A SOFT INTRODUCTION TO BAYESIAN DATA ANALYSIS

Ehssan Ghashim[4], Patrick Boily[1,2,3]

Abstract

Bayesian analysis is sometimes maligned by data analysts, due in part to the perceived element of arbitrariness associated with the selection of a meaningful prior distribution for a specific problem and the (former) difficulties involved with producing posterior distributions for all but the simplest situations. On the other hand, we have heard it said that "while classical data analysts need a large bag of clever tricks to unleash on their data, Bayesians only ever really need one." With the advent of efficient numerical samplers, modern data analysts cannot shy away from adding the Bayesian arrow to their quiver. In this short report, we introduce the basic concepts underpinning Bayesian analysis, and we present a small number of examples that illustrate the strengths of the approach.

Keywords

Bayesian data analysis, Bayesian inference, MaxEnt priors, MCMC methods.

Funding Acknowledgement

Parts of this report were funded by a University of Ottawa grant to develop teaching material in French (2019-2020). These were subsequently translated into English before being incorporated into this document.

[1] Department of Mathematics and Statistics, University of Ottawa, Ottawa
[2] Data Action Lab, Ottawa
[3] Idlewyld Analytics and Consulting Services, Wakefield, Canada
Email: pboily@uottawa.ca

Contents

1. Introduction

Bayesian statistics is a system for describing epistemiological uncertainty using the mathematical language of probability; Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result with a probability distribution on the parameters of the model and on unobserved quantities (such as predictions).

1.1 Background

In 1763, Thomas Bayes published a paper on the problem of induction, that is, arguing from the specific to the general. In modern language and notation, Bayes wanted to use binomial data comprising r successes out of n attempts to learn about the underlying chance θ of each attempt succeeding. Bayes' key contribution was to use a probability distribution to represent uncertainty about θ. This distribution represents 'epistemiological' uncertainty, due to lack of knowledge about the world, rather than 'aleatory' probability arising from the essential unpredictability of future events, as may be familiar from games of chance.

In this framework, a probability represents a 'degree-of-belief' about a proposition; it is possible that the probability of an event will be recorded differently by two different observers, based on the respective background information to which they have access.

Modern Bayesian statistics is still based on formulating probability distributions to express uncertainty about unknown quantities. These can be underlying parameters of a system (induction) or future observations (prediction).

1.2 Bayes' Theorem

Bayes' Theorem provides an expression for the conditional probability of A given B, that is:

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

Bayes' Theorem can be thought of as way of coherently updating our uncertainty in the light of new evidence. The use of a probability distribution as a 'language' to express our uncertainty is not an arbitrary choice: it can in fact be determined from deeper principles of logical reasoning or rational behaviour.

Example 1. Consider a medical clinic.

- A could represent the event "Patient has liver disease." Past data suggests that 10% of patients entering the clinic have liver disease: $P(A)=0.10$.
- B could represent the litmus test "Patient is alcoholic." Perhaps 5% of the clinic's patients are alcoholics: $P(B)=0.05$.
- $B|A$ could represent the scenario that a patient is alcoholic, given that they have liver disease: perhaps we have $P(B|A)=0.07$, say.

According to Bayes' Theorem, then, the probability that a patient has liver disease assuming that they are alcoholic is

$$P(A|B)=\frac{0.07\times 0.10}{0.05}=0.14$$

While this is a (large) increase over the original 10% suggested by past data, it remains unlikely that any particular patient has liver disease.

Bayes' Theorem with Multiple Events

Let D represent some observed data and let A, B, and C be mutually exclusive (and exhaustive) events conditional on D. Note that

$$\begin{aligned}P(D)&=P(A\cap D)+P(B\cap D)+P(C\cap D)\\&=P(D|A)P(A)+P(D|B)P(B)+P(D|C)P(C).\end{aligned}$$

According to Bayes' theorem,

$$\begin{aligned}P(A|D)&=\frac{P(D|A)P(A)}{P(D)}\\&=\frac{P(D|A)P(A)}{P(D|A)P(A)+P(D|B)P(B)+P(D|C)P(C)}.\end{aligned}$$

In general, if there are n exhaustive and mutually exclusive outcomes $A_1,\ldots,A_n$, we have, for any $i\in\{1,\ldots,n\}$:

$$P(A_i|D)=\frac{P(A_i)P(D|A_i)}{\sum_{k=1}^{n}P(A_k)P(D|A_k)}$$

The denominator is simply $P(D)$, the **marginal distribution of the data**.

Note that, if the values of A_i are portions of the continuous real line, the sum may be replaced by an integral.

Example 2. In the 1996 General Social Survey, for males (age 30+):

- 11% of those in the lowest income quartile were college graduates.
- 19% of those in the second-lowest income quartile were college graduates.
- 31% of those in the third-lowest income quartile were college graduates.
- 53% of those in the highest income quartile were college graduates.

What is the probability that a college graduate falls in the lowest income quartile?

Let $Q_i, i=1,2,3,4$ represent the income quartiles (i.e. $P(Q_i)=0.25$) and D represent the event that a male over 30 is a college graduate. Then

$$\begin{aligned}P(Q_1|D)&=\frac{P(D|Q_1)P(Q_1)}{\sum_{k=1}^{4}P(Q_k)P(D|Q_k)}\\&=\frac{(0.11)(0.25)}{(0.11+0.19+0.31+0.53)(0.25)}=0.09.\end{aligned}$$

1.3 Bayesian Inference Basics

Bayesian statistical methods start with existing prior beliefs, and update these using data to provide posterior beliefs, which may be used as the basis for inferential decisions:

$$\underbrace{P(\theta|D)}_{\text{posterior}}=\underbrace{P(\theta)}_{\text{prior}}\times\underbrace{P(D|\theta)}_{\text{likelihood}}/\underbrace{P(D)}_{\text{evidence}},$$

where the evidence is

$$P(D)=\int P(D|\theta)P(\theta)d\theta.$$

In the vernacular of Bayesian data analysis (BDA),

- the **prior**, $P(\theta)$, represents the strength of the belief in θ without taking the observed data D into account;
- the **posterior**, $P(\theta|D)$, represents the strength of our belief in θ when the observed data D is taken into account;
- the **likelihood**, $P(D|\theta)$, is the probability that the observed data D would be generated by the model with parameter values θ, and
- the **evidence**, $P(D)$, is the probability of observing the data D according to the model, determined by summing (or integrating) across all possible parameter values and weighted by the strength of belief in those parameter values.

Example 3. *Application to neuroscience.* Cognitive neuroscientists investigate which areas of the brain are active during particular mental tasks. In many situations, researchers observe that a certain region of the brain is active and infer that a particular cognitive function is therefore being carried out; [41] cautioned that such inferences are not necessarily firm and need to be made with Bayes' rule in mind. The same paper reports the following frequency table of previous studies that involved any language-related task (specifically phonological and semantic processing) and whether or not a particular **region of interest** (ROI) in the brain was activated:

	Language (L)	**Other** ($\bar{L}$)
Activated (A)	166	199
Not Activated ($\bar{A}$)	703	2154

Suppose that a new study is conducted and finds that the ROI is activated (A). If the prior probability that the task involves language processing is $P(L) = 0.5$, what is the posterior probability, $P(L|A)$, given that the ROI is activated?

$$\begin{aligned} P(L|A) &= \frac{P(A|L)P(L)}{P(A|L)P(L) + P(A|\bar{L})P(\bar{L})} \\ &= \frac{(166/(166+703))0.5}{(166/(166+703))0.5 + (199/(199+2154))0.5} \\ &= 0.693 \end{aligned}$$

Notice that the posterior probability of involving language processes is slightly higher than the prior.

Exercises

Exercise 1. (1975 British national referendum on whether the UK should remain part of the European Economic Community). Suppose 52% of voters supported the Labour Party and 48% the Conservative Party. Suppose 55% of Labour voters wanted the UK to remain part of the EEC and 85% of Conservative voters wanted this. What is the probability that a person voting “Yes” to remaining in EEC is a Labour voter? [3]

Exercise 2. Given the following statistics, what is the probability that a woman has cancer if she has a positive mammogram result? [20]

- 1% of women over 50 have breast cancer.
- 90% of women who have breast cancer test positive on mammograms.
- 8% of women will have false positives.

2. Bayesian Methods Applied to Data Analysis

The essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis.

2.1 The 3 Steps of Bayesian Data Analysis

The process of Bayesian data analysis (BDA) can be idealized by dividing it into the following 3 steps:

1. Setting up a full probability model (the **prior**) – a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process (when available).
2. Conditioning on observed data (**new data**) – calculating and interpreting the appropriate posterior distribution (i.e. the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data).
3. Evaluating the fit of the model and the implications of the resulting posterior distribution (the **posterior**) – how well does the model fit the data? are the substantive conclusions reasonable? how sensitive are the results to the modeling assumptions made in step 1? Depending on the responses, one can alter or expand the model and repeat the 3 steps.

The essence of Bayesian methods consists in identifying the **prior beliefs** about what results are likely, and then updating those according to the **collected data**.

