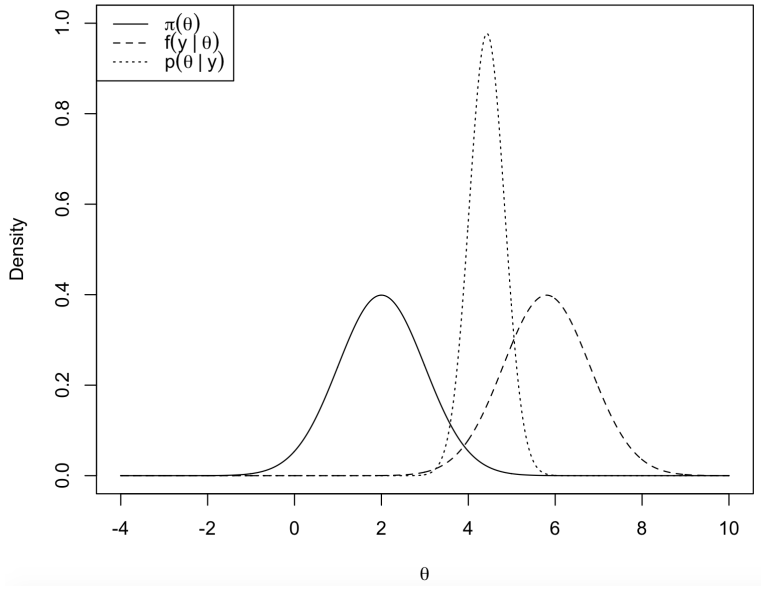

Figure 18: Prior, likelihood, and posterior distribution for the normal/normal model with more observations (see example in the text).

You can find that $\theta_{MAP} = 4.43$ with a range of $(3.21, 5.65)$ (Figure 18). The *MAP* has been shifted towards the MLE as we have more data information in this case.

2.1.2 Monte Carlo sampling

To derive the posterior distribution we can also draw random samples from it, instead of directly calculating the posterior mean and variance (if a normal distribution is considered, for instance). This procedure is often called *Monte Carlo sampling*, after the city famous for its casinos (Figure 19).

Figure 19: Monte Carlo and its famous casino.

In the previous example of the normal/normal model with multiple observations, we calculated the posterior mean (4.43) and posterior variance (0.17). From these parameters, we were able to derive (and plot) the density function, the posterior probability itself. Alternatively, we can randomly sample directly from the posterior distribution. Let's see how we can do this in R.

```
## monte carlo sampling

par(mfrow=c(3,1))

# posterior
x <- seq(2,8,0.01)
postMean <- 4.43
postVar <- 0.16
plot(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)), type="l",
     lty=1, ylab="Density", xlab=expression(theta),
     main=expression(p(theta~"|"~y)), ylim=c(0,1.2), xlim=c(2,8))

# sampling
y_sampled_1 <- rnorm(n=50, mean=postMean, sd=sqrt(postVar))
hist(y_sampled_1, breaks=20, freq=F, lty=2, col="grey",
     ylim=c(0,1.2), xlim=c(2,8), sub="50 samples",
     main=expression(p(theta~"|"~y)), xlab=expression(theta))
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)),
     type="l", lty=1)

# more sampling
y_sampled_2 <- rnorm(n=1e6, mean=postMean, sd=sqrt(postVar))
hist(y_sampled_2, breaks=20, freq=F, lty=2, col="grey",
     ylim=c(0,1.2), xlim=c(2,8), sub="1e6 samples",
     main=expression(p(theta~"|"~y)), xlab=expression(theta))
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)),
     type="l", lty=1, ylab="Density", xlab=expression(theta),
     main=expression(p(theta~"|"~y)), sub="1e6 samples")
```

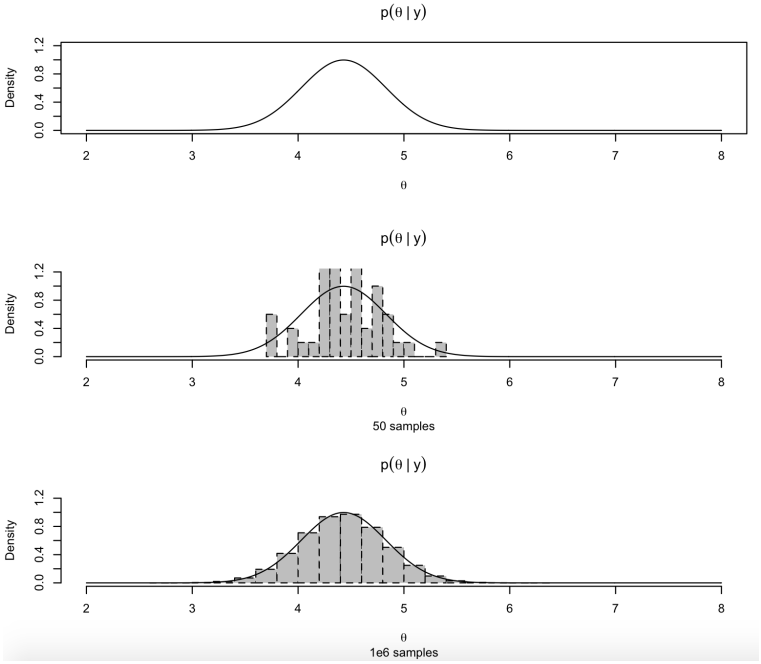

Figure 20: Posterior distribution and Monte Carlo sampling.

The more sampling we do, the closer our sampled distribution will look like the posterior (Figure 20). With 50 samples the empirical posterior mean is 4.444 while with $1e6$ samples we have an empirical posterior mean of 4.429 which is very close to our direct estimate of 4.43. With 50 samples the empirical posterior variance is 0.105 while with $1e6$ samples we have an empirical posterior variance of 0.159 which is very close to our direct estimate of 0.16.

You can understand how, in this simple normal/normal case, Monte Carlo methods are not necessary since the integral in the denominator of Bayes' Theorem can be evaluated in closed forms (as we will see later on). In these simple cases, we prefer to have a smooth curve rather an a histogram of sampled values, and have the corresponding exact answers for the posterior parameters. However there are cases where, given the choice for the likelihood and prior functions, this integral cannot be evaluated. Therefore, in these cases, Monte Carlo methods are to be preferred for estimating the posterior distribution. As any samples can be drawn from any posterior regardless of how many parameters θ you may have, we have the ability to work on problems with (theoretically) unlimited complexity, at the price of not obtaining an exact form for the posterior.

All scripts used in these examples are available at https://github.com/mfumagalli/BayesianMethods/blob/master/Examples_1.R.

2.1.3 CASE STUDY (1): reconstructing genomes from sequencing data

We want to use a Bayesian approach to reconstruct genomes from data produced from high-throughput sequencing machines. Instructions are available at https://github.com/mfumagalli/BayesianMethods/blob/master/Exercise_1.R.

A possible solution is available at https://github.com/mfumagalli/BayesianMethods/blob/master/Solutions_1.R.

2.2 Prior distributions

One of the main feature of Bayesian statistics is that we assign probability distributions not only to data variables $\vec{y}$ but also to parameters $\vec{\theta}$. In other words we quantify whatever feelings or believes we have about $\vec{\theta}$ before observing $\vec{y}$. Using Bayes' Theorem (Equation 2.1), we can obtain a posterior distribution of $\vec{\theta}$, a blend of the information between the data and the prior (e.g. Figures 15, 16, 17).

How can we decide which prior distribution is more appropriate in our study? As you can imagine this can be an arduous task and hampered the use of Bayesian statistics in the past. In general, prior distributions are derived from past information or personal opinions from experts. It is typical to restrict prior distributions to be distributed as familiar distribution families. Alternatively, one limit the prior distribution to bear little information. We now review some possibilities to assign prior distributions.

2.2.1 Elicited priors

The simplest approach to specify $\pi(\theta)$ is to define the collection of θ which are "possible". Then one can assign some probability to each one of these cases and make sure that they sum to 1. If θ is discrete, this looks like a natural approach.

For instance, imagine that your prior distribution represents the number of kits a specific mother rabbit will have in the next litter (Figure 21). Perhaps, you want to make some inference on the biological mechanisms on the number of kits. In this case you may have a likelihood function relating some observations $\vec{y}$ (e.g. genetic or environmental markers) to the number of kits θ. θ is clearly discrete

Figure 21: How many babies do rabbits have in one litter?

and you may have some past information on its distribution. For instance, from a literature search[5] you find that *Rabbits can have anywhere from one to 14 babies, also called kits, in one litter. An average litter size is six. Hereditary and environmental factors play a role in the number of kits born in a litter.* Firstly, you can assign

$$\pi(\theta = 0) = \pi(\theta > 14) = 0$$

[5]Actually, just doing a quick search on Google and reading www.reference.com/pets-animals

Secondly, you can impose that it is more probable that this mother will have 6 kits, as this is the average litter size based on past information. For instance,

$$\pi(\theta = 2) < \pi(\theta = 6) > \pi(\theta = 10)$$

Finally, you must assure that

$$\sum_{i=1}^{14} \pi(\theta = i) = 1$$

On the other hand, it θ is continuous, a simple solution would be to discretise our prior distribution, by assigning masses to intervals on the real line. In other words, you create a histogram prior for θ. Imagine that your prior distribution specifies the recorded temperature in hot springs at Lassen Volcanic National Park. Specifically, you are interested in relating the temperature of different pools at Bumpass Hell (Figure 22) with the occurrence of certain extremophile micro-organisms, capable of surviving in extremely hot environments. For instance, you may want to predict the temperature of the pool from

Figure 22: Bumpass Hell, hot springs and fumaroles at Lassen Volcanic National Park, California.

the observations of such micro-organisms. The latter distribution is the likelihood function while you also want to assign a prior distribution of the pool temperature, θ. Clearly θ is continuous and, from past observations, we know that it has a range of $(80, 110)$ with an average of 88, in Celsius scale. A simple solution here would be to derive a prior histogram of θ. For instance,

$$\pi(80 \geq \theta < 85) < \pi(85 \geq \theta < 90) > \pi(90 \geq \theta < 95)$$

Again, you have to make sure that all these probabilities sum to 1. Also, it is important that the histogram is sufficiently wide, as the posterior will have support only for values that are included in the prior.

Alternatively, we might assume that the prior for θ belongs to a parametric distributional family $\pi(\theta|\nu)$. Here, we choose ν so that $\pi(\theta|\nu)$ closely matches our elicited beliefs. This approach:

- reduces the effort to the elicitee (you don't have to decide a probability for each value θ can have);
- overcomes the finite support problem (as in the case of the histogram);
- may also lead to simplifications in the computation of the posterior (as we will see later).

A limitation of this approach is that it would be impossible to find a distribution that perfectly matches the elicitee's beliefs.

For instance, the prior of temperatures could be normally distributed $N(\mu, \sigma^2)$ bounded as $(80, 110)$:

$$\pi(\theta) = \begin{cases} 0 & \text{for } \theta < 80 \text{ or } \theta > 110 \\ N(\mu, \sigma^2) & \text{for } 80 \leq \theta \leq 110 \end{cases}$$

with $\mu = 88$ and $\sigma^2 = 10$ We can plot such distribution in `R` (Figure 23).

```
## elicited prior
mu <- 88
sigma2 <- 10
x <- seq(80,110,0.1)
plot(x=x, dnorm(x=x, mean=mu, sd=sqrt(sigma2)),
type="l", lty=1, ylab="Density", xlab=expression(theta),
main=expression(pi(theta)))
```

Note that this distribution is not defined outside the interval $(80, 110)$. Therefore, the posterior won't have mass outside this interval either. Overconfidence may result into failing to condition on events outside the range of personal experience or previous observations. For instance, the fact that a temperature lower than 80 was never observed may be better integrating by leaving a small (but greater than 0) probability of occurrence.

