Mattias Villani

Bayesian Learning [rough draft]

A GENTLE INTRODUCTION

Some publisher

PUBLISHED BY SOME PUBLISHER

TYPESET BY LaTeX USING TEMPLATE FROM TUFTE-LATEX.GITHUB.IO

First edition, December 2020

Contents

Preface

Who is this book for?

This book can be used as a first book in Bayesian statistics at the advanced undergraduate or master level. The book is written so that it can accomodate also students in engineering and computer science who are interested in Bayesian learning for applications in the field of Machine Learning.

In fact, the book grew out of a Bayesian course that I taught for groups of heterogenuous students with roughly half of students from statistics and the other half from engineering and computer science, often with an interest in machine learning. To my surprise, I found that it was indeed possible to teach the same material to all students, even if half the class had a much more extensive background in statistics. The course had always very favorable reviews from the students and not a single student has complained over the years on it being too easy or too hard. There are two main explanations for this. First, since most bachelor level Statistics are non-Bayesian in methods and thinking, taking a first course in Bayesian inference is in some way like starting from scratch. Sure, there are several overlapping concepts and probability is of course the underlying technical language (although with highly different interpretations), but there are nevertheless a lot of effort spend in basic statistics courses that are not needed prerequisites for a Bayesian course. Second, my courses are very computational, as is most of the Bayesian field, with a lot of computer labs and also a partly computerized exam. Engineering and particularly computer science students tend to have a comparative advantage in computing and programming. So the additional time that students from statistics had to spend on programming, computer science students could spend on catching up on statistical concepts. In the end, everyone seemed put in the same number of hours and everyone was happy with the learning experience. In order to accomodate both groups of students, my lectures covers also some rather elementary concepts, especially in the early part of the course, but then rather quickly moves over to territory unknown to all stu-

dents. This book is written in the same style using Tufte style margin notes and figures to fill in potential missing gaps in probability and statistics, without breaking the flow of the main text.

Programming is useful for the exercises, or at least basic familiarity with R, Python or Julia or a similar datacentric language. I will use pseudo code for certain smaller algorithms and Julia for real code; Julia is used to present algorithms in the book since the ability to use mathematical symbols in Julia (via unicode) makes the code easy to read, almost like pseudo code. All graphs were made in Julia using the Plots package with GR as backend.

Why the term Bayesian learning?

I have used the term Bayesian *learning* in the book's title instead of Bayesian *inference* of Bayesian *statistics*. There are several reasons for this.

First, I want my courses and this book to be welcoming to students in fields neighboring statistics, such as machine learning, computer science, and parts of engineering. This reflects my strong belief that a modern statistician or machine learner should be a little of a renaissance person that understands both probability and statistical modelling, and computing. The ideal class is therefore a mix of students from nearby disciplines that learn for each others competences as much as they learn from my classes or this book.

Second, the term learning instead of inference was chosen since Bayesian statistics is about learning from data, often in a very sequential way where incrementally collected information updates our knowledge about the world.

Finally, the title is meant to convey the message that this is not a traditional book in statistics. The approach taken here, especially in later chapters, is very computationally driven with many algorithms for real-world data analysis. It is also inspired by machine learning in that much of the focus is given to prediction and decision making, and almost none to hypothesis testing.

Acknowledgment

This section will be much more complete when the book is finished, but I want to note already now that this book has been influenced by many other excellent textbooks on Bayesian methods. This is particularly true for two books that I have used as course literature over the years. I taught my first Bayes course in the year of 2000 using the book *Statistical Inference - An Integrated Approach* by Migon and Gamerman. Second, I have used the book *Bayesian Data Analysis* by Gelman et al. for a number of years while teaching. I imagine that

I have been more influenced by these two books than I know, and I thank the authors for taking the time to write them. I now appreciate them even more: it takes a lot of time to write a book!

The Bayesics

Learning probability models

A central task in statistics and machine learning is to infer an unknown parameter $\theta \in \Theta$ in a probability model $p(X_1, \ldots, X_n|\theta)$ from a dataset of n observations $x_1, \ldots, x_n$. The **parameter space** Θ is the set of allowed parameter values. Some examples of problems with a single parameter are learning the voting share of a political party from exit polls, predicting the number of bugs in a software release and infering a one-dimensional measure of a persons intelligence from IQ tests.

parameter space

Most problems require models with more than one parameter. A prominent example with an extremely large number of parameters are the deep neural network models widely used in artificial intelligence (AI); such models often have millions of network weights that have to be learned from training data. However, to focus on ideas and easy derivations, we will keep things as simple as possible in the first two chapters and only consider models with a single parameter. Later chapters tackle more complex models and present methods specifically designed for models with many parameters.

Figure 1: Artificial intelligence is often based on Bayesian learning.

The initial chapters will be focused on learning parameters in models. It is important to remember that parameter inference is usually an intermediate step toward the final aim of prediction or decision making under uncertainty; for example, the predictions and decisions of a robot are based on a probability model with network weights learned from training data. Throughout this book we will exclusively work with probability models. Probability models have the distinct advantage of giving a precise quantification of uncertainty that can be directly used for decision making in the real world. The Bayesian approach to predictions and decisions will be presented in the chapter Prediction and Decision making.

Statistical distributions and independent data

We will initially assume that the observations $X_1, \ldots, X_n$ are *independent and identically distributed* (**iid**) conditional on θ so that we can write the joint distribution as a product

iid

$$p(X_1, \ldots, X_n|\theta) = \prod_{i=1}^{n} p(X_i|\theta).$$

We denote this by $X_1, \ldots, X_n|\theta \overset{\text{iid}}{\sim} p(X|\theta)$. In this setting we can refer to 'the probability model' as the probability distribution $p(X|\theta)$ for a single observation.

EXAMPLE: A binary random variable $X \in \{0, 1\}$ follows a **Bernoulli distribution** if

Bernoulli distribution

$$\Pr(X = x|\theta) = \begin{cases} \theta & \text{for } x = 1 \\ 1 - \theta & \text{for } x = 0 \end{cases}$$

which can be written more compactly as

$$\Pr(X = x|\theta) = \theta^x (1 - \theta)^{1-x}. \tag{1}$$

A typical example of iid Bernoulli data occurs when a coin is flipped n times (also called **Bernoulli trials**) and the sequence of heads ($x = 1$) and tails ($x = 0$) are recorded. It is common to refer to the outcome $X = 1$ as a success, and $X = 0$ as a failure. The Bernoulli distribution is illustrated in Figure 2.

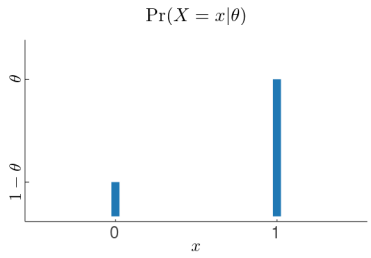


Figure 2: Bernoulli distribution with success probability $\theta = 0.8$.

Bernoulli trials

We make the usual distinction between *random variables* denoted by capital letters and their *realizations* (*data*), so $X = x$ means a random variable X with outcome x. As we will see later on, this distinction will often be less relevant in a Bayesian world where all inferences are conditioned on the observed data; we will therefore be more sloppy with this distinction in later chapters, but no harm will come from this.

The likelihood function and maximum likelihood estimation

The likelihood function is a key component of Bayesian learning, and indeed in all of Statistics. Given a probability model $p(X_1, \ldots, X_n|\theta)$ the **likelihood function** $p(x_1, \ldots, x_n|\theta)$ is the *joint* probability of observing the data set $x_1, \ldots, x_n$ *considered as a function* of the parameter θ. Again, if the data are iid we can express the likelihood in terms of the univariate distributions $p(X|\theta)$ as

likelihood function

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta). \tag{2}$$

EXAMPLE: In the case of iid Bernoulli data the likelihood function is simply obtained by multiplying together the probability of success θ for the observations where $x_i = 1$ and probability of failure $1 - \theta$ when $x_i = 0$, giving the likelihood

$$p(x_1, \ldots, x_n|\theta) = \theta^s(1-\theta)^f, \tag{3}$$

where $s = \sum_{i=1}^{n} x_i$ is the number of successes in the sample, and $f = n - s$ is the number of failures.