For example, if the current success rate of a gambling strategy is 5%, we may say that it's reasonably likely that a small strategy modification could further improve that rate by 5 percentage points, but that it is most likely that the change will have little effect, and that it is entirely unlikely that the success rate would shoot up to 30% (after all, it is only a small modification).

As the data start coming in, we start updating our beliefs. If the incoming data points to an improvement in the success rate, we start moving our prior estimate of the effect upwards; the more data we collect, the more confident we are in the estimate of the effect and the further we can leave the prior behind.

The end result is called the **posterior** – a probability distribution describing the likely effect of the strategy.

3. Prior Distributions

Specifying a model means, by necessity, providing a prior distribution for the unknown parameters. The prior plays a critical role in Bayesian inference through the updating statement :

$$P(\theta|D) \propto P(\theta) \times P(D|\theta).$$

In the Bayesian approach, all unknown quantities are described probabilistically, even before the data has been observed. All priors are subjective in the sense that the decision to use any prior is left completely up to the researcher. But the choice of priors **is no more subjective**

than the choice of likelihood, the selection or collection of a sample, the estimation, or the statistic used for data reduction. The choice of a prior can substantially affect posterior conclusions, however, especially when the sample size is not large.

We now examine several broad methods of determining prior distributions.

3.1 Conjugate Priors

The main challenge of Bayesian methods is that the posterior distribution of the vector θ might not have an analytical form. Specifically, producing marginal posterior distributions from high-dimensional posteriors by repeated analytical integration may be difficult or even impossible mathematically. There are exceptions however, providing easily obtainable computational posteriors through the use of a **conjugate prior**. Conjugacy is a joint property of a prior and a likelihood that implies that the posterior distribution has the same distributional form as the prior, but with different parameter(s).

The table below represents some common likelihoods and their conjugate priors (an extensive list can be found in [27]).

Likelihood	Prior	Hyperparameters
Bernouilli	Beta	$\alpha > 0, \beta > 0$
Binomial	Beta	$\alpha > 0, \beta > 0$
Poisson	Gamma	$\alpha > 0,\ \beta > 0$
Normal for μ	Normal	$\mu \in \mathbb{R},\ \sigma^2 > 0$
Normal for σ^2	Inverse Gamma	$\alpha > 0,\ \beta > 0$
Exponential	Gamma	$\alpha > 0,\ \beta > 0$

For instance, if the probability of s successes in n trials (the likelihood) is given by

$$P(s,n|q) = \frac{n!}{s!(n-s)!} q^s (1-q)^{n-s}, \quad q \in [0,1],$$

and the prior probability for q follows a Beta(α,β) distribution with $\alpha > 0, \beta > 0$ (so that

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

for $q \in [0,1]$), then the posterior distribution for q given s successes in n trials follows a Beta$(\alpha+s, \beta+n-s)$ distribution (so that

$$P(q|s,n) = \frac{P(s,n|q) \times P(q)}{P(s,n)} = \frac{q^{\alpha+s-1}(1-q)^{\beta+n-s-1}}{B(\alpha+s,\beta+n-s)}$$

for $q \in [0,1]$).

Conjugate priors are mathematically convenient, and they can be quite flexible, depending on the specific hyperparameters we use; but **they reflect very specific prior knowledge and should be eschewed unless we truly possess that prior knowledge**.

3.2 Uninformative Prior Distribution

An uninformative prior is one in which little new explanatory power about the unknown parameter is provided by intention. Uninformative priors are very useful from the perspective of traditional Bayesianism seeking to mitigate the frequentist criticism of **intentional subjectivity**. These priors intentionally provide very little specific information about the parameter(s).

A classic uninformative prior is the **uniform prior**. A proper uniform prior integrates to a finite quantity and is thus normalizable. By example, for data following a Bernoulli(θ) distribution, a uniform prior on θ is

$$P(\theta) = 1, \quad 0 \le \theta \le 1.$$

This approach makes sense when θ has bounded support. But for data following a $N(\mu,1)$ distribution, the uniform prior on the support of μ is improper as

$$P(\mu) = 1, -\infty < \mu < \infty$$

diverges; however, such a choice could still be acceptable as long as the resulting posterior is normalizable (i.e. the integral of the posterior converges on its support). As there are instances where an improper prior yields an improper posterior, care is warranted. The rationale for using uninformative prior distributions is often said to be 'to let the data speak for itself,' so that inferences are unaffected by information external to the current data.

3.3 Informative Prior Distributions

Informative priors are those that **deliberately** insert information that researchers have at hand. This seems like a reasonable approach since previous scientific knowledge should play a role in doing statistical inference. However, there are two important requirements for researchers:

1. overt declaration of prior specification, and
2. detailed sensitivity analysis to show the effect of these priors relative to uninformed types.

Transparency is required to avoid the common pitfall of **data fishing**; sensitivity analysis can provide a sense of exactly how informative the prior is. But where do informative priors come from, in the first place? Generally these priors are derived from:

- past studies, published work, researcher intuition;
- interviewing domain experts;
- convenience with conjugacy, and
- non-parametric and other data-derived sources.

Prior information from past studies need not be in agreement. One useful strategy is to construct prior specifications from **competing school-of-thoughts** in order to contrast the resulting posteriors and produce informed statements about the relative strength of each of them.

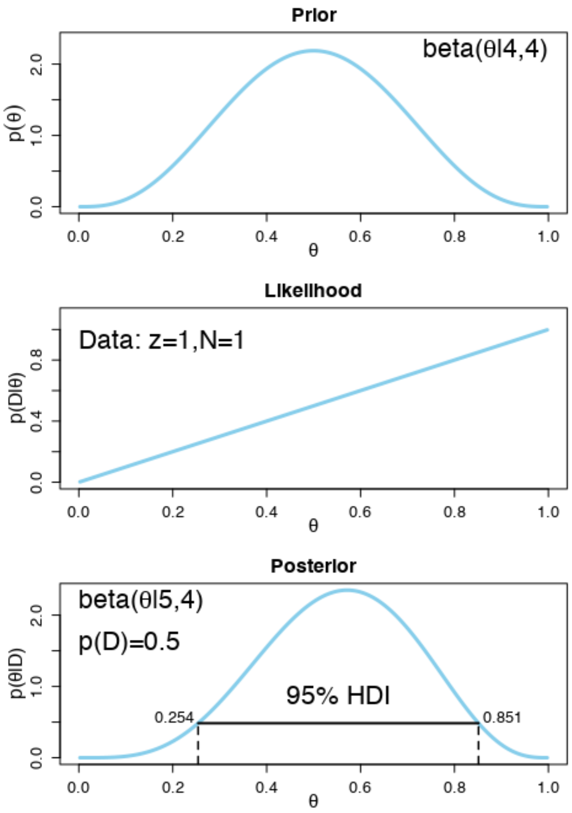
Example 4. *Influence of the prior.* We have noted previously that a Bernouilli likelihood and a Beta prior form a set of conjugate priors. For this exercise, we use the R function `BernBeta())` defined in [10] (notice that the function returns the posterior beta values each time it is called, so returned values can be fed back into the prior during the next function call).

(a) Start with a prior distribution that expresses some uncertainty that a coin is fair: Beta($\theta|4,4$). Flip the coin once; assume that a Head is obtained. What is the posterior distribution of the uncertainty in the coin's fairness θ?

Solution: at the R command prompt, type:

```
> post = BernBeta( c(4,4) , c(1) )
```

This function uses the conjugacy relation from Section 3.1 to determine the posterior distribution Beta for the uncertainty in the fairness of the coin given the parameters of the Beta prior and the observed data assuming a Bernouilli likelihood (1 represents a H(ead) on the flip, 0 a T(ail)). However, we know on theoretical grounds that the posterior follows a Beta($\theta|4+1,4+1-1$) = Beta($\theta|5,4$) distribution:



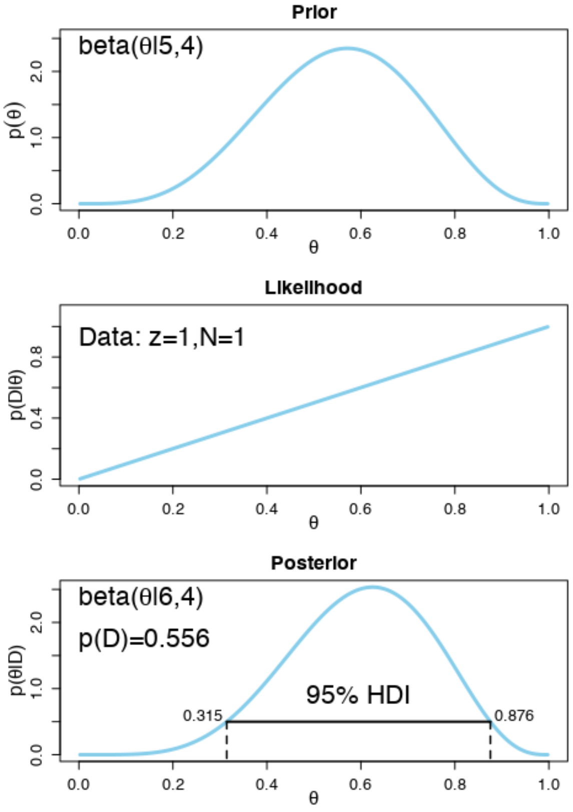
The label on the $y-$axis of the posterior distribution provides the posterior parameters (they are also given by typing `show(post)` at the command prompt).