As a rule of thumb, for elicited priors, it is recommended to 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}). You should also assess the symmetry of your prior. Elicited priors can be updated and reassessed as new information is available. They are very useful for experimental design where some idea on the nature of the studied system must be input.

2.2.2 Conjugate priors

When choosing a prior distribution $\pi(\theta|\nu)$ some family distributions will make the calculation of posterior distributions more convenient than others. It is possible to select a member of that family that is *conjugate* with the likelihood $f(\vec{y}|\theta)$, so that the posterior distribution $p(\theta|\vec{y})$ belongs to the same distributional family as the prior. Let's illustrate this point with an example.

Suppose we are interested in modelling the arrival of herds of elephants to a specific water pond in the savannah in a day during the migratory season

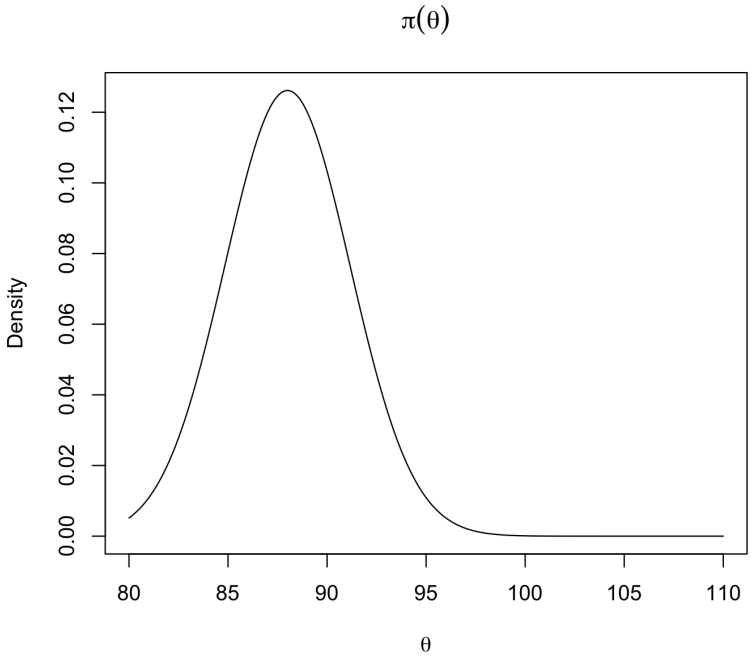

Figure 23: Elicited prior distribution of water temperature at Bumpass Hell pools.

(Figure 24) We may be interested in this estimate for tracking migratory routes or assessing population sizes. Y is the count of distinct elephant herds or groups (not the number of elephants itself!) arriving at the pool in a day during the migration season (not during the whole year!). A Poisson distribution has

Figure 24: Elephants drinking at the pool. What's the arrival rate for distinct herds?

a natural interpretation to model arrival rates for discrete variables. Indeed, the Poisson distribution is a discrete probability distribution that gives the probability of a given number of events occurring in a fixed interval of time (or

space) when such events occur independently and with a known average rate.

The Poisson distribution is an appropriate model under certain assumptions:

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.

Condition 1 is clearly met in our case. Conditions 2 and 4 assumes that different herds do not follow each other and somehow following distinct routes. For the sake of illustrating this distribution, we will assume this to be true. You can see that if Y were the number of elephants (not the herds) then condition 2 is not met as elephants tend to migrate in group. Condition 3 is met when we focus our analysis on the annual period where we expect to see herds, not during the whole year. Condition 5 is easily met, as the number of herds arriving in a week is likely to be higher than the number in a day. If we assume that all these conditions are true, then Y is a Poisson random variable.

The event under study (y, number of herds) can occur $0, 1, 2, ...$ times in the interval (a day). The average number of events (elephants arriving) in an interval (one day) is designated our parameter θ (the parameter of the Poisson distribution is typically written as λ). θ is the event rate, or the rate parameter. The probability of observing y events in an interval is given by the equation:

$$f(y|\theta) = \frac{e^{-\theta}\theta^y}{y!},\ y \in \{0, 1, 2, ...\},\ \theta > 0 \tag{9}$$

This is our likelihood distribution. Once we know θ, then the whole distribution is defined. As you can see, θ has to be positive (not necessarily an integer) and y is a positive integer. Note that Equation 2.2.2 is a probability mass function (pmf), as it is defined only for discrete values of y. For instance, assuming that, based on previous observations, the rate $\theta = 4$ (4 herds per day during migration season), then:

$$p(y = 0) = \frac{e^{-4}4^0}{0!} = e^{-4} = 0.0183$$
$$p(y = 1) = \frac{e^{-4}4^1}{1!} = ... = 0.0733$$
$$p(y = 2) = \frac{e^{-4}4^2}{2!} = ... = 0.147$$

We can plot our likelihood distribution when $\theta = 4$ using `R`.

```r
## Poisson
theta <- 4
y <- seq(0, 20, 1)
plot(x=y, dpois(x=y, lambda=theta), type="p", lty=1,
     xlab=expression(y), main=expression(theta~"="~4),
     ylab=expression(p(Y~"="~y~"|"~theta)))
```

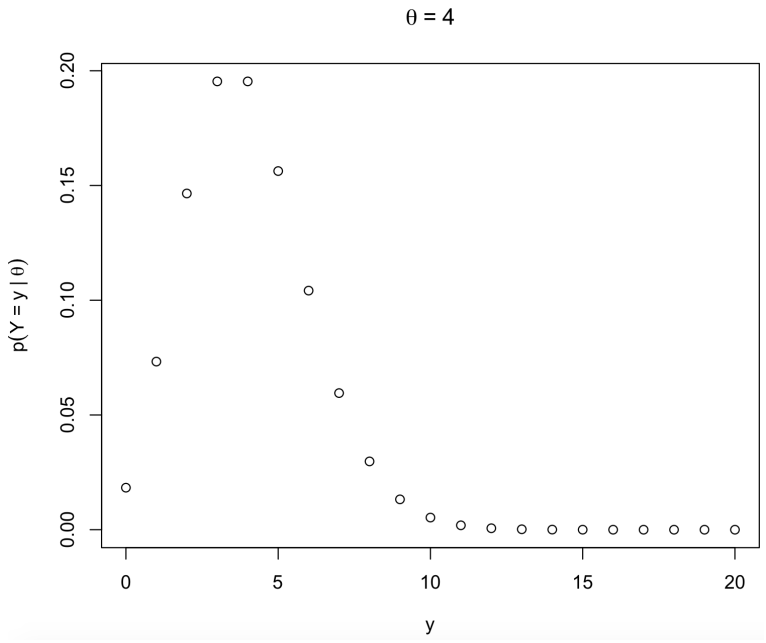

Figure 25: Poisson distribution for $\theta = 4$. This is the likelihood distribution for the number of herds per day with a rate of 4.

As you can see in Figure 25, the highest mass is towards 4 and above 12 the probability is very close to 0. Recall that a Poisson distribution has expected value and variance equal to the rate parameter. Note that we have some probability of having 0 events.

If we want to perform a Bayesian analysis, we need to define a prior distribution for θ having support for positive values. In other words, we assume that we don't know the exact value of θ, but we can make some assumptions on the probability distribution which if belongs to. A reasonable choice (as we will appreciate later) is given by the gamma distribution:

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \; \theta > 0, \; \alpha > 0, \; \beta > 0 \tag{10}$$

This is our prior distribution.

We can write it as $\theta \sim G(\alpha, \beta)$. As you can evince, the gamma distribution is a two-parameter family of continuous probability distributions. Be aware that the common exponential distribution and chi-squared distribution are special cases of the gamma distribution. We have also suppressed the dependency of π to hyperparameters $\nu = (\alpha, \beta)$ since we assume them to be known. α is known as the shape parameter while β is the rate parameter. Also note that

$$E(G(\alpha, \beta)) = \alpha\beta$$
$$Var(G(\alpha, \beta)) = \alpha\beta^2$$

Let's have a look at the shape of the Gamma distribution for different values of its parameters using `R`.

```
## Gamma
alpha <- c(0.5,2,10)
beta <- c(0.5,2)
thetas <- seq(0, 20, 0.1)
mycolors <- topo.colors(6)
plot(x=thetas, dgamma(x=thetas, shape=alpha[1], scale=beta[1]),
    type="l", lty=1, xlab=expression(theta), main="Gamma",
    ylab=expression(pi(theta)), ylim=c(0,1.0), col=mycolors[1],
    lwd=2)
index <- 0
for (i in alpha) {
  for (j in beta) {
      index <- index+1
      points(x=thetas, dgamma(x=thetas, shape=i, scale=j),
          col=mycolors[index], ty="l", lwd=2)
    }
}
names <- cbind(rep(alpha,each=2),rep(beta))
legend(x="topright", legend=apply(FUN=paste, MAR=1, X=names,
    sep="", collapse=", "), col=mycolors, lty=1, lwd=2)
```

As you can see in Figure 26, the gamma distribution is very flexible and it can have one tail ($\alpha \leq 1$) or two tails ($\alpha > 1$). For very large values of α the gamma distribution resembles a normal distribution. The β parameter shrinks or stretches the distribution relative to 0 but it doesn't change its shape.

Using the Bayes' theorem, we can now obtain the posterior probability:

$$\begin{aligned} p(\theta|y) &\approx f(y|\theta)/\pi(\theta) \\ &\approx (e^{-\theta}\theta^y)(\theta^{\alpha-1}e^{-\theta/\beta}) \\ &= \theta^{y+\alpha-1}e^{-\theta(1+1/\beta)} \end{aligned}$$

Since we are only interested in a normalised function of θ, we dropped all functions that do not depend on θ. We can realise that this posterior distribution is actually a gamma distribution $G(\alpha', \beta')$ with $\alpha' = y + \alpha$ and $\beta' = (1 + 1/\beta)^{-1}$. We could do this because gamma is the conjugate family for the Poisson likelihood.

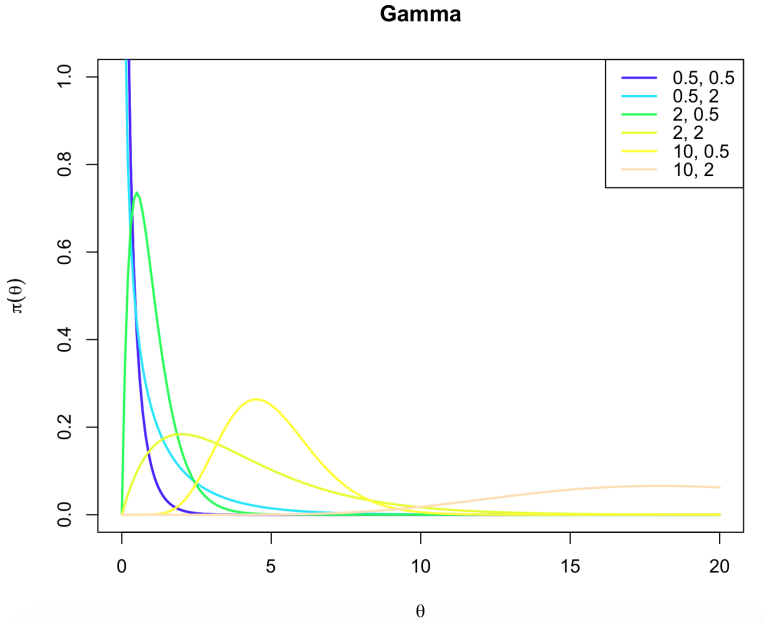

Figure 26: Gamma distribution for different values of shape and rate parameters.