It is absolutely essential to have mental image of the likelihood function when thinking about statistical modeling. Figure 3 illustrates the likelihood function for Bernoulli model when $s = 4$ successes was obtained in $n = 10$ trials. Figure 4 is another example with the same proportion of 40% successes, but here in $n = 100$ trials; note how the large dataset makes the likelihood more concentrated, more informative regarding the plausibility of different θ values. Finally, Figure 5 is again the likelihood from only $n = 10$ trials, but now with $s = 9$ successes; this data is much more likely to have resulted from a data generating process with θ closer to one.

Figures 3-5 nicely illustrates how the likelihood function can inform us about the plausibility of different θ for any given dataset. If we want to select a single value, an estimate of θ, a natural candidate is the widely used **maximum likelihood estimator**

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} p(x_1, \ldots, x_n|\theta). \tag{4}$$

It makes some intuitive sense to estimate θ by the value that maximizes the probability of the observed data; the estimator $\hat{\theta}_{\text{MLE}}$ also enjoys several other attractive properties, particularly in large samples, i.e. when n is large.

It is quite easy to derive $\hat{\theta}_{\text{MLE}}$ for iid Bernoulli data. Rather than maximizing $p(x_1, \ldots, x_n|\theta)$ directly with respect to θ it is often easier to maximize the *log-likelihood function*

$$\log p(x_1, \ldots, x_n|\theta) = s \log \theta + f \log(1-\theta).$$

Since the logarithm is a monotonically increasing function we obtain the same estimator if we maximize the likelihood or the log-likehood function. We can now easily find $\hat{\theta}_{\text{MLE}}$ by taking the first derivate of the log-likelihood function with respect to θ, setting that derivative to zero and solving for θ. Solving

$$\frac{d \log p(x_1, \ldots, x_n|\theta)}{d\theta} = \frac{s}{\theta} - \frac{f}{1-\theta} = 0,$$

gives the unique solution $\hat{\theta}_{\text{MLE}} = s/n$, the fraction of successes in the data. It is straightforward to show that this indeed a maximum by checking that the second derivative is negative at $\hat{\theta}_{\text{MLE}}$.

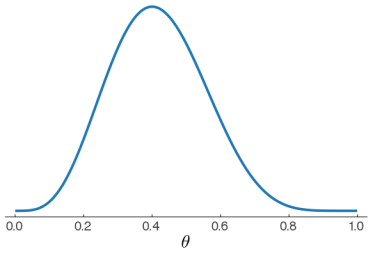


Figure 3: Bernoulli likelihood function for $n = 10$ and $s = 4$.

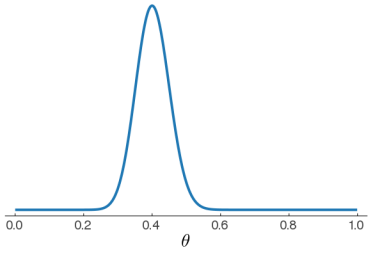


Figure 4: Bernoulli likelihood function for $n = 100$ and $s = 40$.

maximum likelihood estimator

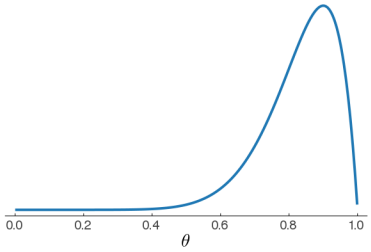


Figure 5: Bernoulli likelihood function for $n = 10$ and $s = 9$.

Binomial distribution

$S \sim \text{Binom}(n, \theta)$
Support: $S \in 0, 1, \ldots, n$

$$p(s) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$$

$\mathbb{E}(X) = n\theta$
$\mathbb{V}(X) = n\theta(1-\theta)$

Figure 6: The binomial distribution.

The maximum likelihood estimator is **unbiased** in this example, i.e. it is correct on average over all possible samples from the model:

unbiased

$$\mathbb{E}\left[\hat{\theta}_{MLE}(X_1,\ldots,X_n)\right] = \mathbb{E}\left(\frac{S}{n}\right) = \frac{n\theta}{n} = \theta,$$

where we have written out explicitly that an estimator is function of the sample. Note that the number of successes is random in this calculation as we are considering the variability over all possible samples, hence the use of capital letter S. We have also used that if $X_1,\ldots,X_n|\theta \overset{iid}{\sim}$ Bernoulli then $S|\theta \sim$ Binomial(n,θ) with mean $E(S) = n\theta$; see Figure 7 for an example of a **Binomial distribution**.

Binomial distribution

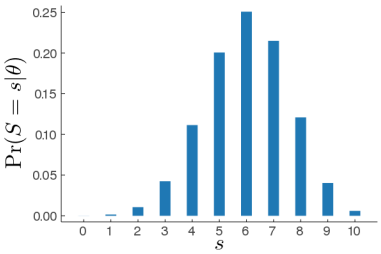


Figure 7: Binomial distribution with $n = 10$ and $\theta = 0.7$.

The **sampling variance** of an estimator is often used to assess the quality of an estimator. It is easily calculated for $\hat{\theta}_{MLE}$ in the Bernoulli example as

sampling variance

$$\mathbb{V}\left[\hat{\theta}_{MLE}(X_1,\ldots,X_n)\right] = \mathbb{V}\left(\frac{S}{n}\right) = \frac{1}{n^2}\mathbb{V}\left(S\right) = \frac{\theta(1-\theta)}{n},$$

since $\mathbb{V}(S) = n\theta(1-\theta)$ when $S|\theta \sim$ Binomial(n,θ).

It is important to understand that the above mean and variance of $\hat{\theta}_{MLE}$ are computed with respect to the **sampling distribution**, i.e. the distribution of the estimator as we repeatedly sample new datasets of size n from the assumed data generating process. They are long run properties of the estimation method, telling us how the estimator would perform on average over many repeatedly sampled datasets. Such long run properties play a very limited role in the Bayesian approach to inference where one can directly condition the inferences on the single dataset that we have observed. While the sampling properties of $\hat{\theta}_{MLE}$ are not used in Bayesian world, the likelihood *function* is at the core of Bayesian learning.

sampling distribution

The likelihood functions in Figures 3-5 *look like* a probability distribution for θ, and it is tempting to compute probabilities for θ, for example $\Pr(\theta \leq c|x_1,\ldots,x_n)$ for some c. Of course, such probabilities only makes sense if θ is a random variable, and we have so far considered θ to be a fixed unknown constant. So while $p(X_1,\ldots,X_n|\theta)$ *is* a probability distribution for a random sample $X_1,\ldots,X_n$ for a fixed θ, the likelihood function is only the probability of a *fixed* sample $x_1,\ldots,x_n$ considered as function of θ; the likelihood is therefore *not* a probability distribution for θ. Figure 8 reminds us of this error.

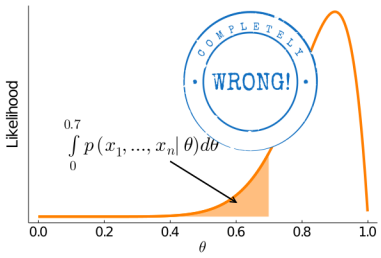
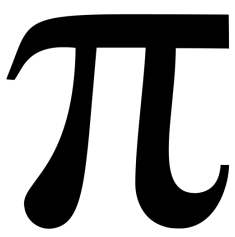


Figure 8: The likelihood function is **not** a probability distribution for θ.

This is somewhat disappointing since having a probability distribution for θ would be very useful, for example when making a decision whose consequences depend on the unknown θ; see Chapter Prediction and Decision making. But again, it only makes sense to speak about probabilities for θ when θ is random. And this is where our Bayesian story begins.