(b) Use the posterior parameters from the previous flip as the prior for the next flip. Suppose we flip again and get a H. What is the new posterior on the uncertainty in the coin's fairness?

Solution: at the R command prompt, type

```
> post = BernBeta( post , c(1) )
```

The posterior distribution is Beta($\theta|6,4$), which is shown below.



(c) Using the most recent posterior as the prior for the next flip, flip a third time and obtain yet again a H. What is the new posterior?

Solution: in this case, we know that the posterior for the coin's fairness follows a Beta($\theta|7,4$) distribution (we won't provide the code or the output, this time!). Does 3 H in a row give you pause? Is there enough evidence to suggest that $\theta \neq 0.5$ (i.e that the coin is not fair)? What if you flipped 18 H in a row from this point on?

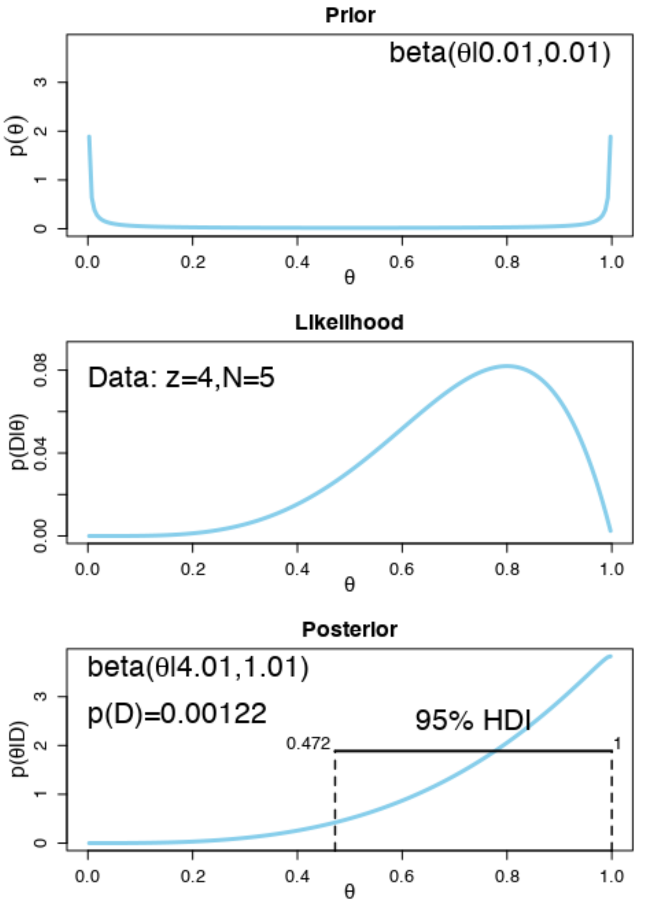
When working on a problem, it can be easy to get sidetracked and confused with the notation. In those cases, it is useful to return to the definition of each of the terms in Bayes' theorem (i.e. $P(\theta|D)$, $P(D)$, $P(D|\theta)$, etc.).

Example 5. *An unusual prior.* Suppose that a friend has a coin that we know comes from a magic store; as a result, we believe that the coin is strongly biased in either of the two directions (it could be a trick coin with both sides being H, for instance), but we don't know which one it favours. We will express the belief of this prior as a Beta distribution. Let's say that our friend flips the coin five times; resulting in 4 H and 1 T. What is the posterior distribution of the coin's fairness θ?

Solution: at the prompt, type

```
> post = BernBeta(c(1,1)/100,c(1,1,1,1,0))
```

yielding the posterior below.



The code above uses a prior given by Beta$(\theta|0.01, 0.01)$. This prior captures our belief that the coin is strongly biased (although we do not know in which direction the bias lies before seeing data). The choice of 0.01 is arbitrary, in a sense; 0.1 would have worked as well, for instance.

The posterior distribution is Beta$(\theta|4.01, 1.01)$ which, as shown above, has its mode essentially at 1.0, and not near the mean ≈ 0.8. Is the coin indeed biased? In which direction? How would your answer change if you had no reason to suspect that the coin was biased in the first place?

3.4 Maximum Entropy Priors

Whether the priors are uninformative or informative, we search for the distribution that best encodes the prior state of knowledge from a set of trial distributions.

Consider a discrete space X of cardinality M with probability density $P(X) = (p_1, ..., p_M)$. The **entropy** of a p, denoted by $H(p)$, is given by

$$H(p) = -\sum_{i=1}^{M} p_i \log p_i,$$ [1] $$\text{with } 0 \cdot \log(0) = 0.$$

The **maximum entropy principle** (MaxEnt) states that, given a class of trial distributions with constraints, the optimal prior is the trial distribution with the largest entropy. As an example, the most basic constraint is for p to lie in the probability simplex, that is, $\sum_i p_i = 1$ and $p_i \geq 0$ for all i in the discrete case, or $\int_\Omega P(Z)dZ = 1$ and $P(Z) \geq 0$ on Ω in the continuous case.

Example 6. Without constraints, the MaxEnt principle yields a prior which solves the optimization problem:

$$\begin{array}{rl} \max & -p_1 \log p_1 - \cdots - p_M \log p_M \\ \text{s.t.} & p_1 + \cdots + p_M = 1 \text{ and } p_1, \ldots, p_M \geq 0 \end{array}$$

Using them method of Lagrange multipliers, this optimization reduces to

$$p^* = \text{argmax}_p \{H(p) - \lambda(p_1 + \cdots + p_M - 1)\},$$

whose solution is $p^* \propto$ constant. Hence, subject to no additional constraints, the uniform distribution is the maximum entropy prior.

Example 7. *Using MaxEnt to build a prior for Bayesian inference.* "The joke about New York is that you can never get a cab, except when you don't need a cab, and then there are cabs everywhere" (quote and example from S.DeDeo's Maximum Entropy Methods tutorial [29]). How could we use Bayesian analysis to predict the cab waiting time? At various moments, head out to the street, say "I need a cab!" and keep track of how long you took before a cab was available. Perhaps the observations (in minutes) look like this

$$6, 3, 4, 6, 2, 3, 2, 6, 4, 4.$$

What can you conclude about the waiting time for a New York City cab? In the best case scenario a cab is waiting for us as we get to the curb ($j = 0$), while in the worst case scenario (a zombie apocalypse, say?), no cab ever comes ($j \to \infty$). But can anything else be said?

To use MaxEnt in this situation, we need to find – among all of the trial distributions that could have generated the

[1] In the case of a continuous pdf $P(X_1, \ldots, X_n)$ on some domain $\Omega \subseteq \mathbb{R}^n$, the entropy is given by $H(p) = -\int_\Omega P(Z)\log(P(Z))dZ$.

observed waiting times – the one with the highest entropy. Unfortunately, there are infinitely many such distributions. We can narrow the search by including a constraint stating that the expected value of the trial distributions should be the same as the mean of the sample, namely 4.

The two constraints translate to

$$g_1(p)=\sum_{j=0}^{\infty} j\cdot p_j-4=0 \quad \text{and} \quad g_2(p)=\sum_{j=0}^{\infty} p_j-1=0,$$

where p_j is the probability of having to wait j minutes for a cab.

The method of Lagrange multipliers reduces the problem to solving

$$\text{argmax}_p \{\{H(p)-\lambda_1 g_1(p)-\lambda_2 g_2(p)\}.$$

This requires solving the gradient equation

$$\nabla_p H(p)=\lambda_1\nabla_p g_1(p)+\lambda_2\nabla_p g_2(p),$$

which gives rise to equations of the form

$$-(\ln p_j+1)=\lambda_1 j+\lambda_2, \quad j=0,1,\ldots,$$

or simply $p_j = \exp(-\lambda_1 j)\exp(-1-\lambda_2)$ for $j = 0,1,\ldots$ Since

$$1=\sum_{j=0}^{\infty} p_j=\exp(-1-\lambda_2)\sum_{j=0}^{\infty}\exp(-\lambda_1 j),$$

so that

$$\exp(1+\lambda_2)=\sum_{j=0}^{\infty}\exp(-\lambda_1 j)=\frac{1}{1-\exp(-\lambda_1)}, \tag{1}$$

assuming that $|\exp(-\lambda_1)|<1$. Similarly,

$$4=\sum_{j=0}^{\infty} jp_j=\exp(-1-\lambda_2)\sum_{j=0}^{\infty} j\exp(-\lambda_1 j),$$

so that

$$4\exp(1+\lambda_2)=\sum_{j=0}^{\infty} j\exp(-\lambda_1 j)=\frac{\exp(-\lambda_1)}{(1-\exp(-\lambda_1))^2}. \tag{2}$$