To finalise our example of elephant herds, suppose that, before looking at the actual data, we have some intuition that we expect to see 3 herds per day. This could be derived from observations on previous years. Let's also assume that we are not very confident and we assign a large variance on it. For instance, our prior distribution is $G(3, 1)$ which points to an expected value and variance of 3. Assuming that we observed 4 herds then $y = 4$. Therefore, our posterior distribution will be a gamma distribution $G(3 + 4, 1/(1 + 1/1)$.

Let's plot these distributions in R.

```
## Gamma posterior
alpha <- 3
beta <- 1
theta <- seq(0, 20, 0.1)
prior <- dgamma(x=theta, shape=alpha, scale=beta)
y <- 4
posterior <- dgamma(x=theta, shape=y+alpha, scale=1/(1+1/beta))
plot(x=theta, y=posterior, xlab=expression(theta), ylab="Density",
    type="l")
lines(theta, prior, lty=3)
postdraw <- rgamma(n=1e5, shape=y+alpha, scale=1/(1+1/beta))
histdraw <- hist(postdraw, breaks=20, plot=F)
lines(histdraw, lty=3, col="grey", freq=F)
```

You can see these distributions in Figure 27. We also plot the posterior distribution by Monte Carlo sampling. As we could appreciate, conjugate priors

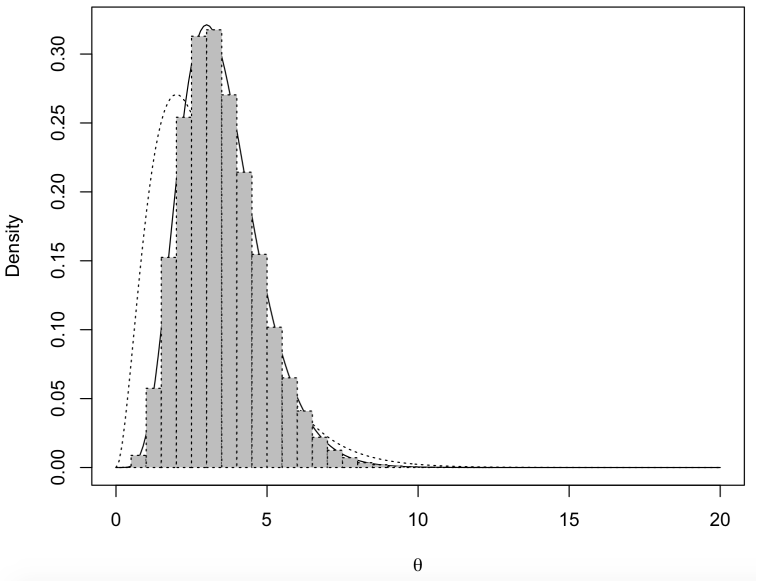

Figure 27: Prior and posterior distribution of θ, the number of herds per day

allow for posterior distributions to emerge without numerical integration.

2.2.3 Hierarchical modelling

There are many ways to build a prior distribution. A posterior distribution is typically obtained with two stages, one for $f(\vec{y}, \vec{\theta})$, the likelihood of the data, and $\pi(\vec{\theta}, \vec{\nu})$, the prior distribution of $\vec{\theta}$ given a vector of *hyperparameters* $\vec{\nu}$. Suppose we are unsure about the values of $\vec{\nu}$. In this case we need an additional stage, a *hyperprior* defining the density distribution of hyperparameters. If we denote this distribution as $h(\vec{\nu})$, then the posterior distribution is

$$p(\vec{\theta}|\vec{y}) = \frac{\int f(\vec{y}|\vec{\theta})\pi(\vec{\theta}|\vec{\nu})h(\vec{\nu})d\vec{\nu}}{\int\int f(\vec{y}|\vec{\theta})\pi(\vec{\theta}|\vec{\nu})h(\vec{\nu})d\vec{\nu}d\vec{\theta}} \tag{11}$$

Another possibility is to replace $\vec{\nu}$ with an estimate $\hat{\vec{\nu}}$ obtained by maximising the marginal distribution $m(\vec{y}|\vec{\nu})$. Now inferences are made on the *estimated posterior* $p(\vec{\theta}|\vec{y}, \hat{\vec{\nu}})$, by putting $\hat{\vec{\nu}}$) in the Bayes' Theorem equation. This approach is called *Empirical Bayesian* analysis as we are using the data to estimate the hyperparameter.

The empirical estimation of the prior seems a violation of Bayesian principles. The update of the prior based on the data would use of data twice, both for the likelihood and the prior. Inferences from such modelling tend to be "overconfident" and methods that ignore this fact are called *naive Empirical Bayesian* approaches.

Furthermore, we can again think of $\vec{\nu}$ depending on a collection of unknown parameters $\vec{\lambda}$, with $h(\vec{\nu}|\vec{\lambda})$ and a third-stage prior $g(\vec{\lambda})$. This procedure of specifying a model over several layers is called *hierarchical modelling*. This framework is very much used in graphical modelling. As we add extra layers and levels of randomness, subtle changes at the top levels (hyperpriors) will not have a strong effect at the bottom level (the data).

2.2.4 Non-informative priors

It is often the case that no reliable prior information on $\vec{\theta}$ is available. Can we employ a Bayesian approach under such circumstances? It is still appropriate if we find a distribution $\pi(\vec{\theta})$ that contains "no information" about $\vec{\theta}$, in the sense that it does not favour one values over another. We refer to such a distribution as a *noninformative prior* for $\vec{\theta}$. All the information in the posterior will arise from the data.

Suppose that the parameter space is discrete and finite $\vec{\Theta} = \{\theta_1, \theta_2, ..., \theta_n\}$. What is a possible noninformative prior? It can be:

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

Such prior distribution does not favour any value and therefore it is noninformative about θ. Moreover,

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

If $\vec{\Theta}$ is continuous and bounded, then $\vec{\Theta} = [a, b]$ with $-\infty < a < b < +\infty$. In this case a uniform prior in the form:

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

is a noninformative prior distribution, although it is less clear than in the discrete case if that's true.

How about the case of $\vec{\Theta}$ being continuous and unbounded, so that $\vec{\Theta} = (-\infty, +\infty)$? A noninformative prior could be set as:

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

However, this distribution is clearly *improper* as

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

This may suggest that in this case a Bayesian approach cannot be used. However a Bayesian inference is still possible if the integral of of the likelihood $f(\vec{y}|\theta)$ with respect to θ equals some finite value K. Indeed,

$$\int \frac{f(\vec{y}|\theta) \cdot c}{\int f(\vec{y}|\theta) \cdot c d\theta} d\theta = 1 \tag{12}$$

Let's go back to our example on elephant herds. Suppose that we don't have information on past rates of arrivals and therefore we cannot (or don't want to) use the gamma distribution as prior distribution as previously done. Instead we use a uniform prior U for the mean arrival θ, our parameter. What is the interval of θ? Theoretically it can be $[0, +\infty)$. However let's limit it to a high number and define a uniform prior distribution as $U(0, 100)$.

Let's plot this distribution in R.

```
## Uniform
theta <- seq(0, 100, 0.1)
prior <- dunif(x=theta, min=0, max=100)
plot(x=theta, y=prior, xlab=expression(theta), ylab="Density",
     type="l")
```

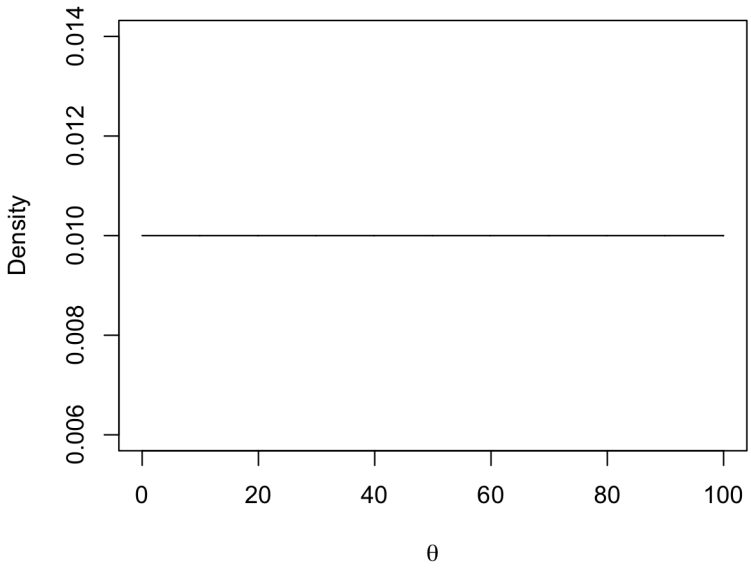

Figure 28: Uniform distribution, a noninformative prior distribution.

You can see the distribution in Figure 28. This choice rules out scenarios that are impossible in real life. This means that the posterior will be truncated at 0 and 100 too. We lack a conjugate model here. In this case we can sample from the posterior to obtain its distribution. We will see later how we can do this.

Noninformative priors are related to the notion of *reference* priors. These are not necessarily noninformative but a convenient, default choice for prior distributions. We will see later how to use non-conjugate priors for deriving the posterior distribution.

2.3 Bayesian inference

Once we have specified the prior, we can use Bayes' Theorem to obtain the posterior distribution of model parameters. However the density (or cumulative) function can be difficult to interpret. Therefore we want to summarise the information enclosed in these distributions. In particular here we discuss how to develop Bayesian techniques for point estimation, interval estimation, and hypothesis testing.

2.3.1 Point estimation

Let's first consider the univariate case. We want to select a summary feature of $p(\theta|\vec{y})$ to obtain a point estimate $\hat{\theta}(\vec{y})$. This feature can be either its mean, median, or mode. These features may behave very differently depending on the distribution, especially when it is asymmetric and one-tailed. Recalling the example of elephant herds, then assuming $y = 1$ and our prior is $G(0.5, 1)$, let's calculate and plot the posterior distribution.

```
## Gamma posterior asymmetric, one-tailed prior
alpha <- 0.5
beta <- 1
theta <- seq(0, 20, 0.1)
prior <- dgamma(x=theta, shape=alpha, scale=beta)
y <- 1
posterior <- dgamma(x=theta, shape=y+alpha, scale=1/(1+1/beta))
plot(x=theta, y=posterior, xlab=expression(theta), ylab="Density",
    type="l")
lines(theta, prior, lty=3)
```

What is the mean, mode and median of the resulting posterior distribution in Figure 29?

Generally speaking, the mode is the easiest to calculate. Indeed, since no normalisation is required, we can work directly with the numerator. Note that if the prior distribution is flat, like in Figure 28, then the *posterior mode* will be equal to the maximum likelihood estimate of θ. In this case, it is called the generalised maximum likelihood estimate of θ.

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 like the mean and it is an intermediate to the mode and the mean.

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

$$Var_{\theta|\vec{y}}(\theta) = E_{\theta|\vec{y}}[\theta - E_{\theta(\vec{y})}]^2 \tag{13}$$

The posterior mean minimises the posterior variance in respect to $\hat{\theta}(\vec{y})$.