Subjective Probability

What is the probability that the 10th decimal of π is 3? This may seem like a silly question since there is nothing intrinsically random about the 10th decimal of π; it is a fixed quantity that does not vary. A Bayesian will however argue that if *you do not know its value* then you should express that uncertainty by a probability distribution. The Italian mathematician Bruno de Finetti, one of the founders of this school of probability, has expressed this well:



> The only relevant thing is uncertainty - the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense determined, or known by other people, and so on, is of no consequence.
>
> **Bruno de Finetti** in his 1974 book 'A Theory of Probability' Vol 1.

Probability is the language of uncertainty and Bayesian learning is based on a **subjective probability**. A subjective probability measures the **personal degree of belief** of a person. Since different person have difference knowledge and experience, such beliefs will vary between persons. A person that has no idea about the 10th decimal of π may use a uniform distribution on the integers 0-9. Someone else may however know this decimal with certainty and assigns a probability of 1 to that outcome. Again, whether or not the event is in some sense intrinsically random or not is of no consequence; the only relevant thing is *your* uncertainty. Einstein's famous statement "God does not play dice with the universe" may interesting to ponder about, but has not bearing on subjective probability and Bayesian learning.

subjective probability

personal degree of belief

Figure 9: Bruno de Finetti, 1906-1985, a founder of subjective probability.

The notion of probability in Bayesian learning is therefore radically different from the frequentist interpretation of probability taught in most basic statistics classes. The **frequentist probability** of an event A is defined as the limiting proportion of times that event A occurs in an (imagined) infinite number of repetitions of an experiment; for example the tossing a coin with the event of interest $\mathrm{A} = \{\mathrm{Heads}\}$. A subjective probability measure is instead defined as the personal degree of belief in the event A for a person. Note that subjective probabilities can be used to quantify uncertainties also for events that are unrepeatable, for example the probability of a nuclear disaster at a particular location. A subjective probability distribution can also contain useful information that may not directly come from observed data. As we will see, the Bayesian approach combines such subjective information with objective data in a natural way.

frequentist probability

Luckily, the computational rules for probabilities are the same for both frequentist and subjective interpretations of probability; for

example $0 \leq \Pr(A) \leq 1$ and $\Pr(\mathrm{A} \cap \mathrm{B}) = \Pr(\mathrm{A}) + \Pr(\mathrm{B})$ when A and B are disjoint events. The rules can be motivated by considering subjective probabilities as the result of pricing of bets. Imagine that you are given the chance to enter a bet where you win \$1 if event A occurs. How much would you be willing to pay that bet? Surely not more than \$1 as then you would loose money with certainty. If you strongly believe that A will occur you would probably be willing to pay closer to \$1, but if you believe that A is nearly impossible your price for the bet would be close to \$0. The highest price that you would be willing to pay for the bet is your subjective probability in the event A. Given this setup one can easily show that your subjective probabilities must satisfy the axioms for probabilities otherwise you would be willing to enter a sequence of bets where you would loose an infinite amount with certainty; this is the so called **dutch book argument**. Objections have been raised against this argument, for example that the utility from the bet may not linearly increasing with the monetary gain, and some people may even get utility just by the excitement in gambling; subsequent refinements of this argument have therefore completely disposed with the notion of money in favor of a more general notion of utility; see the chapter Prediction and Decision making.

dutch book argument

Bayesian Learning

The general recipe for Bayesian learning about an event A is:

- Formulate your subjective *prior beliefs* $\Pr(\mathrm{A})$ about A.
- *Collect data* that inform you about A.
- *Update* your prior beliefs with the observed data.

The big question is *how* to update prior beliefs with data. Bayesian learning gets its name from using Bayes' theorem for this updating. The most basic version of **Bayes' theorem** computes the probability of an event A given the known occurance of some other event B as

$$\Pr(\mathrm{A}|\mathrm{B}) = \frac{\Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{B})}$$

One way to think about this result is that it 'reverses the conditioning', i.e. it computes $\Pr(\mathrm{A}|\mathrm{B})$ from $\Pr(\mathrm{B}|\mathrm{A})$.

Bayes' theorem will be used to infer an unknown parameter θ in a probability model, but let us first use the theorem to solve a simple problem. Imagine that you have taken a test for a specific latent disease and that the test was unfortunately positive. The doctor tells you that $\Pr(\mathrm{B}|\mathrm{A}) = 0.9$ where $A = \{$'Have disease'$\}$ and

Bayes' theorem

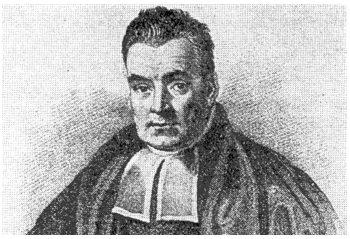


Figure 10: Reverend Thomas Bayes, ca 1701-1761, whose famous theorem was published posthumously. Interestingly, we are not quite sure that the man in the photo actually is Thomas Bayes. Probably not.

$B = \{\text{'Positive test'}\}$ and also $\Pr(\text{B}|\text{A}^\text{c}) = 0.05$, where A^c is the complement to A, i.e. the event that you do not have the disease. Hence, a positive test is very unlikely if you do not have the disease, so you start to worry. But what you really want to know is the probability of having the disease given a positive test, i.e. $\Pr(\text{A}|\text{B})$. To compute this you need to know the so called *prior* probability of A before you took the test. The doctor tells you that only one in ten thousand has the disease and you set $\Pr(\text{A}) = 0.0001$. Bayes' theorem then gives

$$\Pr(\text{A}|\text{B}) = \frac{\Pr(\text{B}|\text{A})\Pr(\text{A})}{\Pr(\text{B})} = \frac{\Pr(\text{B}|\text{A})\Pr(\text{A})}{\Pr(\text{B}|\text{A})\Pr(\text{A}) + \Pr(\text{B}|\text{A}^\text{c})\Pr(\text{A}^\text{c})} \approx 0.0018,$$

where we have expressed $\Pr(\text{B})$ in the numerator using a version of the **law of total probability**. Hence, even though the test has increased the probability of having the disease by a factor of 18 from the initial $\Pr(\text{A}) = 0.0001$, the probability of actually having the disease is still tiny. The lesson here is that prior probabilities matter.

law of total probability

To see how Bayes' theorem can be used for Bayesian learning from data, let us consider the event $\text{B} = \{\text{'Data } x_1, \ldots, x_n \text{ was observed'}\}$ which we write simply as $\text{B} = \{x_1, \ldots, x_n\}$. We can now use Bayes' theorem to update the initial beliefs $\Pr(\text{A})$ about some event A with data $\text{B} = \{x_1, \ldots, x_n\}$ by the formula

$$\Pr(\text{A}|x_1, \ldots, x_n) = \frac{\Pr(x_1, \ldots, x_n|\text{A})\Pr(\text{A})}{\Pr(x_1, \ldots, x_n)}.$$

The initial belief $\Pr(\text{A})$ is called a **prior** since it refer to beliefs about A *before* the data $x_1, \ldots, x_n$ was observed. Likewise $\Pr(\text{A}|x_1, \ldots, x_n)$ is referred to as the **posterior** since it is the probability of A *after* data was observed.

prior

posterior

Let us now show how Bayes' theorem can be used to infer a parameter in a probability model $p(X_1, \ldots, X_n|\theta)$. We first take a simplified approach where the only possible parameter values are on a grid of values $\theta_1, \theta_2, \ldots, \theta_K$. Let $B = \{x_1, \ldots, x_n\}$ be the event of observing a specific dataset and $A_k = \{\theta_k\}$ be the event that $\theta = \theta_k$. The posterior probability for each $A_k = \{\theta_k\}$ is then

$$\Pr(\theta_k|x_1, \ldots, x_n) = \frac{\Pr(x_1, \ldots, x_n|\theta_k)\Pr(\theta_k)}{\sum_{j=1}^K \Pr(x_1, \ldots, x_n|\theta_j)\Pr(\theta_j)}. \tag{5}$$

Note how we again used the law of total probability in the denominator to express $\Pr(\text{B}) = \Pr(x_1, \ldots, x_n)$. This denominator is only there to guarantee that the posterior is a probability distribution, i.e. that $\sum_{j=1}^K \Pr(\theta_j|x_1, \ldots, x_n) = 1$.