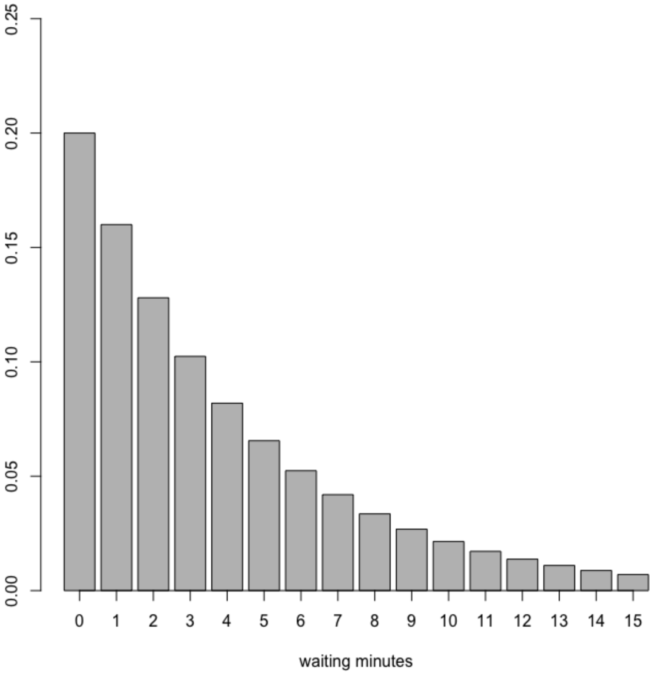
Substituting (1) into (2) and solving for λ_1, we see that $\lambda_1=\ln(5/4)$. Substituting that result back into (1), we obtain $\exp(-1-\lambda_2)=\frac{1}{5}$, so that

$$p_j=\exp(-1-\lambda_2)\exp(-\lambda_1 j)=\frac{1}{5}\left(\frac{4}{5}\right)^j, j=0,\ldots$$

It is easy to see that this defines a distribution; a "verification" is provided by the following code.

```
pmf_maxent <- function(x,lambda=4/5)
    (1-lambda)*(\lambda)^x
sum(pmf_maxent(0:100)) # check if it's a
    distribution
mp <- barplot(pmf_maxent(0:15),
    ylim=c(0,.25), xlab="waiting
    minutes")
axis(1,at=mp,labels=paste(0:15))
```

This distribution (see below) could be used as a prior in a Bayesian analysis of the situation. Notice that some information about the data (in this case, only the sample mean) is used to define the MaxEnt prior.



Exercises

Exercise 3. In this exercise you will study the possible effect that the choice of prior has on conclusions.

(a) Suppose you have in your possession a coin that you know was minted by the federal government and for which you have no reason to suspect tampering of any kind. Your prior belief about fairness of the coin is thus strong. You flip the coin 10 times and record 9 H(eads). What is your predicted probability of obtaining 1H on the 11th flip? Explain your answer carefully; justify your choice of prior. How would your answer change (if at all) if you use a frequentist viewpoint?

(b) A mysterious stranger hands you a different coin, this one made of some strange-to-the-touch material, on which the words "Global Tricksters Association" You flip the coin 10 times and once again record 9H. What is your predicted probability of obtaining 1H on the 11th flip? Explain your answer carefully; justify your choice of prior. Hint: Use the prior from Example 5.

4. Posterior Distributions

The posterior distribution is used to estimate a variety of model parameters of interest, such as the mean, the median, the mode, and so forth.

It is possible to construct **credible intervals/regions** directly from the posterior (in contrast to the "confidence" intervals of frequentist inference).

Given a posterior distribution on a parameter θ, a $1-\alpha$ credible interval $[L,U]$ is an interval such that

$$P(L \leq \theta \leq U|D) \geq 1-\alpha.$$

Because the posterior is a full distribution on the parameters, it is possible to make all sorts of probabilistic statements about their values, such as:

- "I am 95% sure that the true parameter value is bigger than 0.5."
- There is a 50% chance that θ_1 is larger than θ_2 .
- etc.

The best approach is to build the credible interval of θ-values using the **highest density interval** (HDI), i.e. to define a region C_k in the parameter space with

$$C_k = \{\theta : P(\theta|D) \geq k\},$$

where k is the largest number such that

$$\int_{C_k} P(\theta|D)\, d\theta = 1-\alpha.$$

This typically has the effect of finding the smallest (in measure) region C_k meeting the criterion.

The value k can be thought of the height of a horizontal line (or hyperplane, in the case of multivariate posteriors) overlaid on the posterior and whose intersection(s) with the latter define a region over which the integral of the posterior is $1-\alpha$. In most cases, the value k can be found numerically.

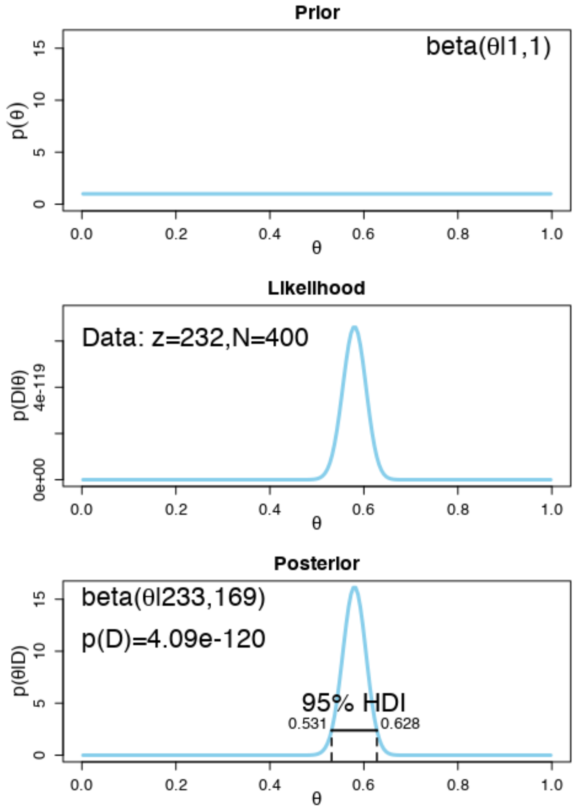
Example 8. *HDIs, elections, and iterative data collection.* It is an election year and you are interested in knowing whether the general population prefers candidate A or candidate B. A recently published poll states that of 400 randomly sampled people, 232 preferred candidate A, while the remainder preferred candidate B.

(a) Suppose that before the poll was published, your prior belief was that the overall preference follows a uniform distribution. What is the 95% HDI on your belief after learning of the poll result?

Solution: let preference for candidate A be denoted by 1, and preference for candidate B by 0. We can use the R function `BernBeta()` as in Example 4. At the prompt, type

```
> post=BernBeta(c(1,1),c(rep(1,232),rep(0,168)))
```

yielding a posterior with a 95% HDI from 0.531 to 0.628 for probability of candidate A.



(b) Based on the poll, is it credible to believe that the population is equally divided in its preferences among candidates?

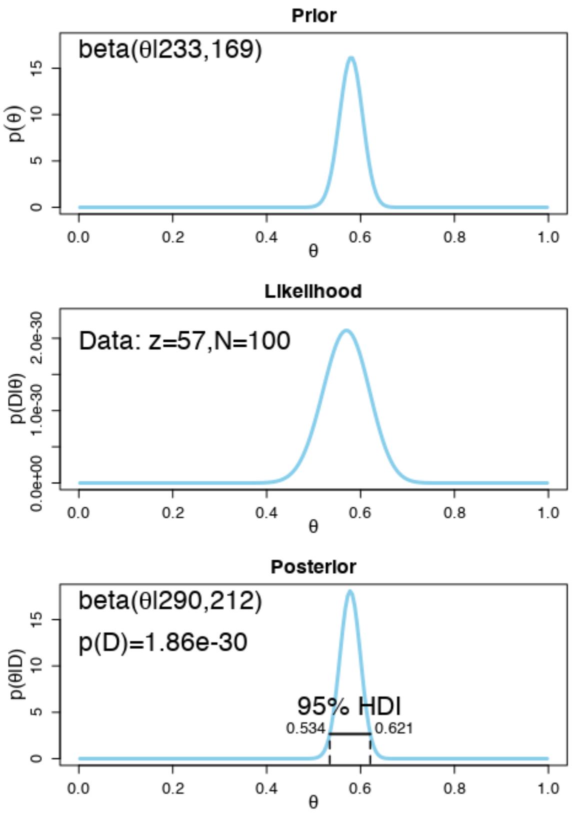
Solution: the HDI from Part (a) shows that $\theta = 0.5$ is not among the credible values, hence it is not credible to believe that the population is equally divided in its preferences (at the 95%) level.

(c) You want to conduct a follow-up poll to narrow down your estimate of the population's preference. In the follow-up poll, you randomly sample 100 people and find that 57 prefer candidate A. Assuming that the opinion of people have have not changed between polls, what is the 95% HDI on the posterior?

Solution: at the prompt, type

```
> post=BernBeta(post,c(rep(1,57),rep(0,43)))
```

yields the figure on the next page. The 95% HDI is a bit narrower for preference for candidate A is a bit nanarrower, from 0.534 to 0.621.



(d) Based on the follow-up poll, is it credible to believe that the population is equally divided in its preferences among candidates?

Solution: the HDI from (c) excludes $\theta = 0.5$; both the follow-up poll and the original poll suggest that the population is not equally divided (and actually prefers candidate A).