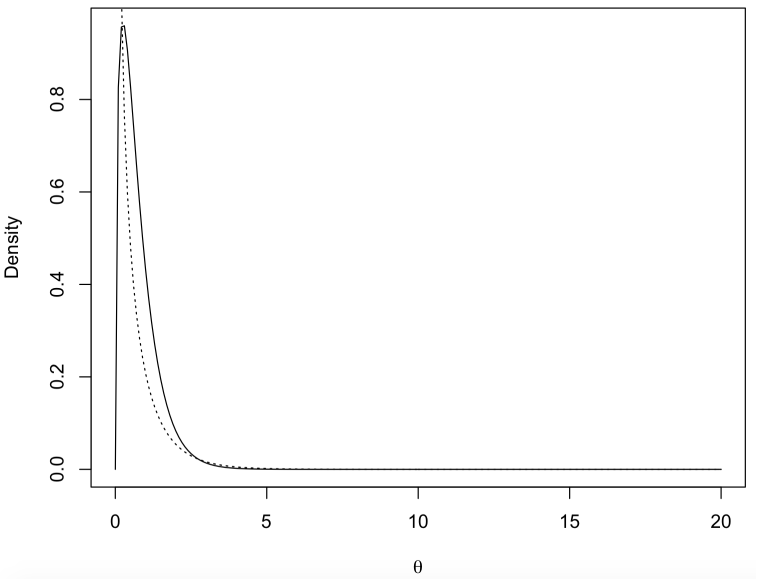

Figure 29: Posterior distribution using a one-tailed gamma prior distribution.

In the multivariate case, we can obtain the posterior mode as $\hat{\vec{\theta}}(\vec{y}) = (\hat{\theta_1}, \hat{\theta_2}, ..., \hat{\theta_k})$. If the mode exists, maximisation methods (e.g. grid search, golden section search, Newton-type methods, ...) are typically employed to locate the maximum. One can again calculate the posterior mean which, again, minimises the *posterior covariance matrix* with respect to $\hat{\vec{\theta}}(\vec{y})$.

2.3.2 Credible intervals

The Bayesian analogue of a frequentist confidence interval is called a *credible set*. A $100 \times (1 - \alpha)$ credible set for $\vec{\theta}$ is a subset C of $\vec{\Theta}$ such that:

$$1 - \alpha \leq P(C|\vec{y}) = \int_C p(\vec{\theta}|\vec{y}) d\vec{\theta} \tag{14}$$

in the continuous case. In the discrete case the integral is replaced by a summation. This definition can express a likelihood of θ falling in C: *"The probability that θ lies in C given the observed data y is at least $(1 - \alpha)$"*. Unlike the frequentist case, the credible set provides an actual probability statement, based on both the observed data and prior opinion.

In discrete settings it may not be possible to find the exact coverage probability $(1 - \alpha)$. In continuous settings we can calculate the *highest posterior density*, or HPD, credible set, defined as:

$$C = \{\theta \in \Theta : p(\theta|y) \geq k(\alpha)\} \tag{15}$$

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

Let's recall the example of elephants herds and assume that the posterior distribution $p(\theta|\vec{y}) \sim G(2,1)$.

```
## Interval estimation
theta <- seq(0, 10, 0.05)
alpha <- 2
beta <- 1
posterior <- dgamma(x=theta, shape=alpha, scale=beta)
plot(x=theta, y=posterior, xlab=expression(theta), ylab="Posterior
    density", type="l")
```

There are several ways to define a credible interval. In the HPD, we choose the narrowest interval, which for a unimodal distribution will involve choosing those values of highest probability density. This will include the mode for a unimodal distribution. For instance, drawing a line at $k(\alpha) = 0.1$ results in a 87% HPD.

```
abline(h=0.1, lty=3)
```

We can calculate the interval of values of θ included in the 87% HPD.

```
library(coda)
x <- rgamma(n=1e5, shape=alpha, scale=beta)
hpd <- HPDinterval(as.mcmc(x), prob=0.87)
abline(v=hpd, lty=1)
```

Another common strategy (often used) is to choose the interval where the probability of being below it is as likely as being above it. For instance, if $a = 1 - 0.87$ assuming a 87% HPD, then the *equal-tailed interval* corresponds to the $\{a/2, 1-a/2\}$- quantiles of the distribution. This interval includes the median. If the distribution is symmetric, both credible intervals will be the same. In R, we can easily calculate such quantiles.

```
a <- 1-0.87
eqi <- quantile(x, probs=c( a/2, 1-(a/2) ) )
abline(v=eqi, lty=2)
```

How can we then summarise our results? Typically, we can report:

- 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, ...

2.3.3 Hypothesis testing

The frequentist approach to compare predictions made by alternative scientific explanations is based on classic ideas of Fisher, Neyman and Pearson. Typically, one formulates a null hypothesis H_0 and an alternative hypothesis H_a. Then an appropriate test statistic is chosen $T(\vec{Y})$. Finally, one computes the *observed significance*, or *p-value*, of the test as the chance that $T(\vec{Y})$ is "more extreme" that $T(\vec{y}_{obs})$, where the "extremeness" is towards the alternate hypothesis. 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.

While widely used, there are few criticisms to such classic approach:

1. it is applied only when two hypotheses are nested, one within the other; typically, H_0 is a simplification of H_a and involves setting one parameter of H_a to some known constant value;

2. it offers evidence *against* the null hypothesis; 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. p-values don't offer a direct interpretation in terms of weight of evidence, but only as a long-term probability; p-values are not the probability that H_0 is true.

The Bayesian approach to hypothesis testing is much simpler and more intuitive. Basically, one calculates the posterior probability that the first hypothesis is correct. In general, one can test as many models as desired, $M_i, i = 1, ..., m$.

Suppose we have two models M_1 and M_2 for data Y and the two models have parameters $\vec{\theta}_1$ and $\vec{\theta}_2$. With prior densities $\pi_i(\vec{\theta}_i)$, with $i = 1, 2$, the marginal distributions of Y are:

$$p(y|M_i) = \int f(y|\theta_i, M_i)\pi_i(\theta_i)d\theta_i \tag{16}$$

Then Bayes' Theorem can be used to calculate the posterior probabilities $P(M_1|y)$ and $P(M_2|y) = 1 - P(M_1|y)$ for the two models. 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:

$$BF = \frac{P(M_1|y)/P(M_2|y)}{P(M_1)/P(M_2)} = \frac{p(y|M_1)}{p(y|M_2)} \tag{17}$$

which turns to be the ratio of observed marginal densities for the two models (note that $p(M_i)$ is the prior probability). If the two models are *a priori* equally probable then:

$$BF = p(M_1|y)/p(M_2|y) \tag{18}$$

which is the posterior odds of M_1.

The interpretation of BF is that it captures the change in the odds in favour of model 1 as we move from the prior to the posterior (Table 2).

Table 2: Bayes factors

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

2.3.4 CASE STUDY / EXERCISE (2): population variation

We now suppose that we have sequenced our bears' genomes and, using the method in Exercise 1, assigned each individual genotype. We now address a further question: what is the frequency of a certain allele at the population level? Be aware that we have only a sample of the entire population of bears but we want to make inferences at the whole population level.

Our sample contains information for 100 individuals with the following genotypes: 63 AA, 34 AT, 3 TT. A frequentist estimate of the frequency of T is given by: $(34 + (3 * 2))/200 = 40/200 = 0.20$. What is the posterior distribution for the population frequency of T?

The first thing we need to do is define our likelihood model. We can think of randomly sample one allele from the population and each time the allele can be either T or not. This is a set of Bernoulli trials and we can use of Binomial distribution as likelihood function.

The Binomial likelihood is:

$$p(k|p, n) = \binom{n}{k} p^k (1-p)^{n-k}$$

where k is the number of successes (i.e. the event of sampling a T), p is the proportion of T alleles we have (i.e. the probability of a success), and n is the number of alleles we sample, and:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Note that the combinatorial term does not contain p.

What is the maximum likelihood estimate of p? You may recall that it is $\hat{p} = \frac{k}{n}$. Note that the combinatorial terms does not affect this estimate.

The second thing we need to do is define a prior probability for p. What is the interval of values that p can take? It is $[0, 1]$.

It may be convenient to choose conjugate prior to the binomial. What is a convenient conjugate prior probability for the binomial distribution? A Beta distribution is a conjugate prior.

Are certain values of p more likely to occur without observing the data, a priori? If we assume that it's not the case, can we use the Beta distribution to generate a noninformative prior? We can choose $Beta(\alpha = 1, \beta = 1)$, which is defined as:

$$p(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

where $\frac{1}{B(\alpha,\beta)}$ is simply a normalisation term which does not depend on p.

The full model can be expressed as $p(p|k, n) \approx P(k|p, n)P(p)$. What is the closed form for the posterior distribution given our choices for the likelihood and prior functions? It is:

$$p(p|k, n) \approx p^{k+\alpha-1}(1-p)^{n-k+\beta-1}$$

.

The posterior distribution (beta-binomial model) is a Beta distribution with parameters $k + \alpha$ and $n - k + \beta$. If we set $\alpha = \beta = 1$ then $p(p|k, n) = Beta(k + 1, n - k + 1)$. What is k and n again?

Write a R code to plot this posterior probability. Then calculate the maximum a posteriori value, 95% credible intervals, and notable quantiles. What happens if we have only 10 samples (with the sample allele frequency of 0.20)?

Now let's think of a more informative prior. Look at the genome-wide distribution of allele frequencies for human populations (Figure 30). This is called a site frequency spectrum (SFS) or allele frequency spectrum (AFS). We can have another view by plotting the minor allele counts (MAC) distribution (Figure 31).

Does this distribution fit with a uniform prior? Can we use a conjugate (beta) function to model this distribution? For instance, choosing $\alpha = 0.5$ and $\beta = 2$ will put more weights on low-frequency variants. However, we don't know a priori whether the allele we are interested in is the minor allele. Therefore a prior distribution with more density at both low and high frequencies might be more appropriate. For instance, this could be achieved by setting $\alpha = 0.1$ and $\beta = 0.1$.

Recalculate the posterior distribution of p using an informative prior both in the case of 100 and 10 samples. Compare the results.

Assuming that for diagnostic use, one wants to classify a mutation as rare or medium depending on its allele frequency (e.g. 10%). From the posterior distributions above (both using a noninformative and informative prior), calculate Bayes factors for models $p(p|k, n) \geq 0.1$ and $p(p|k, n) < 0.1$. Therefore, $M_1 : p \geq 0.1$ and $M_1 : p < 0.1$. and:

$$BF = \frac{P(M_1|k, n)/P(M_2|k, n)}{P(M_1)/P(M_2)}$$

Use only 10 individuals and add a prior distribution with more density for intermediate allele frequencies (e.g. $\alpha = \beta = 2$).

Write a code in R to calculate these Bayes factors. Create a table with 2.5%, 50% and 97.5% percentiles for the posterior distribution as well as $p(p|k, n) \geq 0.1$ and Bayes factors.