The really interesting stuff is however in the numerator of (5) and we will therefore often write Bayes' theorem in proportional form

$$\Pr(\theta_k|x_1, \ldots, x_n) \propto \Pr(x_1, \ldots, x_n|\theta_k)\Pr(\theta_k), \tag{6}$$

where the symbol $\propto$ is read as 'is proportional to', i.e. a multiplicative normalizing constant is missing in the expression. Now here is the really crucial thing: the factor $\Pr(x_1, \ldots, x_n|\theta_k)$ in Equation (6) is the *likelihood function* evaluated in the point θ_k. Equation (6) therefore expresses the fundamental idea in Bayesian learning:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}. \tag{7}$$

Figure 12-14 illustrates the updating from prior to posterior for the Bernoulli model with data $n = 10$ and $s = 9$ over a grid of θ values. Note how the posterior is a compromise between the prior information and the data information (likelihood).

Finally, taking a finer and finer grid in Equation 5 we get the following Bayes' theorem for a continuous parameter θ in the limit

$$p(\theta|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{\int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta'} \tag{8}$$

where $p(\theta)$ is now a continuous **prior density** that gets updated with new data via the likelihood function $p(x_1, \ldots, x_n|\theta)$ to a **posterior density** $p(\theta|x_1, \ldots, x_n)$. The normalizing constant is now given by an integral over θ and is a continuous version of the law of total probability. We can again hide the unimportant normalizing constant to get the nicer form

$$p(\theta|x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n|\theta)p(\theta). \tag{9}$$

It is important to note that the posterior distribution $p(\theta|x_1, \ldots, x_n)$ *is* a probability distribution for the parameter θ; it completely describes the knowledge about θ for a person with the prior $p(\theta)$ after having observed the data $x_1, \ldots, x_n$. Remember that the likelihood can not be used to compute probabilities for θ. With a posterior distribution we actually *can* compute $\Pr(\theta \leq c|x_1, \ldots, x_n) = \int p(\theta \leq c|x_1, \ldots, x_n)d\theta$ or any other probability of interest. It is the prior $p(\theta)$ that makes it possible to use Bayes' theorem to revert the conditioning in the likelihood $p(x_1, \ldots, x_n|\theta)$ into the conditional probability that we really care about, the posterior $p(\theta|x_1, \ldots, x_n)$; but you need the prior to get the posterior. As Leonard Jimmie Savage, a founder of Bayesian analysis, has famously said:

> You can't cook the Bayesian omelet
> without breaking the Bayesian eggs.
>
> **Leonard Jimmy Savage**

The ability to use prior information is a strength, especially when one has to make an decision based by very weak data. Later in the book we will see how priors can be used to convey the idea that

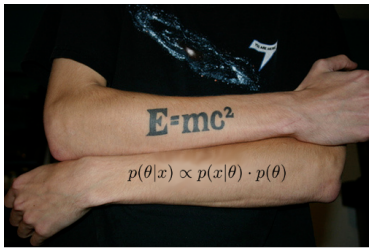


Figure 11: Great theorems make great tattoos.

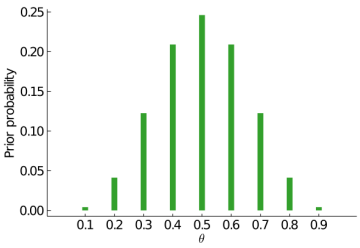


Figure 12: Prior for Bernoulli example.

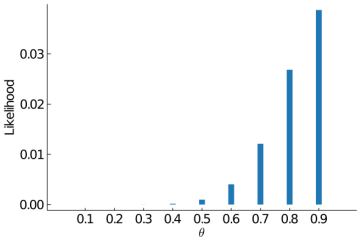


Figure 13: Likelihood for Bernoulli model with $n = 10$ and $s = 9$.

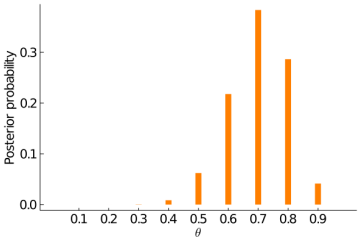


Figure 14: Posterior for Bernoulli model with $n = 10$ and $s = 9$.

prior density

posterior density

Figure 15: Making a Bayesian omelet.

a functional relationship between two variables is in some sense smooth, and how this can prevent models from overfitting the data. Nevertheless, the subjective elements of a Bayesian analysis can complicate the reporting of scientific evidence, where objectivity is the ideal. One can argue that objectivity is simply unattainable, and that the supposedly objective alternatives to Bayesian learning just sweeps the subjective elements under the carpet. A more pragmatic Bayesian approach for scientific communication is presented in Section Invariant priors where priors are intentionally chosen to be 'non-informative', in the sense of having a minimal influence on the posterior.

There are also two aspects of a Bayesian approach that gives it a clear scientific character. The prior distribution is subjective, and therefore varies from person to person, but the rule that updates the beliefs with new data is objective: we *should* use Bayes' theorem and the data *should* enter the updating *only through the likelihood function*. The word 'should' is emphasized here since one can mathematically derive this result from some simple axioms, and it can be proved to be the optimal way to process information; see Bernardo and Smith [2009] and Section Bayesian learning and the likelihood principle. Second, one can prove that the effect of the prior vanishes asymptotically as the sample size n grows large; objectivity is attained by a **subjective consensus**: persons with wildly different priors will eventually reach the same posterior distribution as we collect more data. This result is given in chapter Classification and we will see an empirical demonstration of this effect already in the next chapter.

subjective consensus

EXERCISES

1. This is the first problem.
2. **Computer exercise.** This is the first computer exercise.

Single-parameter models

Now that we know the basics of Bayesian updating of prior beliefs with new data, we can start to analyze models with a single parameter. This will allow to practice on deriving the posterior distribution in simple settings. The drawback of simple models is that they do not show anywhere near the full potential of Bayesian methods. But you need to crawl before you can walk, and some patience is required before we come to more useful models, such as regression and classification models in later chapters.

Bernoulli data

Let us return to iid Bernoulli data:

$$x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{Bern}(\theta). \tag{10}$$

We first need a prior distribution $p(\theta)$ for θ. There are a number of ways to do **prior elicitation**, i.e. to extract a prior distribution from a person, for example an expert. Such methods involve ideas from psychology and usually consist of asking a series of questions to the expert, followed by checks for internal consistency of the elicitated prior beliefs. One can in principle elicit any distribution, e.g. in the form of a histogram, but the most common approach is to first settle on a distributional family and then elicit the hyperparameters within the family. Since $\theta \in [0,1]$, the **Beta distribution** is a suitable two-parameter family with quite a lot of flexibility; Figure 17 plots a few members of the Beta family. Note that $\text{Beta}(1,1)$ is the **uniform distribution**. We will now show that the Beta family is particularly convenient as a prior for the iid Bernoulli model.

A nice feature of Bayesian inference is that one always know where to start. To derive the posterior distribution of a parameter θ we start with Bayes' theorem (9):

$$p(\theta | x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n | \theta) p(\theta),$$

where $p(x_1, \ldots, x_n | \theta) = \theta^s (1-\theta)^f$ is the likelihood for iid Bernoulli

Beta distribution

$X \sim \text{Beta}(\alpha, \beta)$ for $X \in [0,1]$.

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

$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$$

$$\mathbb{V}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$\text{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, where $\Gamma(\alpha)$ is the Gamma function.

Figure 16: The beta distribution.