4.1 Markov Chain Monte Carlo (MCMC) Methods

The true power of Bayesian inference is most keenly felt when the model specifications lead to a posteriors that cannot be manipulated analytically; in that case, it is usually possible to recreate a synthetic (or **simulated**) set of values that share the properties with a given posterior. Such processes are known as **Monte Carlo simulations**.

A **Markov chain** is an ordered, indexed set of random variables (a stochastic process) in which the values of the quantities at a given state depends probabilistically only on the values of the quantities at the preceding state. **Markov chain Monte Carlo** (MCMC) methods are a class of algorithms for sampling from a probability distribution based on the construction of a Markov chain with the desired distribution as its equilibrium distribution. The state of the chain after a number of steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps.

MCMC techniques are often applied to solve integration and optimization problems in large-dimensional spaces. These two types of problem play a fundamental role in machine learning, physics, statistics, econometrics and decision analysis. For instance, given variables $\theta \in \Theta$ and data D, the following (typically intractable) integration problems are central to Bayesian inference:

- **normalisation** – in order to obtain the posterior $P(\theta|D)$ given the prior $P(\theta)$ and likelihood $P(D|\theta)$, the normalizing (denominator) factor in Bayes' theorem needs to be computed

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(D|\theta)P(\theta)d\theta}.$$

- **marginalisation** – given the joint posterior of (θ, x), we may often be interested in the marginal posterior

$$P(\theta|D) = \int P(\theta, x|D)dx.$$

- **expectation** – the final objective of the analysis is often to obtain summary statistics of the form

$$E(f(\theta)) = \int_\Theta f(\theta)P(\theta|D)d\theta$$

for some function of interest (i.e. $f(\theta) = \theta$ (mean), or $f(\theta) = (\theta - E(\theta))^2$ (variance)).

The Metropolis-Hastings (MH) Algorithm

The Metropolis-Hastings (MH) algoirthm is a specific type of Monte Carlo process; it is likely among the ten algorithms that have had the greatest influence on the development and practice of science and engineering in recent times.

MH generates a random walk (that is, it generates a succession of posterior samples) in such a way that each step in the walk is **completely independent** of the preceding steps; the decision to reject or accept the proposed step is also independent of the walk's history.

Any process for which the current step is independent (forgetful) of the previous states, namely

$$P(X_{n+1} = x|X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = x|X_n = x_n)$$

for all n, X_j and x_j, $j = 1, \ldots, n$, is called a **(first order) Markov process**, and a succession of such steps is a **(first order) Markov chain**.

MH uses a candidate or proposal distribution for the posterior, say q$(\cdot, \theta)$, where θ is a vector of parameters that is fixed by the user-called tuning parameters; MH then constructs a Markov Chain by proposing a value for θ from this candidate distribution, and then either accepting or rejecting this value (with a certain probability).

Theoretically the proposal distributions can be nearly any distribution, but in practice it is recommended that (really) simple ones be selected: a normal if the parameter of interest can be any real number (e.g. μ), or a log-normal if it has positive support (e.g. σ^2), say.

The **Metropolis-Hastings** (MH) algorithm simulates samples from a probability distribution by making use of the full joint density function and (independent) proposal distributions for each of the variables of interest.

Algorithm 1: Metropolis-Hastings Algorithm

```
Initialize  $x^{(0)} \sim q(x)$
for $i = 1, 2, \cdots$ do
    Propose:  $x^* \sim q(x^{(i)}|x^{(i-1)})$
    Acceptance Probability:
        $\alpha(x^*|x^{(i-1)}) = min\left\{1, \frac{q(x^{(i-1)}|x^*)\pi(x^*)}{q(x^*|x^{(i-1)})\pi(x^{(i-1)})}\right\}$
    $u \sim U(0,1)$
    if $u < \alpha$ then
        Accept the proposal:  $x^{(i)} \leftarrow x^*$
    else
        Reject the proposal:  $x^{(i)} \leftarrow x^{(i-1)}$
    end
end
```

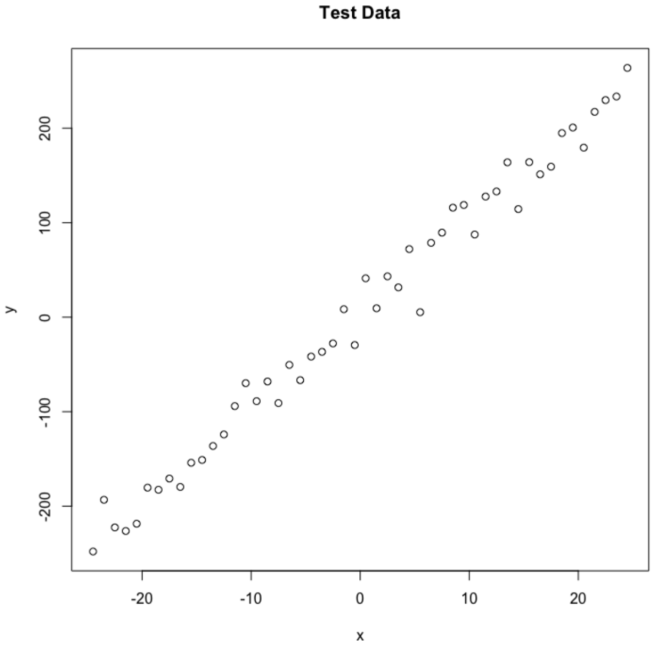
The first step is to **initialize the sample value** for each random variable (often obtained by sampling from the variable's prior distribution). The main loop of Algorithm 1 consists of three components:

- **generate a candidate sample** x^* from the proposal distribution $q(x^{(i)}|x^{(i-1)})$;
- **compute the acceptance probability** via the acceptance function $\alpha(x^*|x^{(i-1)})$ based on the proposal distribution and the full joint density $\pi(\cdot)$;
- **accept the candidate sample** with probability α, the acceptance probability, or **reject it** otherwise.

Example 9. *The MH algorithm and simple linear regression.* The **test data** for this example is genreated using the following code.

```
t.A <- 10 # true slope
t.B <- 0 # true intercept
t.sd <- 20 # true noise
s.Size <- 50 # sample size
# create independent x-values
x <- (-(s.Size-1)/2):((s.Size-1)/2)
# create dependent values according to
    ax + b + N(0,sd)
y <- t.A * x + t.B +
    rnorm(n=s.Size,mean=0,sd=t.sd)
plot(x,y, main="Test Data")
```

Notice that the x values are balanced around zero to "de-correlate" slope and intercept. The result should look like the chart below.



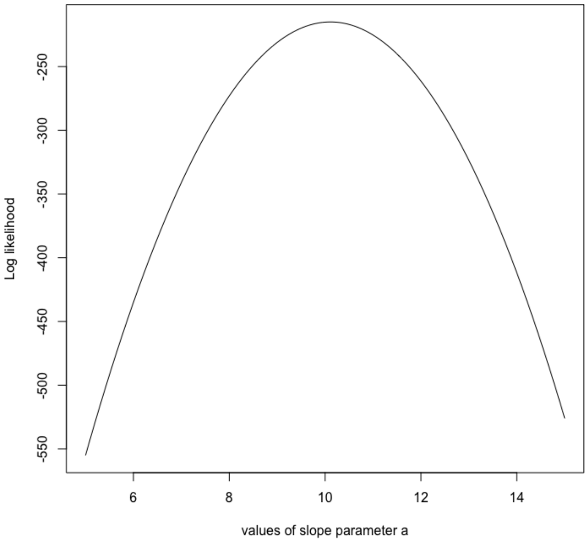
Defining the statistical model. The next step is to specify the statistical model. We already know that the data was created with a linear relationship $y = ax + b$ together with a normal error model $N(0, sd)$ with standard deviation sd, so we might as well use the same model for the fit and see if we can retrieve our original parameter values. Note however that, in general, the generating model is unknown.

Deriving the likelihood function from the model. A linear model of the form $y = ax + b + N(0, sd)$ takes the parameters (a, b, sd) as inputs. The output should be the probability of obtaining the test data under this model: in this case, we only need to calculate the difference between the predictions $y = ax + b$ and the observed y, and then look up the probability (using `dnorm`) for such deviations to occur.

```
likehd <- function(param){
   a = param[1]
   b = param[2]
   sd = param[3]
   pred = a*x + b
   singlelikelihoods = dnorm(y, mean =
       pred, sd = sd, log = T)
   sumll = sum(singlelikelihoods)
   return(sumll) }
# Example: plot the likelihood profile
    of the slope a
s.values <-
    function(x){return(likehd(c(x, t.B,
    t.sd)))}
s.likehds <- lapply(seq(1/2*t.A,
    3/2*t.A, by=.05), s.values )
```

```
plot(seq(1/2*t.A, 3/2*t.A, by=.05),
    s.likehds , type="l", xlab = "values
    of slope parameter a", ylab = "Log
    likelihood")
```

As an illustration, the last lines of the code plot the Likelihood for a range of parameter values of the slope parameter a. The result should look like the image below.