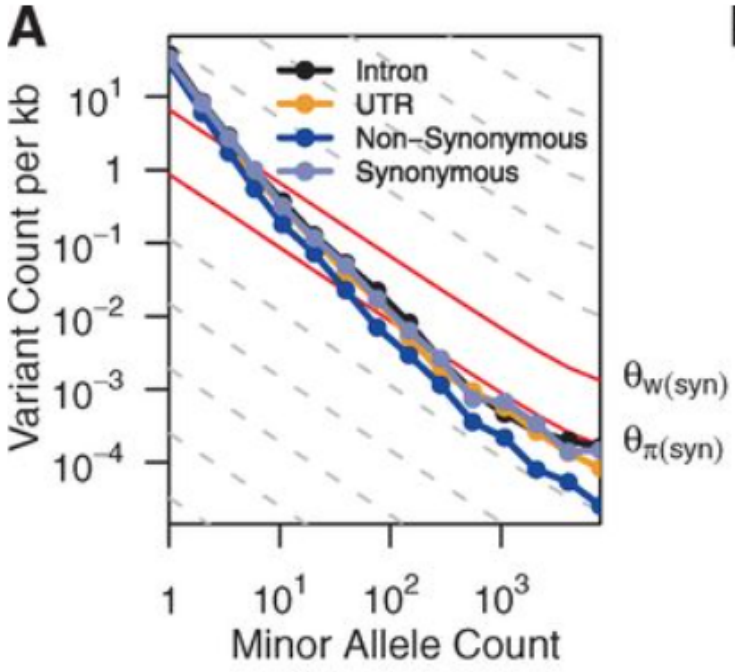

Figure 30: From Nelson et al. 2012 Science. *Frequency spectrum of variants relating the number of variants per kilobase within minor allele counts. Solid red lines provide expectations from nucleotide diversity (α_π) and the number of segregating sites (α_W).*

3 Bayesian computation

As seen before, the calculation of posterior distributions often involves the evaluation of complex high-dimensional integrals (e.g. the denominator of classic Bayes' formula). This is particularly true when a conjugate prior is not available or appropriate. The two ways of addressing this issue are (i) asymptotic methods for approximating the posterior density and (ii) numerical integration.

3.1 Asymptotic methods

When there are many data points, the likelihood will be peaked and the posterior distribution won't be affected by the prior too much. Therefore, small changes in the prior will have little effect on the posterior and the likelihood will be concentrated in a small region.

Let's recall again the case of the beta-binomial model of allele frequencies. If we have more data (individuals) then the posterior (and the likelihood) will be skewer (Figure 32).

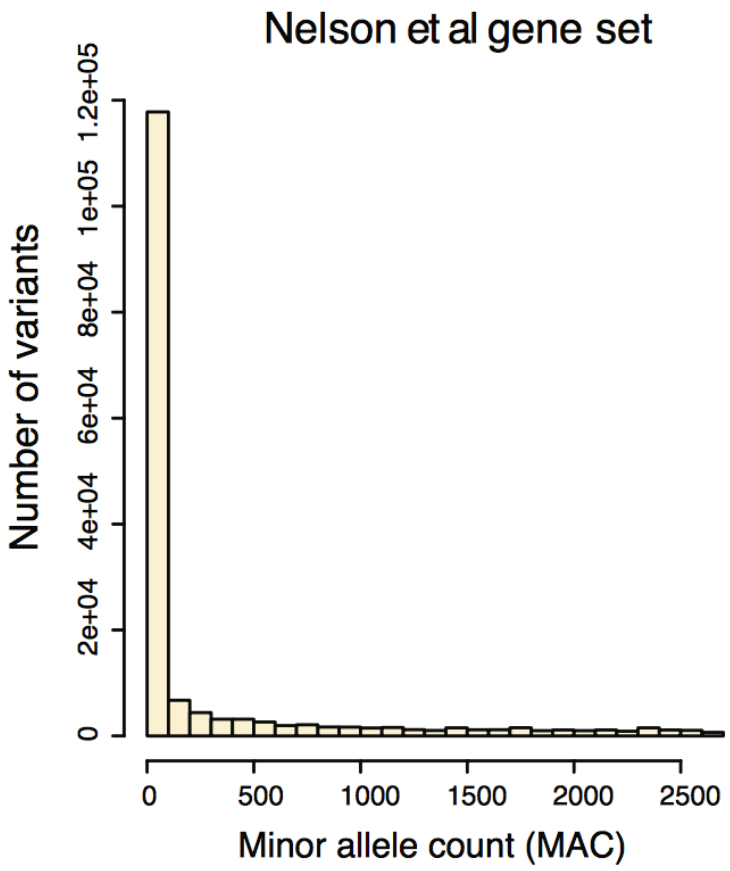

Figure 31: From Moutsianas et al. 2015 PLoS Genetics. Minor allele counts in Nelson et al. 2012.

```
# beta-binomial model of allele frequencies
p <- seq(0, 1, 0.01)

# 1000 chromosomes (20 derived alleles)
k <- 200
n <- 1000
alpha <- k+1
beta <- n-k+1
plot(x=p, y=dbeta(p, shape1=alpha, shape2=beta), ylab="Posterior
    density", xlab="Population frequency of T", type="l")

# 100 chromosomes
k <- 20
n <- 100
alpha <- k+1
```

```
beta <- n-k+1
points(x=p, y=dbeta(p, shape1=alpha, shape2=beta), type="l", lty=2)

# 10 chromosomes
k <- 2
n <- 10
alpha <- k+1
beta <- n-k+1
points(x=p, y=dbeta(p, shape1=alpha, shape2=beta), type="l", lty=3)

legend("topright", legend=c(1e3,1e2,1e1), lty=1:3)
```

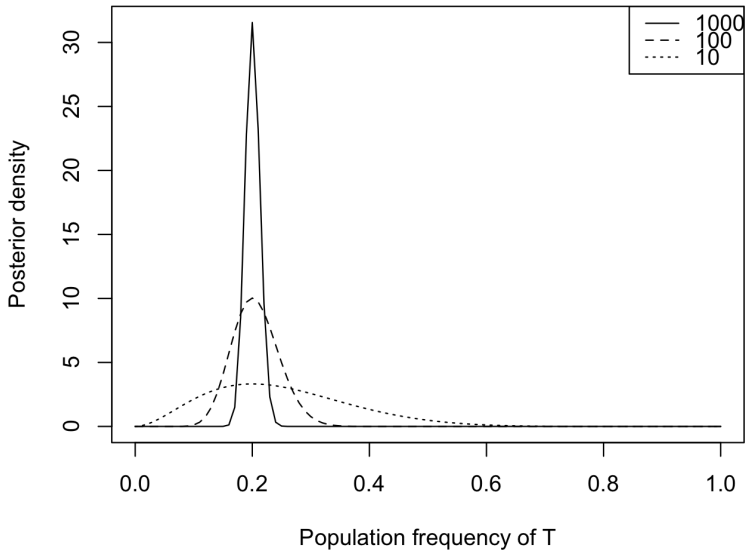

Figure 32: Posterior distribution of allele frequencies for increasing data points (n, number of chromosomes). For large n the posterior can be approximated by a Normal distribution.

In this situation, $p(\theta|x)$ will be approximately normally distributed. This is given by the *Bayesian Central Limit Theorem*. More formally, for large data points n 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.

For instance, recalling our beta-binomial model for allele frequencies, if we use a flat beta prior, we have $p(\theta|x) \approx \theta^x(1-\theta)^{n-x}$. The approximation follows the procedure:

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 $\hat{\theta}$, $-\frac{n}{\hat{\theta}} - \frac{n}{1-\hat{\theta}}$

4. take the minus inverse, $\frac{\hat{\theta}(1-\hat{\theta})}{n}$

5. $p(\theta|x) \sim N(\hat{\theta}^{\pi}, \frac{\hat{\theta}(1-\hat{\theta})}{n})$

Remember that the derivative of $log(x)$ is $1/x$ and that the derivative of $log(1-x)$ is $1/(x-1)$.

Let's plot the exact posterior distribution and the normal approximation.

```
# beta-binomial model of allele frequencies
p <- seq(0, 1, 0.01)

# 100 chromosomes (20 derived alleles)
k <- 20
n <- 100

# exact posterior with flat prior
alpha <- k+1
beta <- n-k+1
plot(x=p, y=dbeta(p, shape1=alpha, shape2=beta), ylab="Density",
    xlab="Population frequency", type="l")

# normal approximation
thetaHat <- k/n
var <- thetaHat*(1-thetaHat)/n
points(x=p, y=dnorm(p, mean=thetaHat, sd=sqrt(var)), type="l",
    lty=2)

legend("topright", c("exact","approx"), lty=1:2)
```

You can see how the modes are very similar (Figure 33) but the approximated curve fails to fully capture the asymmetry of the tails.

Other normal approximations are used. If the flat is prior, then we can replace the posterior mean by the MLE. Alternatively, we can replace the posterior mode by the posterior mean.

These approximations are called *model approximations* or *first order approximations* as the estimate θ by the mode and the error goes to 0 at a rate proportional to $1/n$. Estimates of posterior moments and quantiles can be obtained simply as the corresponding features of the approximated normal density. In the previous example, we can easily compare the exact and approximated quantiles with these few lines:

```
# data
k <- 20
n <- 100

# exact posterior with flat prior
alpha <- k+1
beta <- n-k+1
exact <- rbeta(1e5, shape1=alpha, shape2=beta)
quantile(exact)
```

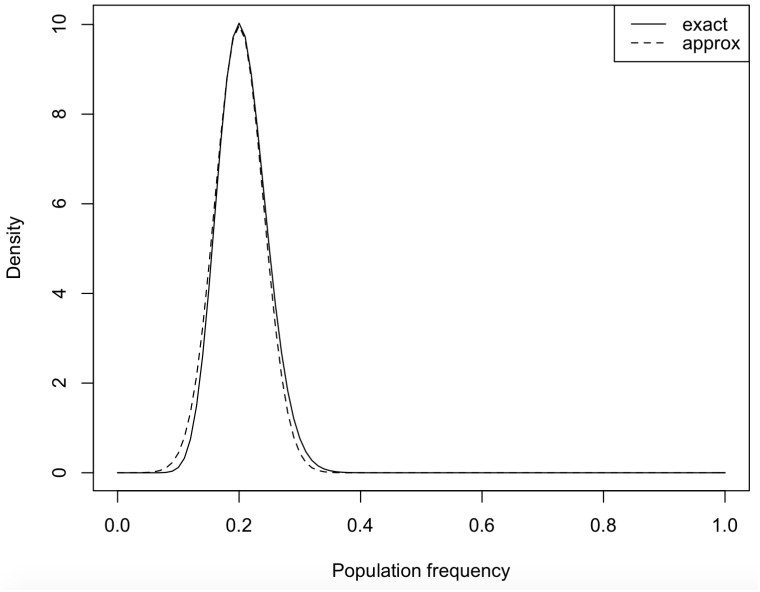

Figure 33: Normal approximation of a beta-binomial posterior distribution.

```
10
11 # normal approximation
12 thetaHat <- k/n
13 var <- thetaHat*(1-thetaHat)/n
14 approx <- rnorm(1e5, mean=thetaHat, sd=sqrt(var))
15 quantile(approx)
```

However, the estimates of moments and quantiles may be poor if the posterior differ from normality. The *Laplace's Method* provides a second order approximation to the posterior mean, with an error that decreases at a rate $1/n^2$.

The advantages of asymptotic methods are:

- 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 have also 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.

For these reasons, researchers now use iterative methods based on Monte Carlo sampling, which are longer to run but more general and fairly easy to implement.

3.2 Non-iterative Monte Carlo methods

Direct sampling of the posterior density can be done using a Monte Carlo integration. The basic definition is the following one. Suppose that $\vec{\theta} \sim h(\vec{\theta})$ with $h(\vec{\theta})$ being a posterior distribution. We aim at estimating γ the posterior mean of $c(\vec{\theta})$, where $\gamma \equiv E[c(\vec{\theta})] = \int c(\vec{\theta})h(\vec{\theta})d\vec{\theta}$. If $\vec{\theta}_1, \vec{\theta}_2, ..., \vec{\theta}_N$ are independent and identically distributed (iid) as $h(\vec{\theta})$, then:

$$\hat{\gamma} = \frac{1}{N}\sum_{i=1}^{N} c(\vec{\theta}_i) \tag{19}$$

which converges to $E[c(\vec{\theta})]$ with probability 1 as $N \to \infty$. The computation of posterior expectations requires only a sample of size N from the posterior distribution.