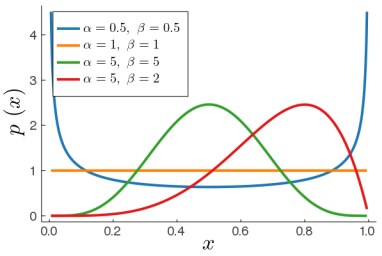


Figure 17: Some Beta distributions.

prior elicitation

Beta distribution

uniform distribution

Uniform distribution

$X \sim \text{Uniform}(a,b)$, $X \in [a,b]$.

$$p(x) = \frac{1}{b-a}$$

$$\mathbb{E}(X) = \frac{a+b}{2}$$

$$\mathbb{V}(X) = \frac{(b-a)^2}{12}$$

Figure 18: The uniform distribution.

data and $p(\theta)$ is the $\theta \sim \text{Beta}(\alpha, \beta)$ prior. So,

$$p(\theta|x_1, \ldots, x_n) \propto \theta^s(1-\theta)^f \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\text{B}(\alpha,\beta)} \tag{11}$$

$$\propto \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}, \tag{12}$$

where the second line puts the Beta function $\text{B}(\alpha, \beta)$ into the missing proportionality constant. Note that $1/\text{B}(\alpha, \beta)$ is a multiplicative constant and *not* a function of θ and will therefore not affect the shape of the posterior distribution, just scale it vertically. In the final step will recover the normalizing constant so that $p(\theta|x_1, \ldots, x_n)$ integrates to one over its support, as required. Now, from the pdf of the Beta distribution we see that the expression in (11) can be recognized as proportional to a Beta distribution. We see this as the expression is of the form $\theta^{a-1}(1-\theta)^{b-1}$ where $a = \alpha + s$ and $b = \beta + f$. The posterior for θ is therefore the $\text{Beta}(\alpha + s, \beta + f)$ distribution and the missing proportionality constant in (11) is then known to be $1/\text{B}(\alpha + s, \beta + f)$. The prior-to-posterior updating for the Bernoulli model is summarized in Figure 19. Note that the random variables in the model are written with lowercase letters for simplicity.

Conjugate analysis - Bernoulli model

Model: $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{Bern}(\theta)$
Prior: $\theta \sim \text{Beta}(\alpha, \beta)$
Posterior: $\theta|x_1, \ldots, x_n \sim \text{Beta}(\alpha + s, \beta + f)$
where $s = \sum_{i=1}^{n} x_i$ and $f = n - s$.

Figure 19: Prior-to-Posterior updating for the Bernoulli data with a Beta prior.

Using a Beta prior for the Bernoulli parameter is convenient since the posterior distribution then belongs to the *same distributional family* as the prior distribution; the posterior is also a Beta distribution. The beta family is said to be *conjugate* to the Bernoulli model, or that the beta distribution is the **conjugate prior** for the Bernoulli model. Conjugate priors are easy to use since all we have to do when updating a Beta prior with Bernoulli data is to add the number of successes s to α and the number of failures f to β. The way that α and β enter the posterior also shows that the information in a $\text{Beta}(\alpha, \beta)$ prior corresponds to a prior dataset with α successes and β failures. We usually do not have an explicit prior sample at hand, and α and β need not even be integers, but we can nevertheless think about the prior information as being equivalent to an **imaginary prior sample**.

conjugate prior

imaginary prior sample

Similar conjugate results for several other models will be presented in this book, but there are many models for which a known conjugate prior do not exist. For such models, the posterior is often

not available in closed form, but several easy-to-use approximation or simulation methods are presented in later chapters.

It is interesting to compare a Bayesian analysis of Bernoulli data with the maximum likelihood estimator $\hat{\theta}_{\text{MLE}} = s/n$. A common **Bayes estimator**, or Bayesian point estimator, is the posterior mean $\mathbb{E}(\theta|x_1,\ldots,x_n) = \frac{\alpha+s}{\alpha+\beta+n}$, which follows directly from the formula for the mean of a Beta distribution. Let us also assume a uniform prior for θ as some sort of non-informative prior, i.e. our prior is the $\text{Beta}(1,1)$ distribution. Consider the case when we have observed no successes ($s = 0$) in a small number of trials n. We then have the quite unreasonable MLE of $\hat{\theta}_{MLE} = 0$, whereas the Bayes estimator is $\mathbb{E}(\theta|x_1,\ldots,x_n) = 1/(n+2) > 0$. We will return to this example and the idea of a non-informative prior in Section Invariant priors.

Bayes estimator

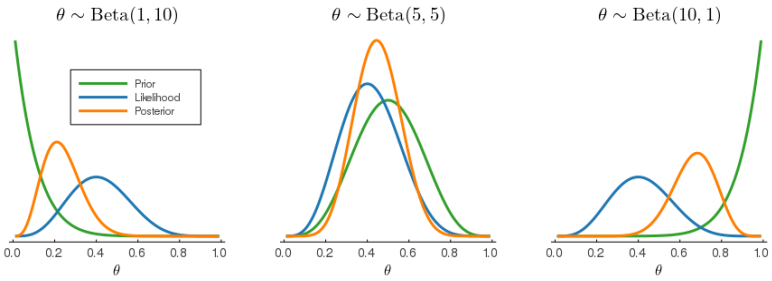


Figure 20: Bayesian analysis of $n = 10$ randomly chosen emails from the SpamBase data using three different priors. The likelihood is normalized.

EXAMPLE: SPAM EMAILS. The **SpamBase dataset** from the UCI repository[1] consists of 4601 emails that have been manually classified as *spam* (junk email) or *ham* (non-junk email). The dataset also contains a vector of covariates/features for each email, such as the number of capital letters or $-signs; this information can be used to build a spam filter that automatically separates spam from ham. We will in this chapter only analyze the proportion of spam emails without using the covariates; we return to the more interesting case with features in the Classification chapter. So, let $x_1,\ldots,x_n|\theta \overset{\text{iid}}{\sim} \text{Bern}(\theta)$ for the $n = 4601$ emails, where $x_i = 1$ if the email is spam and $x_i = 0$ for ham. The unknown quantity θ is the probability of spam.

SpamBase dataset

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml/datasets/Spambase/

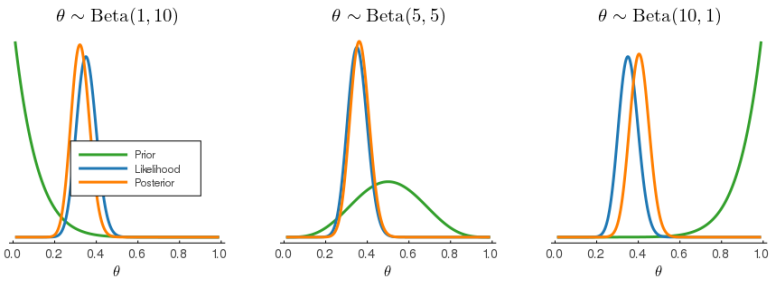


Figure 21: Bayesian analysis of $n = 100$ randomly chosen emails from the SpamBase data using three different priors. The likelihood is normalized.

To illustrate the incremental learning process in Bayesian learning

we start off by analyzing only $n = 10$ randomly sampled emails, out of which $s = 4$ were spam. Figure 20 shows the posterior distribution of θ for three persons with very different priors. With only $n = 10$ data points, the three persons' posteriors are of course very different. The results in Figure 21 are based on $n = 100$ randomly sampled emails, including the 10 emails used in Figure 20. The posteriors are now in rather close but not perfect agreement. Finally, Figure 22 shows the posterior for the full dataset with $n = 4601$; here there is a complete subjective consensus between the three persons that initially had very different beliefs about the spam probability.

From this dataset we have thus learned that around 40% or all emails are spam, and we are also quite certain about this percentage as the posterior distribution is very concentrated around 0.4. This information is not useful for building a spam filter where one instead needs the spam probability for each email to be a function of the text in that specific email (e.g. the number of $-signs). We will acheive this in chapter Classification when derive the posterior for a binary regression and use the methods in chapter Prediction and Decision making to construct Bayesian spam predictions from such a model.