Defining the priors. In Bayesian analysis, the next step is always required: we have to specify a prior distribution for each of the model parameters. To keep things simple, we will use a uniform distribution for and normal distributions for all three parameters are used.[2]

```
# Prior distribution
prior <- function(param){
    a = param[1]
    b = param[2]
    sd = param[3]
    aprior = dunif(a, min=0, max=2*t.A,
        log = T)
    bprior = dnorm(b, mean=t.B, sd = 5,
        log = T)
    sdprior = dunif(sd, min=0,
        max=2*t.sd, log = T)
    return(aprior+bprior+sdprior)
}
```

The posterior. The product of prior by likelihood is the actual quantity that MCMC works with (it is not, strictly speaking, the posterior as it is not normalized).

```
posterior <- function(param){
   return (likehd(param) + prior(param))
}
```

Applying the MH algorithm. One of the most frequent applications of MH (as in this example) is sampling from the posterior density in Bayesian statistics.[3] The aim of the algorithm is to jump around in parameter space, but in such a way as to have the probability to land at a point be proportional to the function we sample from (this is usually called the target function). In this case, the target function is the posterior defined previously.

This is achieved by

1. starting with a random parameter vector;
2. choosing a new parameter vector near the old value based on some probability density (the proposal function), and
3. jumping to this new point with a probability $\alpha = \min\{1, g(\text{new})/g(\text{old})\}$, where g is the target.

The distribution of the parameter vectors MH visits converges to the target distribution g.

```
######## MH ################
proposalfunction <- function(param){
    return(rnorm(3,mean = param, sd=
        c(0.1,0.5,0.3)))
}

run_metropolis_MCMC <-
    function(startvalue, iterations){
    chain = array(dim = c(iterations+1,3))
    chain[1,] = startvalue
    for (i in 1:iterations){
        proposal =
            proposalfunction(chain[i,])

        probab = exP(posterior(proposal) -
            posterior(chain[i,]))
        if (runif(1) < probab){
            chain[i+1,] = proposal
        }else{
            chain[i+1,] = chain[i,]
        }
    }
    return(chain)
}

startvalue = c(4,0,10)
chain = run_metropolis_MCMC(startvalue,
    10000)

burnIn = 5000
acceptance =
    1-mean(duplicated(chain[-(1:burnIn),]))
```

The first steps of the algorithm may be biased by the initialization process; they are usually discarded for the analysis (this is referred to as the **burn-in time**).

[2] We work with the logarithms of all quantities, so that the likelihood is a sum and not a product as would usually be the case.

[3] The algorithm may be used to sample from any integrable function.

An interesting output to study is the acceptance rate: how often was a proposal rejected by the MH acceptance criterion? The acceptance rate can be influenced by the proposal function: generally, the nearer the proposal is to the latest value, the larger the acceptance rate.

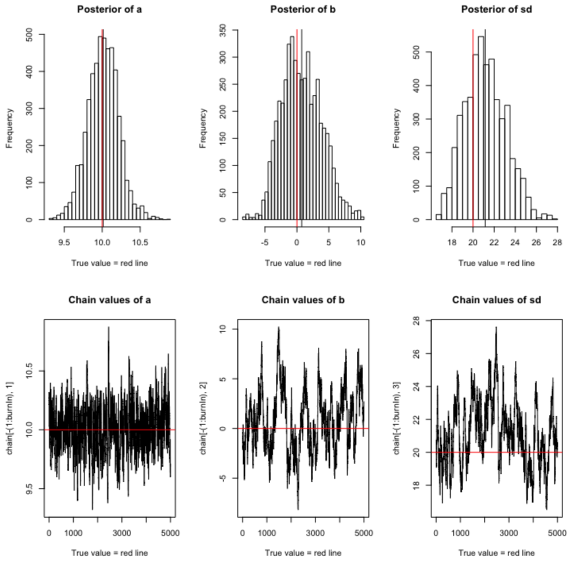
Very high acceptance rates, however, are usually not beneficial, as this implies that the algorithms is "staying" in the same neighbourhood or point, which results in suboptimal probing of the parameter space (there is very litte **mixing**). Acceptance rates between 20% and 30% are considered optimal for typical applications [25].

Finally, we can plot the results.

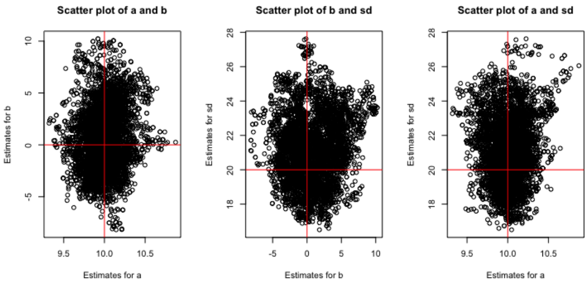
```
### Summary: #######################
 par(mfrow = c(2,3))
hist(chain[-(1:burnIn),1],nclass=30, ,
    main="Posterior of a", xlab="True
    value = red line" )
abline(v = mean(chain[-(1:burnIn),1]))
abline(v = t.A, col="red" )
hist(chain[-(1:burnIn),2],nclass=30,
    main="Posterior of b", xlab="True
    value = red line")
abline(v = mean(chain[-(1:burnIn),2]))
abline(v = t.B, col="red" )
hist(chain[-(1:burnIn),3],nclass=30,
    main="Posterior of sd", xlab="True
    value = red line")
abline(v = mean(chain[-(1:burnIn),3]) )
abline(v = t.sd, col="red" )
plot(chain[-(1:burnIn),1], type = "l",
    xlab="True value = red line" , main
    = "Chain values of a", )
abline(h = t.A, col="red" )
plot(chain[-(1:burnIn),2], type = "l",
    xlab="True value = red line" , main
    = "Chain values of b", )
abline(h = t.B, col="red" )
plot(chain[-(1:burnIn),3], type = "l",
    xlab="True value = red line" , main
    = "Chain values of sd", )
abline(h = t.sd, col="red" )

# for comparison:
summary(lm(y~x))
```

The resulting plots should look something like those seen in the column on the right: the upper row shows posterior estimates for the slope a, intercept (b) and standard deviation of the error (sd); the lower row shows the Markov Chain of parameter values. We retrieve (more or less) the original parameters that were used to create the data, and there is a certain area around the highest posterior values that also show some support by the data, which is the Bayesian equivalent of confidence intervals.



The posterior distributions above are **marginal distributions**, the joint distributions are shown below.



By way of comparison, the `lm()` function in R yields the following estimates: a – 9.9880 (se: 0.2092), b – 0.5840 (se: 3.0185), and sd – 21.34 (48 d.f.).

Exercises

Exercise 4. A group of adults are doing a simple learning experiment: when they see the two words "radio" and "ocean" appear simultaneously on a computer screen, they are asked to press the F key on the keyboard; whenever the words "radio" and "mountain" appear on the screen, they are asked to press the J key. After several practice repetitions, two new tasks are introduced: in the first, the word "radio" appears by itself and the participants are asked to provide the best response (F or J) based on what they learned before; in the second, the words "ocean" and "mountain" appear simultaneously and the participants are once again asked to provide the best response. This is repeated with 50 people. The data shows that, for the first test, 40 participants answered with F and 10 with J; while for the second test, 15 responded with F and 35 with J. Are people biased toward F or toward J for either of the two tests? To answer this question, assume a uniform prior, and use a 95% HDI to decide which biases can be declared to be credible.

5. Uncertainty

According to [12],

> the central feature of Bayesian inference is the direct quantification of uncertainty.

Bayesian approach to modeling uncertainty is particularly useful when:

- the available data is limited;
- there is some concern about overfitting;
- some facts are more likely to be true than others, but that information is not contained in the data, or
- the precise likelihood of certain facts is more important than solely determining which fact is most likely (or least likely).

The following example represents a Bayesian approach to dealing with the uncertainty of the so-called **envelope paradox**.

Example 10. You are given two indistinguishable envelopes, each containing a cheque, one being twice as much as the other. You may pick one envelope and keep the money it contains. Having chosen an envelope at will, but before inspecting it, you are given the chance to switch envelopes. Should you switch? What is the expected outcome in doing so? Explain how this game leads to infinite cycling.

Solution: let V be the (unknown) value found in the envelope after the first selection. The other envelope then contains either $\frac{1}{2}V$ or $2V$, both with probability 0.5, and the expected value of trading is

$$E[\text{trade}] = 0.5 \times \frac{1}{2}V + 0.5 \times 2V = \frac{5}{4}V > V;$$

and so it appears that trading is advantageous. Let the (still unknown) value of the cheque in the new envelope be W. The same argument shows that the expected value of trading *that* envelope is $\frac{5}{4}W > W$, so it would make sense to trade the envelope once more, and yet once more, and so on, leading to infinite cycling.

There is a Bayesian approach to the problem, however. Let V be the (uncertain) value in the original selection, and W be the (also uncertain) value in the second envelope. A proper resolution requires a joint (prior) distribution for V and W. Now, in the absence of any other information, the most we can say about this distribution using the maximum entropy principle is that $P(V < W) = P(V > W) = 0.5$.

By definition, if $V < W$, then $W = 2V$; if, on the other hand, $V > W$ then $W = \frac{V}{2}$. We now show that the expected value in both envelopes is the same, and thus that trading envelope is no better strategy than keeping the original selection.