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). With higher dimensionality of $\vec{\theta}$, more samples are needed but the structure remains the same.

The variance of $\hat{\gamma}$ can be estimated from the sample variance of the $c(\vec{\theta}_i)$ values. The standard error estimate for $\hat{\gamma}$ is:

$$\hat{se}(\hat{\gamma}) = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}[c(\vec{\theta}_i) - \hat{\gamma}]^2} \tag{20}$$

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

In the univariate case, a histogram of the sampled θ_is estimates the posterior itself, as the probability of each bin converges to the true bin probability. Indeed, an estimate of $p \equiv P\{a < c(\theta) < b\}$ is given by

$$\hat{p} = \frac{\text{number of } c(\theta_i) \in (a, b)}{N} \tag{21}$$

One can even use a kernel density estimate to smooth the histogram, using a normal or rectangular distribution. Given a sample from the posterior distribution, almost any quantity can be estimated.

What happens if we can't directly sample from this distribution? There are methods for *indirect* sampling of the posterior distribution. The most commonly

used ones are: (i) importance sampling, (ii) rejection sampling, (iii) weighted bootstrap. In the importance sampling, an importance function is derived to approximate the normalised likelihood times prior. A weight function is then used for sampling. In the weighted bootstrap, instead of resampling from the set $\{\theta_1, ..., \theta_N\}$ with equal probabilities, some points are sampled more often than others because of unequal weighting.

In the rejection sampling, a smooth density called the envelope function to "cover" rather than approximate the posterior distribution. Suppose we can identify an *envelope function* $g(\vec{\theta})$ and a constant $M > 0$ such that $L(\vec{\theta})\pi(\vec{\theta}) < Mg(\vec{\theta})$ for all $\vec{\theta}$ Then the algorithm is the following (in the univariate case).

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.

Repeat this procedure until N samples are obtained. The members of this sample will be random variables from $h(\theta)$. The intuition for the rejection sampling algorithm is given in Figure 34. It is hard to sample from the true posterior but it is easier to sample from the envelope function.

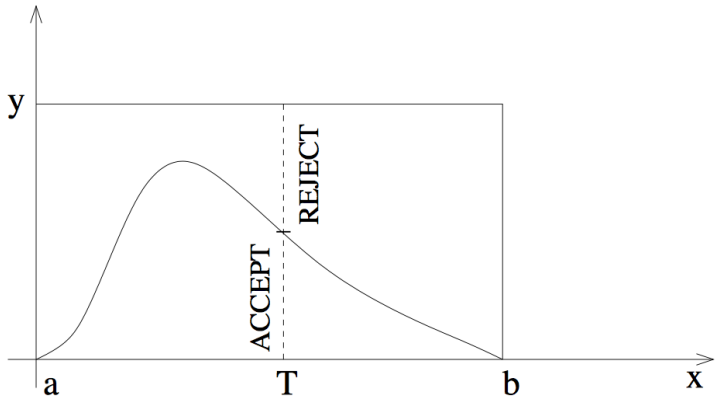

Figure 34: Rejection sampling algorithm.

Let's make a simple example in R when we approximate a Beta distribution using a uniform envelope function.

```
# calculate density from true posterior
calcTrueDensity <- function(x) dbeta(x, 3, 10)

# g is a uniform prior, M is the maximum density value of the
    posterior (if known)
x <- seq(0,1,0.01)
epsilon <- 1e-3
```

```
M <- max(calcTrueDensity(x)) + epsilon

thetas <- c()

# we want N samples
N <- 1e4
rawDensity <- rbeta(N, 3, 10)

while (length(thetas) < N) {

  theta_j <- runif(1, 0, 1)
  U <- runif(1, 0, 1)

    if (M*U < calcTrueDensity(theta_j)) thetas <- c(thetas,
        theta_j)

}

# check qq-plot
qqplot(rawDensity, thetas)
abline(0,1)
```

3.3 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, for instance, that is acceptable to approximate the (log) posterior.

In these cases it is now standard practice to use *Markov chain Monte Carlo* (MCMC) methods. The rationale is that these methods can sequentially sample parameter values from a Markov chain whose stationary distribution is the desired posterior distribution.

A Markov process is a mathematical object following a a stochastic (or random process), typically usually defined as a collection of random variables (Figure 35). A Markov process has the property that the next value of the process depends only on the current value, but it is independent of the previous values. In other words, the future value will depend only on the current state. 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.

The great increase of generality of MCMC methods comes at the price of requiring an assessment of *convergence* of the Markov chain to its stationary distribution. The stationary distribution is the probability distribution to which the process converges for large values of steps, or iterations. Convergence is usually assessed using plots or numerical summaries of the sampled distribution from the chain. The majority of Bayesian MCMC computation is based on two algorithms: the *Gibbs sampler* and the *Metropolis-Hastings (M-H)* algorithm.

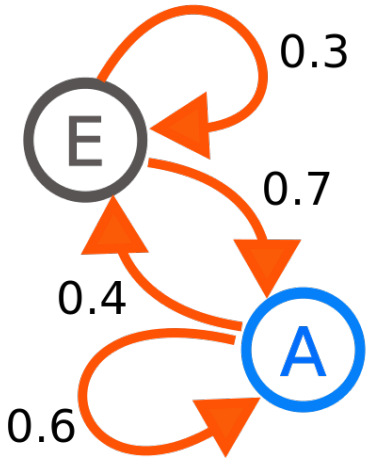

Figure 35: A diagram of a two-state Markov process, with the states labelled E and A. Each number represents the probability of the Markov process changing from one state to another state, with the direction indicated by the arrow. Source: Wikipedia.

3.3.1 Gibbs sampler

Suppose our model has k parameters so that $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_k)$. We assume that we can sample from the full conditional distributions. The collection of full conditional distributions uniquely determines the joint posterior distribution $p(\vec{\theta}, \vec{y})$ and therefore all marginal posterior distributions $p(\theta_i, \vec{y})$, for $i = 1, ..., k$.

Given an arbitrary set of starting $\{\theta_2^{(0)}, ..., \theta_2^{(k)}\}$, the algorithm is:

- for $(t = 1, ..., T)$, repeat:

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

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

  ...

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

Under most conditions, $(\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_k^{(t)})$ converges to a draw from the true joint posterior distribution $p(\theta_1, \theta_2, ..., \theta_k|\vec{y})$. This implies that for sufficiently large $t > t_0$ then $\{\vec{\theta^{(t)}}, t = t_0 + 1, ..., T\}$ is a correlated sample from the true posterior.

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 pos-

terior mean:

$$\hat{E}(\theta_i|\vec{y}) = \frac{1}{T-t_0}\sum_{t=t_0+1}^{T}\theta_i^{(t)} \tag{22}$$

The time $0 <= t <= t_0$ is called the $burn-inperiod$.

To speed up the computational time, we can run m chains (typically $m \leq 5$ in parallel. This procedure will also be useful for assessing samples convergence. In this case the posterior mean is estimated as:

$$\hat{E}(\theta_i|\vec{y}) = \frac{1}{m(T-t_0)}\sum_{j=1}^{m}\sum_{t=t_0+1}^{T}\theta_{i,j}^{(t)} \tag{23}$$

The entire marginal posterior density of θ_i is estimated as:

$$\hat{E}(\theta_i|\vec{y}) \approx \frac{1}{m(T-t_0)}\sum_{j=1}^{m}\sum_{t=t_0+1}^{T}p(\theta_i|\vec{\theta}_{l\neq i,j}^{(t)},\vec{y}) \tag{24}$$

A minimal requirement for Gibbs convergence is that the parameter space must be fully *connected*, without "holes". Imagine a joint posterior distribution for two univariate parameters θ and ν with two disconnected regions of support. The posterior is either defined for ($\theta > 0$ and $\nu > 0$ or for ($\theta < 0$ and $\nu < 0$. If we choose $\theta^{(0)} > 0$ then we will have $\mu^{(1)} > 0$ and subsequentially $\theta^{(1)} > 0$ and so on. The chain won't be able to "escape" the first quadrant. Luckily, the vast majority of commonly used statistical models in ecology and evolution have continuous features and fully connected parameter spaces.

An additional issue appears when θ and ν are highly correlated as this will lead to autocorrelation. The chain will therefore have a "slow mixing" and might get trapped in one part of the joint distribution. A solution is called as *thinning* and relies on retaining only every m^{th} iterate after convergence as obtained. If m is large enough, then the samples will be uncorrelated. This procedure simplifies the assessment of the quality of our MCMC estimates.

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. This conditionally conjugate prior specification allows the use of the Gibbs sampler.

3.3.2 Metropolis-Hastings algorithm

The Gibbs sampler requires that you can sample from each of the full conditional probabilities. When the priors and likelihoods are not conjugate pairs one of more of these conditional probabilities may not be available in closed form. However even under these conditions the full conditional probabilities will be available up to a proportionality constant.

The *Metropolis* algorithm is a rejection algorithm which requires only a function proportional to the distribution to be sampled, at the cost of a rejection step

from a candidate density function. Therefore in this algorithm we generate samples from a joint posterior distribution $h(\vec{\theta})$ such as $p(\vec{\theta}|\vec{y}) \propto h(\vec{\theta}) \equiv f(\vec{y}|\vec{\theta})\pi(\vec{\theta})$.

The algorithm begins by proposing a *candidate*, or *proposal*, symmetric density $q(\vec{\theta^*}|\vec{\theta}^{(t-1)})$ which satisfies $q(\vec{\theta^*}|\vec{\theta}^{(t-1)}) = q(\vec{\theta}^{(t-1)}|\vec{\theta^*})$. From a starting value $\vec{\theta^{(0)}}$ at iteration $t = 0$, for $(t = 1, ..., T)$ the algorithm repeats:

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

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

Metropolis algorithm is flexible in the selection of the candidate density q but may be less efficient than the Gibbs sampler if not properly tuned. The usual approach is to set the candidate density as:

$$q(\vec{\theta^*}|\vec{\theta}^{(t-1)}) = N(\vec{\theta^*}|\vec{\theta}^{(t-1)}, \tilde{\Sigma}) \tag{25}$$

This distribution is symmetric and is "self-correcting" as candidates are always centered around the current value of the chain. As such, this approach is also called *random walk Metropolis*.

The posterior variance is represented by $\tilde{\Sigma}$ which can be empirically estimated from a preliminary run. A skewed q density will increase the acceptance rate but also generate more autocorrelated samples and therefore it may explore only a small proportion of the parameter space. A rule of thumb is to choose $\tilde{\Sigma}$ so that around 50% of the candidates are accepted. However often the choice of $\tilde{\Sigma}$ is done *adaptively*. One can keep track of the proportion of accepted candidates and tune $\tilde{\Sigma}$ accordingly. This is usually done during the burn-in period and it is called *pilot adaptation*.