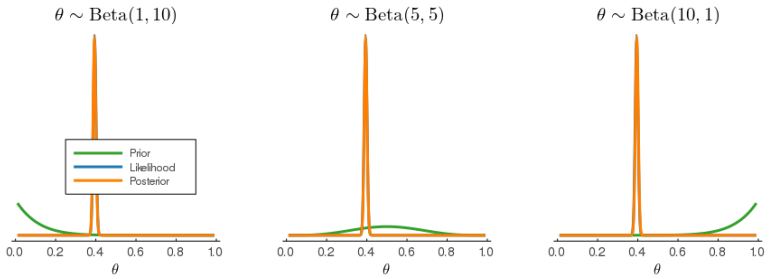


Figure 22: Bayesian analysis of all $n = 4601$ emails from the SpamBase data using three different priors. The likelihood is normalized.

Bayesian learning and the likelihood principle

We will use the Bernoulli example to demonstrate an important feature of Bayesian learning. Consider the following three experiments, all resulting in s successes in n trials:

- **Experiment 1**: sample data from $X_1, \ldots, X_n|\theta \sim \text{Bern}(\theta)$, where n is a predetermined number of trials.
  Data: the outcome in each trial: $x_1, \ldots, x_n$.
- **Experiment 2**: sample data from $X_1, \ldots, X_n|\theta \sim \text{Bern}(\theta)$, where n is a predetermined number of trials.
  Data: the total number of successes: $s = \sum_{i=1}^{n} x_i$
- **Experiment 3**: sample data from $X_i|\theta \sim \text{Bern}(\theta)$ until exactly s, a predetermined number of successes, have been obtained.
  Data: the number of trials, n, until s successes have been obtained.

The above three experiments show that we need to careful in defining exactly *which* data to use in the likelihood function. We know from before that the likelihood from Experiment 1 is

$$p(x_1, \ldots, x_n|\theta) = \theta^s(1-\theta)^{n-s}, \tag{13}$$

In the second experiment we only get to observe that there was s successes in n trials, but the exact sequence $x_1, \ldots, x_n$ is not recorded. So the data is here represented as the outcome of a random variable $S = \sum_{i=1}^{n} X_i \sim \text{Binom}(n, \theta)$. The likelihood for experiment 2 is therefore given by the binomial distribution

$$p(s) = \binom{n}{s}\theta^s(1-\theta)^{n-s}. \tag{14}$$

This is different from the likelihood in Experiment 1 since the outcome $S = s$ can be obtained from several different observed data sequences $x_1, \ldots, x_n$, each with exactly s successes. The exact number of such possible sequences is given by the binomial factor $\binom{n}{s}$.

Finally, the random variable in Experiment 3 is the number of performed trials, which follows the **negative binomial distribution**. The likelihood from Experiment 3 is therefore

negative binomial distribution

$$p(n) = \binom{n-1}{s-1}\theta^s(1-\theta)^{n-s}. \tag{15}$$

The factor $\binom{n-1}{s-1}$ counts the number of ways we can order the $s-1$ successes in the first $n-1$ trials; we know that the nth trial must have been a success since the experiment terminated after n trials. Note that there are several versions of the negative binomial distribution depending on whether we count the number of trials or the number of failures until s successes.

Now, the likelihood functions in (13)-(15) differ only by a constant that does not depend on θ, i.e. the likelihoods are proportional. The likelihood for the jth experiment can therefore be written as $c_j f(\theta)$, where $f(\theta) = \theta^s(1-\theta)^{n-s}$, $c_1 = 1$, $c_2 = \binom{n}{s}$ and $c_3 = \binom{n-1}{s-1}$. The posterior distribution of θ from the jth experiment is then by (8)

$$p_j(\theta|x_1, \ldots, x_n) = \frac{c_j f(\theta)p(\theta)}{\int c_j f(\theta)p(\theta)d\theta} = \frac{f(\theta)p(\theta)}{\int f(\theta)p(\theta)d\theta}.$$

The posterior distribution for θ is therefore the same in all three experiments. It is now obvious that Bayesian inference always satisfies the following likelihood principle.

Definition. ***Likelihood principle.*** *Two experiments that result in (proportionally) equal likelihood functions should give the same inferences.*

Likelihood principle

Informally, the likelihood principle says that all relevant information in an experiment about θ is contained in the likelihood function. The importance of the likelihood principle is that it can be mathematically derived from two simpler principles that everyone hold as self evident. Hence the word *should* in the principle; see Casella and Berger [2002, ch. 6.2] for a discussion of this famous **Birnbaum's theorem**.

Birnbaum's theorem

Many frequentist methods violate the likelihood principle. The maximum likelihood *estimate* is easily seen to be $\hat{\theta}_{\text{MLE}} = s/n$ for all three experiments for a given data set. However, the sampling variability of the maximum likelihood *estimator*, $\mathbb{V}(\hat{\theta}_{\text{MLE}})$, will be different in Experiment 3 from that in Experiment 1 and 2. This is a consequence of the estimator being S/n in Experiment 1 and 2, but s/N in Experiment 3; note the difference in random variables (capital letters) in these estimators.

In summary, Bayesian inference *conditions on the observed data* and does not rely on repeated sampling properties. The data only enters through the likelihood function and Bayesian inference respects the likelihood principle.

Gaussian data - known variance

In this section we derive the posterior distribution for the mean in the iid Gaussian model $x_1, \ldots, x_n | \theta, \sigma^2 \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$. Since this chapter is about models with a single parameter we will assume the variance σ^2 to be known; this is rarely the case in practise and we return to the Gaussian model with both parameters unknown in Chapter Multi-parameter models.

Normal distribution

$X \sim \text{N}(\mu, \sigma^2)$
Support: $X \in (-\infty, \infty)$

$$p(x) = \frac{\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)}{\sqrt{2\pi\sigma^2}}$$

$\mathbb{E}(X) = \mu$

$\mathbb{V}(X) = \sigma^2$

Figure 23: The Gaussian distribution.

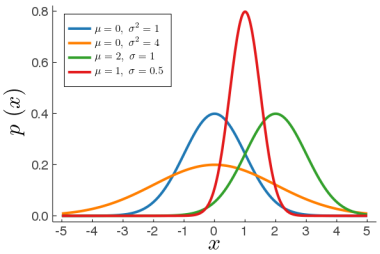


Figure 24: Some Normal distributions.

Uniform prior

We will first derive the posterior for a so called non-informative prior, i.e. a prior that is supposed to contain no, or at least very little, prior information. The most common non-informative prior for θ is a uniform distribution $p(\theta) = c$ for $\theta \in \mathbb{R}$ where $c > 0$ is a constant; the idea is that this distribution does not favor any particular value for θ. A uniform distribution over an unbounded space is not a proper distribution since $\int_{\infty}^{\infty} p(\theta) d\theta = \infty$. It is nevertheless possible to use this somewhat strange prior since the resulting posterior is proper after observing a single data point. We can also think about the uniform prior as a limiting normal distribution with a variance that tends to infinity.