Using Bayes' Theorem, we compute that

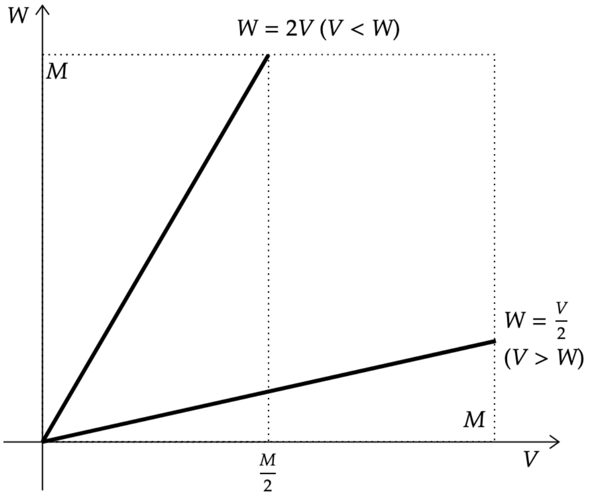
$$\begin{aligned} E[W] &= E[W|V < W]P(V < W) + E[W|V > W]P(V > W) \\ &= E[2V|V < W] \cdot 0.5 + E[0.5V|V > W] \cdot 0.5 \\ &= E[V|V < W] + 0.25 \cdot E[V|V > W], \end{aligned}$$

while

$$\begin{aligned} E[V] &= E[V|V < W]P(V < W) + E[V|V > W]P(V > W) \\ &= 0.5 \cdot E[V|V < W] + 0.5 \cdot E[V|V > W]. \end{aligned}$$

Before we can proceed any further, we must have some information about the joint distribution $P(V,W)$ (note, however, that $E[W]$ will not typically be equal to $\frac{5}{4}V$, as had been assumed at the start of the solution).

The domain Ω of the joint probability consists of those pairs (V,W) satisfying $V = 2W$ $(V > W)$ or $W = 2V$ $(V < W)$ for $0 < V,W < M$, where $M < \infty$ is some upper limit on the value of each cheque.[4]



We have assumed that the probability weight on each branch of Ω is $1/2$; if we further assume, say, that the cheque value is as likely to be any of the allowable values on these branches, then the joint distribution is

$$P(V,W) = \begin{cases} \frac{1}{M} & \text{if } V < W \\ \frac{1}{2M} & \text{if } V > W \\ 0 & \text{otherwise} \end{cases}$$

and the expectations listed above are

$$E[V|V < W] = \int_{V<W} V \cdot P(V,W)\,d\Omega = \int_0^{M/2} V \cdot \frac{1}{M}\,dV = \frac{M}{8}$$

[4] In the worst case scenario, M would have to be smaller than the total amount of wealth available to humanity throughout history, although in practice M should be substantially smaller. Obviously, a different argument will need to be made in the case $M = \infty$.

and

$$E[V|V>W]=\int_{V>W} V\cdot P(V,W)\,d\Omega=\int_0^M V\cdot\frac{1}{2M}\,dV=\frac{M}{4}.$$

Therefore,

$$E[W]=\frac{M}{8}+0.25\cdot\frac{M}{4}=\frac{3M}{16}$$

and

$$E[V]=0.5\cdot\frac{M}{8}+0.5\cdot\frac{M}{4}=\frac{3M}{16},$$

and switching the envelope does not change the expected value of the outcome. There is no paradox; no infinite cycling.

Example 11. *Bayes in the courtroom.* After the sudden death of her two baby sons, Sally Clark was sentenced by a U.K. court to life in prison in 1996. Among other errors, expert witness Sir Roy Meadow had wrongly interpreted the small probability of two cot deaths as a small probability of Clark's innocence. After a long campaign, which included the refutation of Meadow's statistics using Bayesian statistics, Clark was released in 2003. While Clark's innocence could not be proven beyond the shadow of a doubt using such methods, her culpability could also not be established beyond reasonable doubt and she was cleared. An interesting write-up of the situation can be found online [39].

6. Why Use Bayesian Methods

As discussed previously, Bayesian methods have a number powerful features: they allow analysts to

- incorporate specific previous knowledge about parameters of interest;
- logically update knowledge about the parameter after observing sample data;
- make formal probability statements about parameters of interest;
- specify model assumptions and check model quality and sensitivity to these assumptions in a straightforward manner;
- provide probability distributions rather than point estimates, and
- treat the data values in the sample as interchangeable.

6.1 Problems and Solutions

In particular, Bayesian methods are indicated in order to solve a number of problematic challenges in data analysis.

1. The dataset is small, but external related information is available: use the information in a prior.

2. The model is extremely flexible (high-variance model) and so is prone to **overfitting**: use priors that with peaks close to 0 (this is roughly equivalent to the concept of regularization in machine learning).

3. There is an interest in determining the likelihood of parameter values, rather than just producing a "best guess": construct the full posterior for the parameters/variable of interest.

6.2 Bayesian A/B Testing

A/B testing is an excellent tool for deciding whether or not to roll out incremental features. To perform an A/B test, we divide users randomly into a test and control group, then provide the new feature to the test group while letting the control group continue to experience the current version of the product.

If the randomization procedure is appropriate, we may be able attribute any difference in outcomes between the two groups to the changes we are rolling out without having to account for other sources of variation affecting the user behaviour. Before acting on these results, however, it is important to understand the likelihood that any observed differences is merely due to chance rather than to product modification.

For example, it is perfectly possible to obtain different H/T ratios between two fair coins if we only conduct a limited number of tosses; In the same manner, it is possible to observe a change between the A and B groups even if the underlying user behavior is identical.

Example 12. (derived from [28]) Wakefield Tiles is a company that sells floor tiles by mail order. They are trying to become an active player into the lucrative Chelsea market by offering a new type of tile to the region's contractors. The marketing department have conducted a pilot study and tried two different marketing methods:

- A – sending a colourful brochure in the mail to invite contractors to visit the company's showroom;
- B – sending a colourful brochure in the mail to invite contractors to visit the company's showroom, while including free tile samples.

The marketing department sent out 16 mail packages of type A and 16 mail packages of type B. Four Chelseaites that received a package of type A visited the showroom, while 8 of those receiving a package of type B did the same. The company is aware that:

- a mailing of type A costs 30$ (includes the printing cost and postage);
- a mailing of type B costs 300$ (additionnaly includes the cost of the free tile samples);
- a visit to the showroom yields, on average, 1000$ in revenue during the next year.

Which of the methods (A or B) is most advantageous to Wakefield Tiles?

Solution: the Bayesian solution requires the construction of a prior distribution and of a **generative model**; as part of the generative model, we will need to produce n replicates of samples from the binomial distribution (which can be done in R using `rbinom(n,size,prob)`).

The binomial distribution simulates `n` times the number of "successes" when performing `size` trials (mailings), where the probability of a "success" is `prob`. A commonly used prior for `prob` is the uniform distribution $U(0,1)$, from which we can sample in R *via* `runif(1, min = 0, max = 1)`.

```
# Number of replicates from the prior
n.draws <- 200000

# Prior
# This generates a probability of
# success for mailings A and B,
# for each of the replicates
prior <- data.frame(p.A = runif(n.draws,
    0, 1), p.B = runif(n.draws, 0, 1))

# Generative model
# This tells us how many visitors to
    expect
# for mailing types A, B
generative.model <- function(p.A, p.B) {
  visitors.A <- rbinom(1, 16, p.A)
  visitors.B <- rbinom(1, 16, p.B)
  c(visitors.A = visitors.A, visitors.B
      = visitors.B)
}

# Simulate data using the parameters
# from the prior and the gen. model
# This generates the actual number of
# visitors for each replicate
sim.data <- as.data.frame(
    t(sapply(1:n.draws, function(i) {
  generative.model(prior$p.A[i],
      prior$p.B[i])})))

# Only those prior probabilities for
# which the generative model match the
# observed data are retained
posterior <- prior[sim.data$visitors.A
    == 4 & sim.data$visitors.B == 8, ]

# Visualize the posteriors
par(mfrow = c(1,3))
hist(posterior$p.A, main = "Posterior --
    probability of success with mailing
    A", xlab="p.A")
hist(posterior$p.B, main = "Posterior --
    probability of success with mailing
    B", xlab="p.B")
plot(posterior,main = "Scatterplot of
    probabilitie of success for mailing
    types A and B", xlab="p.A",
    ylab="p.B")
```

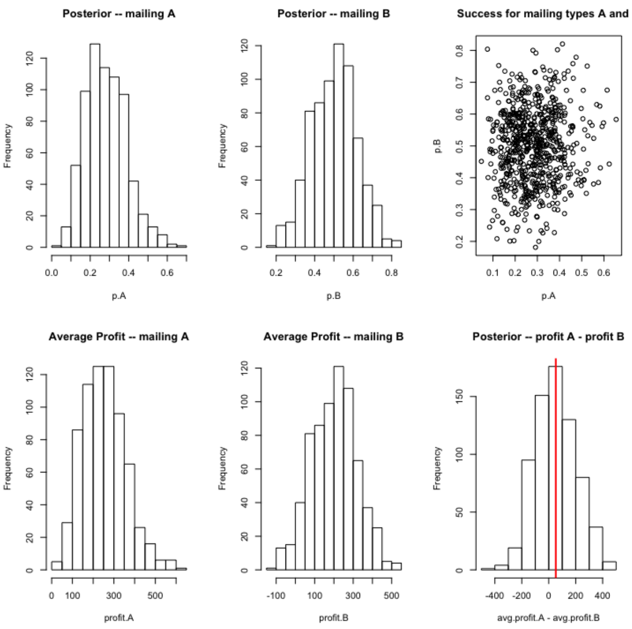
The posterior distributions for the probability of success for each mailing types are shown in the figure below.