A generalisation of the Metropolis algorithm drops the requirement that the candidate density must be symmetric. For instance, for bounded parameter spaces (e.g. $\theta > 0$) a Gaussian density is not appropriate. In the *Metropolis-Hastings* algorithm when $q(\vec{\theta} * |\vec{\theta}^{(t-1)}) \neq q(\vec{\theta}^{(t-1)}|\vec{\theta}*)$ we replace the acceptance rate r by:

$$r = \frac{h(\vec{\theta^*})q(\vec{\theta}^{(t-1)}|\vec{\theta^*})}{h(\vec{\theta}^{(t-1)})q(\vec{\theta^*}|\vec{\theta}^{(t-1)})} \tag{26}$$

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

An alternative is to set $q(\vec{\theta^*}|\vec{\theta}^{(t-1)}) = q(\vec{\theta^*})$ where the proposal ignores the current value of the variable. This algorithm is called *Hastings independence chain*. In this case the acceptance rate becomes:

$$r = \frac{h(\vec{\theta^*})/q(\vec{\theta^*})}{h(\vec{\theta}^{(t-1)})/q(\vec{\theta}^{(t-1)})} \tag{27}$$

which is the weight function in the importance sampling.

There are many variants and types of MCMC algorithms. In the *Langevin-Hastings* algorithm we introduce a systematic drift in the candidate density. Another approach is to use auxiliary variables to expand the parameter space, as in the *slice sampler* algorithm. An appropriate enlargement of the parameter space can broaden the class of distributions we can sample and accelerate convergence. A nice feature of MCMC algorithms is that they can be combined in a single problem using *hybrid* forms, resulting in a mixture of algorithm. Probably the best practise is to use a mixture of MCMC algorithms Finally, *adaptive* algorithms attempt to use the early output from a chain to refine and improve the sampling as it progresses.

Convergence Convergence is an important issue for MCMC algorithms as their output is random and autocorrelated. When the output is safely thought to come from a true stationary distribution of the Markov chain for all $t > T$ then the MCMC algorithm has converged at time T. There are both theoretical basis and diagnostic tools to assess whether the chain has reach convergence.

A possible diagnostic strategy might be:

- run few parallel chains with starting points drawn from an overdispersed (wide) distribution with respect to the stationary distribution;
- visually inspect these chains on a common graph for each parameter (Figure 36);
- for each graph calculate the scale reduction factor (to check whether the variation within chains are approximately equal to the total variation, Figure 37);
- investigate crosscorrelations among parameters suspected of being confounded.

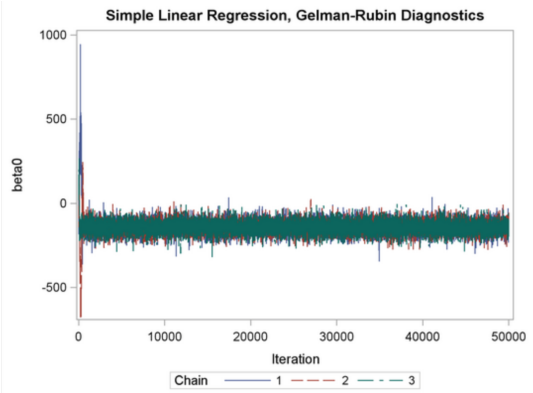

Figure 36: Three chains mixing for increasing t.

Obs	Parameter	Between-chain	Within-chain	Estimate	UpperBound
1	beta0	5384.76	1168.64	1.0002	1.0001
2	beta1	1.20	0.30	1.0002	1.0002
3	sigma2	8034.41	2890.00	1.0010	1.0011

Figure 37: Example of Gelman Rubin (1992) statistic as a diagnostic tool for convergence. Values refer to Figure 36.

The use of multiple methods is often helpful. Likewise considering multiple models is also useful for both understanding our data and check convergence.

Software There are several software that implement MCMC algorithms for generating samples from posterior distributions. A common program is called `OpenBUGS`, which is the free and open-source version of `WinBUGS`. The `BRugs` package in `R` calls `WinBUGS`. Likewise `JAGS` and its R interface `rjags` are valid alternatives.

4 Approximate Bayesian Computation

Bayesian computation involves modelling the joint density of parameters values θ and data x. The aim is then to compute the posterior probability

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \tag{28}$$

which can be difficult as the marginal likelihood

$$p(x) = \int p(x|\theta)\pi(\theta)d\theta \tag{29}$$

can involve a high dimensional integral. If the likelihood can be evaluated up to a normalising constant, Monte Carlo methods can be used to sample from the posterior.

As the models become more complicated, the likelihood function becomes difficult to define and compute. Under these circumstances it is easier to *simulate* data samples from the model given the value of a parameter. If the data are discrete and of low dimensionality, then it is possible to sample from the posterior density of the parameter without an explicit likelihood function and without approximation.

4.1 Rejection algorithm

Under the discrete case and low dimensionality, the algorithm to sample from the posterior from simulations is the following. 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 $p(\theta|x)$.

The posterior distribution gives the probability distribution of the parameter values that gave rise to the observations. To calculate summaries of this distribution it is possible to draw a histogram and derive notable percentiles.

Let's use again the example of elephant herds. Suppose we observe 4 herds arriving, so $y = 4$. Recalling how we modelled this process, the likelihood function was Poisson-distributed with a Gamma-shaped prior $G(3, 1)$. The posterior distribution was Gamma distributed with shape parameter $3 + 4 = 7$ and scale 0.5. Let's assume that we don't know the posterior distribution as we cannot derive it. However we assume that we know how to simulate y given a certain value of our parameter θ, the average arrival of herds per day. Let's write some code to sample from the posterior in a likelihood-free way.

```r
# Rejection algorithm
N <- 1e5
y <- 4
simulate <- function(param) rpois(n=1, lambda=param)

thetas <- c()
while (length(thetas) <= N) {

  # 1. draw from prior (discrete, bounded, uniform)
  theta <- sample(0:10, 1)

  # 2. simulate observations
  ysim <- simulate(theta)

  # 3. accept/reject
  if (ysim == y) thetas <- c(thetas, theta)

}
hist(thetas)
quantile(thetas, c(0.025,0.25,0.5,0.75,0.975))
```

Here we assume that we don't know the likelihood function but we can simulate data under this unknown function.

What happens in the continuous case? If the data are of low dimension we can modify the previous algorithm accordingly:

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.

Let's make a further example, recalling the case of water temperature at Bumpass Hell. θ is continuous and according to our previous example it has a prior distribution $U(80, 110)$. We assume we don't know the likelihood function (which is normally distributed) we can simulate observations that are distributed according to it. Finally, we assume we have an observation of the temperature $y = 91.3514$. What is $\rho(\cdot)$? For instance we can use the Euclidean distance:

$$\rho(x_i, y) = \sqrt{(x_i - y)^2} \tag{30}$$

Let's write some `R` code to estimate θ using this algorithm.

```r
# Rejection algorithm in the continuous case
N <- 1e4
y <- 91.3514
epsilon <- 1e-1

# again this is supposed to be an unknown function
```

```
simulate <- function(param) rnorm(n=1, mean=param, sd=sqrt(10))

# euclidean distance
rho <- function(x,y) sqrt((x-y)^2)

thetas <- c()
while (length(thetas) <= N) {

  # 1. draw from prior (continuous, bounded, uniform)
  theta <- runif(1, min=50, max=150)

  # 2. simulate observations
  ysim <- simulate(theta)

  # 3. accept/reject
  if (rho(ysim,y)<=epsilon) thetas <- c(thetas, theta)

  cat(theta, ysim, rho(ysim,y), (rho(ysim,y)<=epsilon),
      length(thetas), "\n")
  Sys.sleep(2)
}
hist(thetas)
quantile(thetas, c(0.025,0.25,0.5,0.75,0.975))
```

You can appreciate that the more the prior is different from the unknown likelihood function, the lower the acceptance rate.

An alternative to choose a value for ϵ is to rank all distances and select only a proportion of the lowest ones. In this case one sets the number of simulations to be performed (not the number of accepted simulations) and the proportions of simulations to retain. It is convenient to investigate the distribution of ranked distances to be sure to retain true outliers in the distribution.

As a quick exercise, and recalling the previous example, let's write some R code to estimate θ. We assume that:

1. our observations are $Y = \{91.34, 89.21, 88.98\}$
2. θ has prior $N(\mu = 90, \sigma^2 = 20)$ defined only in $80 \geq \theta \leq 110$
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. the total number of simulations is $10,000$ and we want to accept the lowest 5% of distance

We want to do the following 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% (using the library **coda** and function **HPDinterval(as.mcmc(x), prob=0.95)**)

A possible solution is the following.

```
# Rejection algorithm with proportions of simulations to accept
N <- 1e4
Y <- c(91.34, 89.21, 88.98)
th <- 0.05

simulate <- function(param) rnorm(n=1, mean=param, sd=sqrt(10))

# distance function
rho <- function(x,y) sum(sqrt((x-y)^2))/length(y)

thetas <- distances <- c()
for (i in 1:N) {

  # 1. draw from prior (continuous, bounded, uniform)
  theta <- 0
  while (theta<80 | theta>110) {
    theta <- rnorm(1, mean=90, sd=sqrt(20))
    }
  thetas <- c(thetas, theta)

  # 2. simulate observations
  ysim <- simulate(theta)

  # 3. calculate and retain distances
  distances <- c(distances, rho(ysim,Y))

}

# prior / parameter
hist(thetas)
quantile(thetas)

# relationship between parameter and distance
head(cbind(thetas,distances))

# distances
hist(distances)
accepted <- which(rank(distances,
    ties.method="random")/length(distances)<=th)
range(distances[accepted])
abline(v=max(distances[accepted]), lty=2)

# plot prior/posterior
par(mfrow=c(2,1))
hist(thetas, xlim=c(80,110), main="Prior")
hist(thetas[accepted], xlim=c(80,110), main="Posterior")
```

```r
quantile(thetas[accepted])

# 95% HPD
library(coda)
HPDinterval(as.mcmc(thetas[accepted]), prob=0.95)
```

What happens if we simulate more observations? The posterior will be skewer.

When the data become high dimensional (e.g. multivariate measurements) then it is necessary to reduce the dimensionality via the use of summary statistics. For instance, the complete genome of many samples has high dimensionality as it may have up to $N * L$ genotypes with N samples and L number of sites per-genome. One can calculate summary statistics $S(y)$ to describe some features of the data (e.g. indexes of genetic diversity in the case of multiple genomes).

In these case, the following is the prototype for the rejection-ABC algorithm, where ABC stands for Approximate Bayesian Computation. 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 $\rho(S(x_i), S(y)) > \epsilon$

The rejection-based ABC approach is depicted in Figure 38.

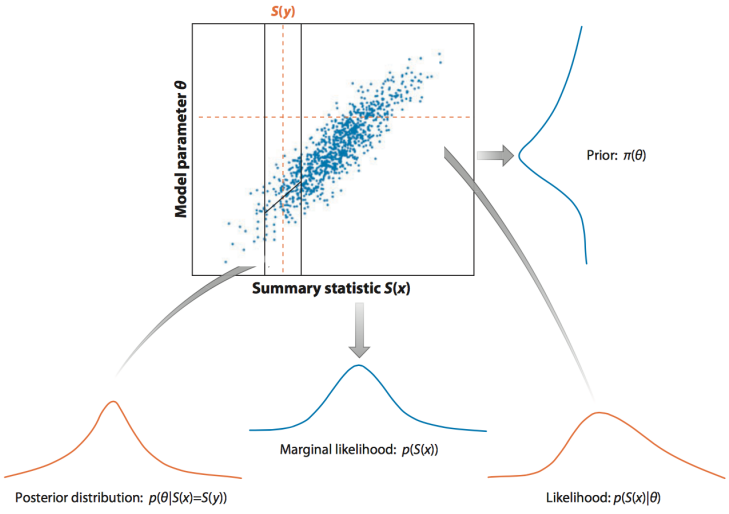

Figure 38: From Beaumont 2010 Annu Rev Ecol Evol Syst. Rejection- and regression-based approximate Bayesian computation (ABC).