By Bayes' theorem, the posterior distribution for θ under a uniform prior is

$$\begin{aligned} p(\theta | x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n | \theta) p(\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x_i - \theta)^2 \Big) \cdot c \\ &\propto \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \Big). \end{aligned}$$

Let $\bar{x}_n = \sum_{i=1}^{n} x_i$ be the sample mean, then

$$\sum_{i=1}^{n} (x_i - \theta)^2 = \sum_{i=1}^{n} \big(x_i - \bar{x} - (\theta - \bar{x}) \big)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\theta - \bar{x})^2,$$

since the cross term $2(\theta - \bar{x}) \sum_{i=1}^{n} (x_i - \bar{x}) = 0$. Note that the term $\sum_{i=1}^{n} (x_i - \bar{x})^2$ does not depend on θ and we therefore get

$$p(\theta | x_1, \ldots, x_n) \propto \exp\Big(-\frac{n}{2\sigma^2} (\theta - \bar{x})^2 \Big), \tag{16}$$

and hence that the posterior for θ can be recognized as

$$\theta | x_1, \ldots, x_n \sim \text{N}\Big(\bar{x}, \frac{\sigma^2}{n} \Big).$$

Normal prior

Consider now a normal prior, $\theta \sim N(\mu_0, \tau_0^2)$; following Gelman et al. [2013] the subscript 0 is used to denote that these are **hyperparameters** in the prior, i.e. based on 0 observations. The user must decide the most probable value for θ, μ_0, and also how sure she is by setting the prior standard deviation, τ_0. One way to elicit these prior hyperparameters is to ask the user for a 95% probability interval for θ and then back out μ_0 and τ_0; see Exercise 2.

hyperparameters

By Bayes' theorem and the rewrite of the likelihood in (16) we have

$$p(\theta|x_1,\dots,x_n) \propto \exp\Big(-\frac{n}{2\sigma^2}(\theta-\bar{x})^2\Big) \times \exp\Big(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\Big)$$

In Exercise 4 you are asked to complete the squares in this expression to prove that this expression is proportional to a normal density of the form given in Figure 25.

Conjugate analysis - Gaussian model with known variance

Model: $x_1,\dots,x_n|\theta \overset{\text{iid}}{\sim} \mathrm{N}(\theta,\sigma^2)$, σ^2 known
Prior: $\theta \sim \mathrm{N}(\mu_0,\tau_0^2)$
Posterior: $\theta|x_1,\dots,x_n \sim \mathrm{N}(\mu_n,\tau_n^2)$.
Posterior precision: $\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$
Posterior mean: $\mu_n = w\bar{x} + (1-w)\mu_0$, where $\bar{x} = \sum_{i=1}^n x_i$
Weight: $w = \frac{n/\sigma^2}{n/\sigma^2+1/\tau_0^2}$

Figure 25: Prior-to-Posterior updating for normal data with known variance and normal prior for the mean.

The normal prior is therefore conjugate to the normal model with known variance (i.e. a normal prior gives a normal posterior). The interpretation of the posterior mean μ_n and τ_n^2 in Figure 25 are quite intuitive. Note first that the expression for the posterior variance τ_n^2 is written in terms of precision = 1/variance. The first term $n/\sigma^2 = 1/(\sigma^2/n)$ is the precision in the data. This can be seen in several ways, for example by the sampling variance being $\mathbb{V}(\bar{x}) = \sigma^2/n$. Hence the formula for the posterior precision $\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$ can be read

Posterior precision = Data precision + Prior precision.

The posterior mean $\mu_n = w\bar{x} + (1-w)\mu_0$ is a weighted average of the data mean $\bar{x}$ and the prior mean. The weight w on $\bar{x}$ in Figure 25 is the data precision relative to the prior precision. The posterior therefore puts more emphasis on the data when n is large, σ small or τ_0 is large. It will not always be possible to get this clear a view of the prior-to-posterior updating in other models, but the same logic will apply also there.

Example: Internet connection speed

The maximum internet connection speed downstream in my home is 50 Mbit/sec. This maximum will typically never be reached, but my internet service provider (ISP) claims that the average speed is *at least* 20Mbit/sec. To test this, I collect a total of five measurements over the course of five consecutive using an speed testing internet service; I will call this the **Internet speed dataset**. The measurements are assumed to be $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$, where θ is the average speed; we ignore for simplicity that the measurements cannot be negative. The measurements are reported to have a standard deviation of $\sigma = 5$ by speed testing service. I will use a prior centered on the average claimed by the ISP, $\mu_0 = 20$, with a prior standard deviation of $\tau_0 = 5$. My prior beliefs are therefore that $\theta \in [10, 30]$ with approximately 95% probability.

Figure 26 displays the prior, normalized likelihood and posterior of θ based on only the first measurement $x_1 = 15.770$ Mbit/sec; the probability of interest $\Pr(\theta \geq 20 | x_1, \ldots, x_n) \approx 0.275$ is marked out by the shaded orange region. Since the prior precision happend to be equal to the data precision of a single observation, the weight on the data in the posterior mean μ_n is exactly $w = 0.5$. Figure 27 shows the results based on first two observations $x_1 = 15.770, x_2 = 20.501$. The posterior has shifted a little to the right and $\Pr(\theta \geq 20 | x_1, \ldots, x_n)$ is now 0.333. Finally, using all $n = 5$ data points with $\bar{x} = 16.001$, Figure 28 shows that we are beginning to be rather confident that the ISP's claim that $\theta \geq 20$ is false since we now have $\Pr(\theta \geq 20 | x_1, \ldots, x_n) \approx 0.051$. The weight w is now 0.833 so that data is starting to dominate the prior.

Figures 26-28 illustrate a situation where the posterior is computed by combining the prior at day 0, $\text{N}(\mu_0, \tau_0^2)$, with the likelihood for all $x_1, \ldots, x_n$ data points; hence the posterior on day n is computed as

$$p(\theta | x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n | \theta) p(\theta). \tag{17}$$

We can however equally well compute this posterior by updating yesterday's posterior $(\theta | x_1, \ldots, x_{n-1})$ with today's measurement x_n by

$$p(\theta | x_1, \ldots, x_n) \propto p(x_n | \theta) p(\theta | x_1, \ldots, x_{n-1}). \tag{18}$$

The updating in (17) and (18) give the same result, but (18) can be used sequentially in what is often called **online learning**, where "yesterday's posterior becomes today's prior". This online learning is illustrated in Figure 29 for the internet speed data.

The same online learning holds also for dependent data, e.g. time

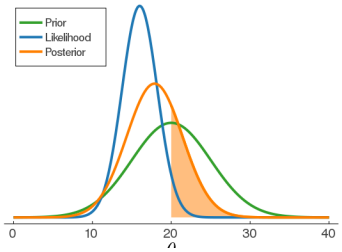


Figure 26: Internet speed data. Posterior updating based on $n = 1$ observations. The orange shaded region marks out $\Pr(\theta > 20 | x_1, \ldots, x_n)$

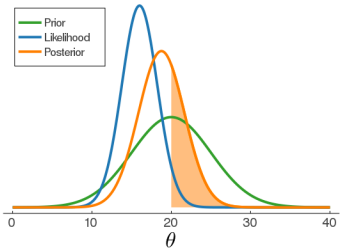


Figure 27: Internet speed data. Posterior updating based on $n = 2$ observations. The orange shaded region marks out $\Pr(\theta > 20 | x_1, \ldots, x_n)$.

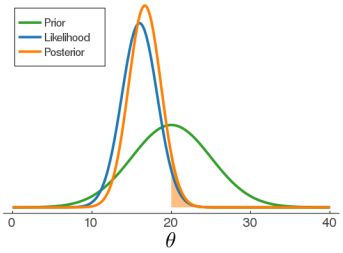


Figure 28: Internet speed data. Posterior updating based on $n = 5$ observations. The orange shaded region marks out $\Pr(\theta > 20 | x_1, \ldots, x_n)$.

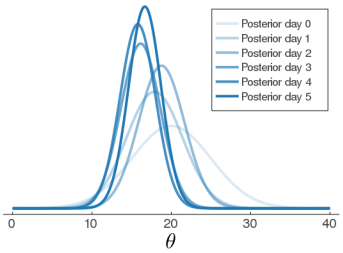


Figure 29: Internet speed data. Bayesian online learning.

online learning

series, as is easily proved as follows

$$\begin{aligned} p(\theta|x_1,\ldots,x_n) &\propto p(x_1,\ldots,x_n|\theta)p(\theta) \\ &= p(x_n|\theta,x_1,\ldots,x_{n-1})p(x_1,\ldots,x_{n-1}|\theta)p(\theta) \\ &\propto p(x_n|\theta,x_1,\ldots,x_{n-1})p(\theta|x_1,\ldots,x_{n-1}), \end{aligned} \tag{19}$$

where the second line follows from the decomposition results in Figure 30. For iid data we have the additional simplication $p(x_n|\theta,x_1,\ldots,x_{n-1}) = p(x_n|\theta)$, hence showing the equivalence of (17) and (18).