In order to estimate the average profit for each mailing type, we use the posterior distributions for the probability of success.

```
# Compute the estimated average profit
    per mailing type
avg.profit.A <- -30 + posterior$p.A *
    1000
avg.profit.B <- -300 + posterior$p.B *
    1000
hist(avg.profit.A, main = "Average
    Profit -- mailing A",
    xlab="profit.A")
hist(avg.profit.B, main = "Average
    Profit -- mailing B",
    xlab="profit.B")
```

The expected profit is thus given by the following code:

```
# Total expected profit
hist(avg.profit.A - avg.profit.B)
expected.avg.profit.diff <-
    mean(avg.profit.A - avg.profit.B)
abline(v = expected.avg.profit.diff ,
    col = "red", lwd =2)
```



The expected profit for mailing type A is around 52$ higher than for mailing type B (your numbers may vary). Keeping it simple seems to be a better idea in this context.

7. Summary

What?

- Bayesian data analysis is a flexible method to fit any type of statistical model.
- Maximum likelihood is actually a special case of Bayesian model fitting.

Why?

- Makes it possible to define highly customizable models.
- Makes it possible to include information from many sources, such as data and expert knowledge.
- Quantifies and retains the uncertainty in parameter estimates and predictions.

How?

- R! Using ABC, MCMCpack, JAGS, STAN, R-inla, Python, etc.

Exercises – Solutions

- **Exercise 1:**

$$P(L|Y) = \frac{P(Y|L)P(L)}{P(Y)}$$

Note that

$$P(Y) = P(Y,L)+P(Y,\overline{L}) = P(Y|L)P(L)+P(Y|\overline{L})P(\overline{L}),$$

so

$$P(L|Y) = \frac{(.55)(.52)}{(.55)(.52)+(.85)(.48)} = 0.41.$$

- **Exercise 2:**

Step 1: assign events to A or X. You want to know what a woman's probability of having cancer is, given a positive mammogram. For this problem, actually having cancer is A and a positive test result is X.

Step 2: list out the parts of the equation (this makes it easier to work the actual equation):

$$P(A) = 0.01, P(\overline{A}) = 0.99, P(X|A) = 0.9, P(X|\overline{A}).$$

Step 3: insert the parts into the equation and solve. Note that as this is a medical test we have

$$\frac{0.9 \cdot 0.01}{(0.9)(0.01)+(0.08)(0.99)} = 0.10.$$

The probability of a woman having cancer, given a positive test result, is thus 10%.

- **Exercise 3**

1. To justify a prior, we might say that our strength of fairness is equivalent to having previously seen the coin flipped 100 times and coming up heads in 50% of those flips. Hence the prior would be Beta$(\theta|50,50)$ (this is not the only correct answer, of course; you might instead be more confident, and use, say, Beta$(\theta|500,500)$ if you suppose you've previously seen 1,000 flips with 50% heads).

   The posterior is then Beta$(\theta|50+9,50+1)$, which has a mean of $\frac{59}{59+51} = 0.536$. This is the predicted probability of heads for the next (11th) flip.

2. In this case, we use a Beta$(\theta|0.5,0.5)$ prior, like the one used in Example5, because it expresses a belief that the coin is either head-biased or tail-biased.

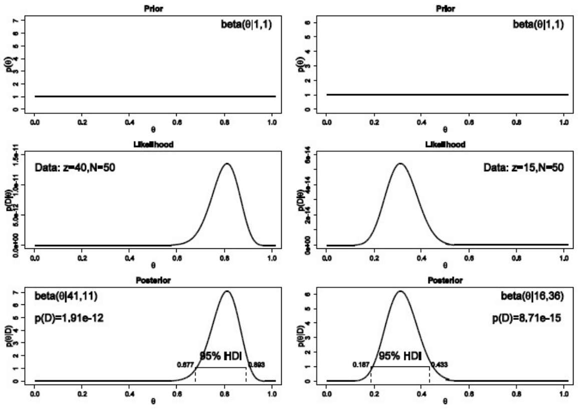
   The posterior is Beta$(\theta|0.5+9,0.5+1)$, which has a mean of $\frac{9.5}{9.5+1.5} = 0.863$. This is the predicted probability of heads for the next (11th) flip. Notice that it is quite different than the conclusion from Part 1.

- **Exercise 4**

The commands

```
> post = BernBeta(c(1,1),
c(rep(1,40),rep(0,10)))
> post = BernBeta(c(1,1),
c(rep(1,15),rep(0,35)))
```

yield the display below.



In both cases, the 95% HDI excludes $\theta = 0.5$, and so we conclude that people are indeed biased in their responses, toward F in the first case and toward J in the second case.

References

[1] Bayes, T. [1763], An Essay towards solving a Problem in the Doctrine of Chances, *Phil. Trans. Royal Society London*.

[2] Gill, J. [2002], Bayesian Methods for the Social and Behavioral Sciences, Boca Raton, Florida: CRC Press.

[3] Hitchcock, D. [2014], Introduction to Bayesian Data Analysis, Department of Statistics, University of South Carolina, US, course notes.

[4] Berger, J.O. [1985], Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York, 2nd edition.

[5] Robert, C.F. [2006], Le choix bayésien - Principes et pratique, Springer-Verlag France, Paris.

[6] Kruschke, J.K. [2009], Highlighting: A canonical experiment. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 153–185). Elsevier Academic Press.

[7] Sivia, D.S., Skilling, J. [2006], Data Analysis: A Bayesian Tutorial (2nd ed.), Oxford Science.

[8] Silver, N. [2012], The Signal and the Noise, Penguin.

[9] Jaynes, E.T. [2003], Probability Theory: the Logic of Science, Cambridge Press.

[10] Kruschke, J.K. [2011], Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan (2nd ed.), Academic Press

[11] Barber, D. [2012], Bayesian Reasoning and Machine Learning, Cambridge Press.

[12] Gelman, A., Carloin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. [2013], Bayesian Data Analysis (3rd ed.), CRC Press.

[13] Baath, R. [2015], Introduction to Bayesian Data Analysis with R, UseR!

[14] Oliphant, T.E. [2006], A Bayesian perspective on estimating mean variance, and standard-deviation from data, All Faculty Publications 278, BYU.

[15] **Reference Priors and Maximum Entropy** (lecture notes)

[16] **Bayesian Inference** (lecture notes)

[17] **MCMC algorithms for fitting Bayesian models** (lecture notes)

[18] **Bayesian Inference: Metropolis-Hastings Sampling**

[19] **Bayesian Statistics** (scholarpedia article)

[20] **Bayes' Theorem Problems, Definition and Examples** (statisticshowto.com)

[21] **Bayesian AB Testing** (Lyst)

[22] **Introduction to Bayesian Inference** (datascience.com)

[23] **Maximum Entropy Priors** (moreisdifferent.com)

[24] **Maximum Entropy**

[25] **MCMC chain analysis and convergence diagnostics with coda in R** (theoreticalecology.wordpress.com)

[26] **A simple Metropolis-Hastings MCMC in R** (theoreticalecology.wordpress.com)

[27] **Conjugate Priors** (Wikipedia)

[28] **Bayesian A/B Testing for Swedish Fish Incorporated** (tutorial)

[29] **Maximum Entropy Methods Tutorial: A Simple Example: The Taxicab** (video)

[30] **Maximum Entropy Methods Tutorial: MaxEnt applied to Taxicab Example Part 1** (video)

[31] **PYMC3** (documentation)

[32] Chipman, H.A., George, E.I., McCulloch, R.E. [2010], BART: Bayesian additive regression trees, Annals of Applied Statistics 6 (1), 266–298.

[33] Chipman, H.A., George, E.I., McCulloch, R.E. [1998], Bayesian CART model search (with discussion and a rejoinder by the authors), J. Amer. Statist. Assoc. 93 935–960.

[34] Hernandez, B., Raftery, A.E., Pennington, S.R., Parnell, A.C. [2015], Bayesian Additive Regression Trees Using Bayesian Model Averaging, Technical Report no. 636, Department of Statistics, University of Washington.

[35] Kabacoff, R.I. [2011], R in Action, Second Edition: Data analysis and graphics with R.

[36] **BayesTree** (CRAN package)

[37] **BART** (slides)

[38] **Doing Bayesian Data Analysis** (R Code).

[39] **Sally Clark is Wrongly Convicted of Murdering Her Children** (Bayesians Without Borders).

[40] Wilkinson, D.J. [2007], **Bayesian Methods in Bioinformatics and Computational Systems Biology**, Briefings in Bioinformatics, v.8, n.2, 109–116.

[41] Poldrac, R.A. [2006], 'Can cognitive processes be inferred from neuroimaging data?, *Trends Cogn. Sci.* 10(2):59-63.