The function $S(\cdot)$ can be a vector. The choice of summary statistics is considered as 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:

$$p(\theta|x) = p(\theta|S(x)) \tag{31}$$

The first example of an ABC approach was introduce by Pritchard et al. (1999). He summarised information for 445 Y-chromosome genes copies at eight microsatellites (and therefore 445 times 8 dimensions) into three numbers. The distance was chosen to be a normalised Chebyshev distance:

$$\max_j |\frac{S_j(x)}{S_j(y)} - 1| \tag{32}$$

for $j = 1, ..., s$ summary statistics.

You can clearly see on of the first issues, related to the curse of the dimensionality when using more than a few summary statistics. If summary statistics are uncorrelated, using the above distance, then we will reject many simulations with increasing number of summary statistics.

Solutions have been proposed in order to (i) use a wider acceptance tolerance and/or (ii) perform a better sampling from the prior.

4.2 Regression-based estimation

Another possibility to derive posterior using ABC is based on local linear regression, in order to obtain a potentially wider set of accepted points. This is the algorithm:

1. Given observation y repeat the following until M points have been generated:

   Draw $\theta_i \sim \pi(\theta)$

   Simulate $x_i \sim p(x|\theta_i)$

2. Calculate $S_j(x)$ for all j and k_j, the empirical standard deviation of $S_j(x)$

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 tolerance ϵ such that the proportion of accepted points $P_\epsilon = \frac{N}{M}$

5. Weight the simulated points $S(x_i)$ using $K_\epsilon(\rho(S(x_i), S(y)))$ where

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

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 approximation of the posterior distribution $p(\theta|y)$.

There are problems with regression-based methods too. When the observed summary statistics lies outside the unknown likelihood distribution (model misspecification), then regression is an extrapolation rather than an interpolation. In these cases posterior draws (after regression adjustments) can be outside the prior range. This problem occurs when the observations lie at the boundaries of the unknown likelihood (called prior-predictive distribution in the ABC context).

4.3 MCMC-ABC

Another possibility to increase the performance of ABC estimation is to do a better sampling. Indeed, the great majority of simulated parameter values may not give rise to summary statistics that are similar enough to the observed data. Efficiency will be slow as many points will be rejected or given negligible weight. We therefore want a procedure whereby parameters are sampled from a distribution that is closer to the posterior than from the prior. There are two main ways to do this, one via Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) sampling.

An MCMC-ABC algorithm is the following:
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 $\theta(t) = \theta^{(t-1)}$.

A good proposal distribution should resemble the actual posterior distribution of the parameters. A normal proposal distribution often works well in practice, centred in $\theta^{(t-1)}$. This is also called the *jumping* distribution. In this algorithm, at convergence the average distribution of proposed θ' is dominated by the posterior itself. It is also possible to apply any regression-adjustment methods on the MCMC sample to obtain more accurate estimates. Compared

to the classic MCMC with likelihoods, this algorithm has higher rejection rate. To circumvent this problem, the tolerance ϵ can be initially high and then reduced during the burn-in phase.

It was later proposed a method called Sequential Monte Carlo (SMC) for iteratively improving on an ABC approximation. This approach consisted of two main features: (i) weighted resampling from the set of points already drawn and (ii) successive reduction in the tolerance ϵ.

4.4 Model assessment in ABC

Classical Bayesian inference can be applied to models and parameters in the ABC framework.

4.4.1 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:

$$\frac{p(\mu_i|x)}{p(\mu_j|x)} \div \frac{p(\mu_i)}{p(\mu_j)} \tag{33}$$

Typically, Bayes factors can be computed only if the parameters within the models have priors that integrate to one. Therefore, Bayesian model choice can be strongly affected by the prior. Notably, Bayesian model choice automatically penalised models with many parameters. Therefore, one does not need to account for different number of parameters between models.

4.4.2 Hierarchical model

ABC can also be adopted in a hierarchical Bayesian model. A potential difficulty here is that summary statistics should capture information from each unit so that the hyperparameters can be well inferred. However if there are many summary statistics, then it is unlikely that simulated data will closely match the observations.

4.4.3 Choice of summary statistics

A very much arbitrary area of ABC modelling lies in the choice of summary statistics. In some fields, there is a history of relating summary statistics to model parameters. In general there is no need of a strong theory relating summary statistics to model parameters. One issue here is the effect of summary statistics on inferences and whether some choices may bias the outcome of model choice. This may happen if chosen summary statistics have little relation to parameters in other models. Typically, some summary statistics may cover some aspects of the model while other statistics may cover different aspects, making the choice of a finite set of informative units problematic (Figure 39).

Figure 39: Choosing summary statistics: the issue of pulling a short blanket.

The main idea is that as more summary statistics are used, then they should be jointly sufficient for the likelihood. Also, summary statistics may be correlated to each other and to the parameters. However the accuracy and stability of ABC decreases rapidly with increasing numbers of summary statistics.

How one can choose the optimal set of summary statistics? For instance, 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. Alternatively, different summary statistics can be weighted differently according to their correlation with some model parameters. The number of summary statistics can also be reduced via multivariate dimensional scaling (e.g. using partial least-squares or principal component analysis). Finally, summary statistics should be scaled in order to have equal mean and variance, if normally distributed, to avoid putting a different weight to sparser distributions.

4.4.4 Model validation

A very important component of Bayesian modelling is validation and testing. Validation is the assessment of goodness-of-fit of the model and comparing alternative models. In ABC it is essential to distinguish errors due to the approximation from errors caused by the choice of the model. If available, one can compare the results of a simulation with expectations based on the theory.

Often the marginal (or joint) distribution 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. This issue does not occur in likelihood-based approaches.

A similar test is to compare the observations with the posterior predictive distribution. This can be done by simulating data with parameters drawn randomly from the current posterior distribution.

4.5 Applications of ABC in ecology and evolution

The initial applications of ABC have been mainly in population genetics, using a rejection or regression algorithm. Later on, a number of other areas in ecology, epidemiology and systems biology have seen an increase in the use ABC. The more recent applications use MCMC or SMC algorithms.

In population genetics, the data consists of frequencies of alleles or haplotypes in one or more populations. The goal is usually to estimate the demographic history of populations in terms of changes of population sizes, divergence times, migration rates, and so on.

For instance, a number of studies on inferring human evolution have been using ABC methods. For instance, Patin et al. (2009) compared different demographic models that explain the genetic differentiation within different African populations (Figures 40,41).

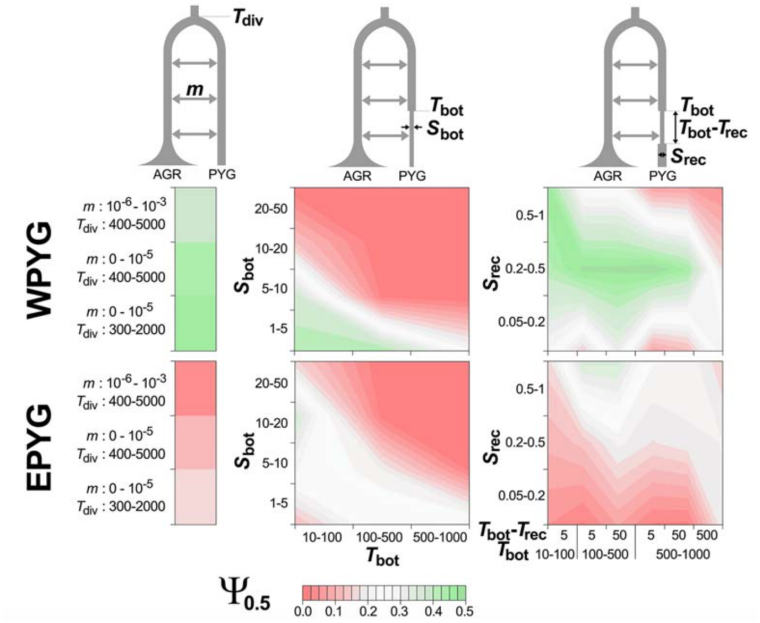

Figure 40: Patin et al. (2009). Different models simulating the demographic regime of the African groups and the mean proportion of small distances ($\Omega_{0.5}$) obtained in comparisons with simulated statistics.

Some features of ecology, epidemiology and systems biology appear to be very similar. Many aspects are captured by systems of partial or ordinary differential equations or stochastic differential equations. Data often consist of time series and/or spatial data. The goal here is to compare between hypothesised models that could explain the observed patterns and to infer parameters. Toni et al. (2009) provide an example using a Lotka-Volterra system on prey-

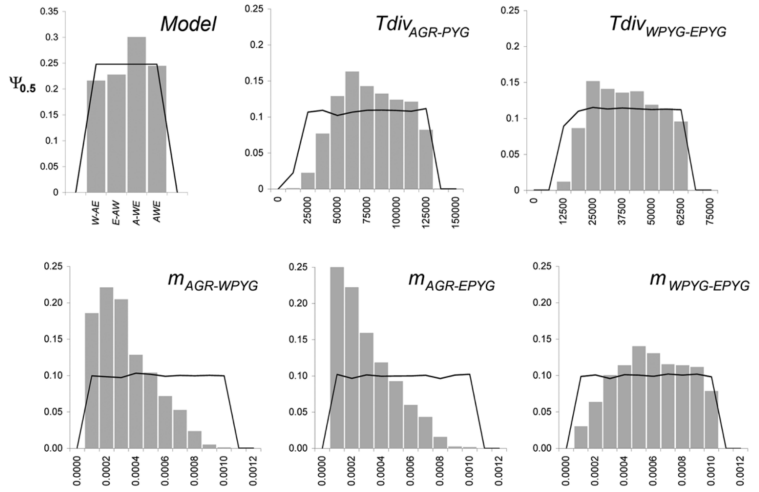

Figure 41: Patin et al. (2009). Prior and approximated posterior distributions of the model and parameters under the best-fit model.

predator dynamics from time series data on abundances (Figure 42). ABC has also been used for agent-based models, protein interaction networks, speciation rates under a neutral ecological model, extinction rates from phylogenetic data, epidemiology (e.g. transmission).

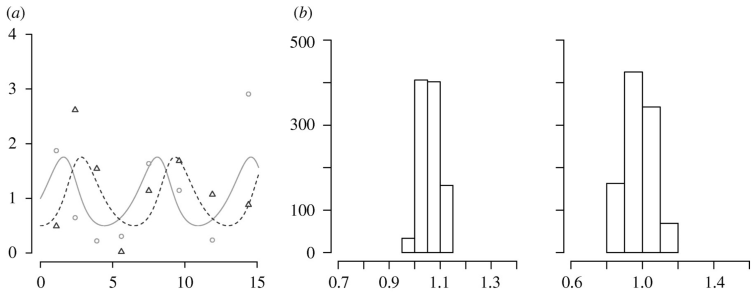

Figure 42: Toni et al. (2009). *(a) Trajectories of prey (solid curve) and predator (dashed curve) populations of the deterministic LV system and the data points (circles, prey data; triangles, predator data). (b) Parameters inferred by the ABC rejection sampler.*

In conclusion, 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 alternative. Many scenarios that evolutionary biologists or ecologists are interested to can be generated by simulations, making ABC very appealing. ABC can also be

useful for initial exploratory phase.

4.6 CASE STUDY / EXERCISE (3: estimating divergence time in bears

List of Figures

List of Tables

References