By the same proof we also see that Bayesian methods are directly applicable in **batch learning**, where the posterior can be incrementally updated using batches of several observations, since for any $1 \leq m \leq n-1$

$$p(\theta|x_1,\ldots,x_n) \propto p(x_{m+1},\ldots,x_n|\theta)p(\theta|x_1,\ldots,x_m). \tag{20}$$

batch learning

Implementing online or batch learning is straighforward for conjugate models since:

- any intermediate posterior $p(\theta|x_1,\ldots,x_m)$ belongs to the same distribution family as the original prior $p(\theta)$ and
- the prior is conjugate to the likelihood for any data, and therefore also to the likelihood of the new batch $p(x_{m+1},\ldots,x_n|\theta)$.

Decomposing distributions

For two random variables X, Y

$$p(x,y) = p(y|x)p(x)$$

For n random variables

$$p(x_1,\ldots,x_n) = p(x_1)p(x_2|x_1)\times \ldots \times p(x_n|x_1,\ldots,x_{n-1})$$

and conditional on θ

$$p(x_1,\ldots,x_n|\theta) = p(x_1|\theta)\times \ldots \times p(x_n|x_1,\ldots,x_{n-1},\theta)$$

Figure 30: Marginal-Conditional decomposition of a joint distribution.

In the case of the iid normal model with known variance we have the recursions for observation $i = 1, 2, \ldots$

$$\begin{aligned} \frac{1}{\tau_i^2} &= \frac{1}{\sigma^2} + \frac{1}{\tau_{i-1}^2} \\ w_i &= \frac{\sigma^{-2}}{\sigma^{-2} + \tau_{i-1}^{-2}} \\ \mu_i &= w_i x_i + (1-w_i)\mu_{i-1}. \end{aligned}$$

When the prior is not conjugate one has to resort to numerical methods that can be more or less computationally attractive in online mode; see in the chapters Posterior simulation and Variational inference.

Poisson data

Count data $X \in 0, 1, 2, \ldots$ is a quite frequently occuring data type in many applications; some examples are the number of software bugs, the number of lethal car accidents in a region, or the number of scooters available at a given pick-up station. The most commonly used model for count data is the **Poisson distribution**. The mean and variance of a Poisson variable are always equal, which can be

Poisson distribution

$X \sim \text{Pois}(\theta)$ for $X \in 0, 1, 2, \ldots$

$$p(x) = \frac{\theta^x e^{-\theta}}{x!}$$
$$\mathbb{E}(X) = \theta$$
$$\mathbb{V}(X) = \theta$$

Figure 31: The Poisson distribution.

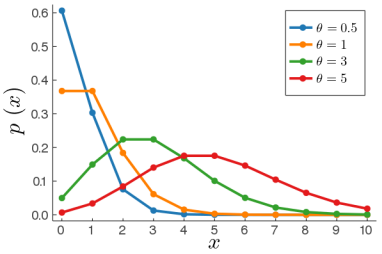


Figure 32: Some Poisson distributions.

Poisson distribution

restrictive in some application, but the model often fits many real datasets surprisingly well or can be extended to do so.

Figure 33.

The likelihood function for $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{Pois}(\theta)$, is

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \propto \theta^{\sum_{i=1}^n x_i} e^{-n\theta}. \tag{21}$$

Comparing the functional form of the likelihood in (21) with a list of common probability distributions we can see that the likelihood from iid Poisson data looks very much like a **Gamma distribution** in θ. Even more, the form of the Gamma distribution tells us that a Gamma prior may indeed combine nicely with this likelihood. So let us try if $\theta \sim \text{Gamma}(\alpha, \beta)$ is conjugate to the iid Poisson model:

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n|\theta) p(\theta) \\ &\propto \theta^{\sum_{i=1}^n x_i} e^{-n\theta} \cdot \theta^{\alpha-1} e^{-\beta\theta} \\ &= \theta^{\alpha+\sum_{i=1}^n x_i - 1} e^{-(\beta+n)\theta}, \end{aligned}$$

where we have directly written up the $\text{Gamma}(\alpha, \beta)$ prior without normalization constant. This expression is indeed proportional to a Gamma distribution and we have the following result:

Conjugate analysis - Poisson model

Model: $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{Pois}(\theta)$
Prior: $\theta \sim \text{Gamma}(\alpha, \beta)$
Posterior: $\theta|x_1, \ldots, x_n \sim \text{Gamma}(\alpha + \sum_{i=1}^n x_i, \beta + n)$

EXAMPLE: INTERNET AUCTION DATA. The **eBayCoin dataset** collected by Wegmann and Villani [2011] and made available in the UCI repository[2] consist of data from 1000 eBay auctions of collectors coins. For each auction, the dataset records the final price of the auctioned coin, the number of bidder in the auction and a number of covariates such as the quality of the sold coin, the lowest price that the seller would agree to sell for etc. We will here analyze the number of bidders using an iid Poisson model without covariates. We return to this dataset in Chapter Classification where we make use of the covariates in a Poisson regression model for predicting the number of bidders.

To compute the posterior distribution for θ, the average number of bidders in an auction we need the summary statistic $\sum_{i=1}^n x_i = 3635$. The sample mean in the $n = 1000$ auctions is therefore $\bar{x} = 3.635$

Gamma distribution

$X \sim \text{Gamma}(\alpha, \beta)$ for $X > 0$.

$$\begin{aligned} p(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \\ \mathbb{E}(X) &= \frac{\alpha}{\beta} \\ \mathbb{V}(X) &= \frac{\alpha}{\beta^2} \end{aligned}$$

Figure 33: Gamma distribution.

Gamma distribution

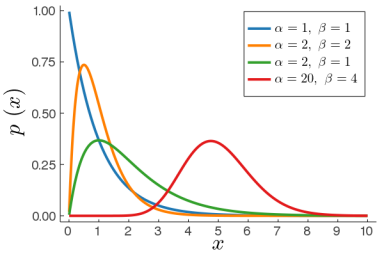


Figure 34: Some Gamma distributions.

Figure 35: Prior-to-Posterior updating for the Poisson data with a Gamma prior.

eBayCoin dataset

[2] http://archive.ics.uci.edu/ml/datasets/eBayCoin/

bidders per auction. I will use the gamma prior with $\alpha = 2$ and $\beta = 1/2$ since this implies a prior mean of $\mathbb{E}(\theta) = 4$ and prior standard deviation of $\mathbb{S}(\theta) = 2.283$, which I find matches quite well with my prior beliefs. This prior and the posterior updated with data from $n = 1000$ auctions are shown in Figure 36. Note the different scales on the horizontal axis. We are now more or less certain that the average number of bidders is in the interval $\theta \in [3.4, 3.9]$.

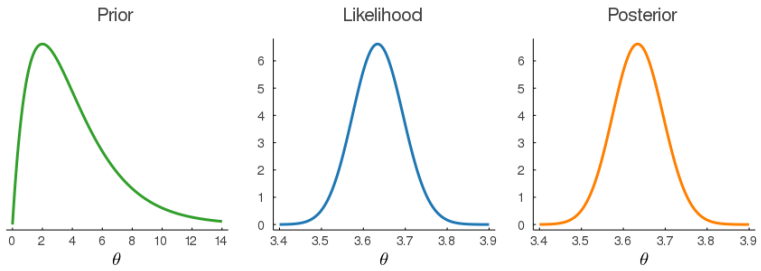


Figure 36: Bayesian analysis of the numbers of bidders in $n = 1000$ eBay coin auctions.

Figure 37 a) plots the fitted Poisson distribution with θ set equal to the posterior mean against the observed data. It is obvious that the Poisson distribution is too restrictive as the fit is terrible.

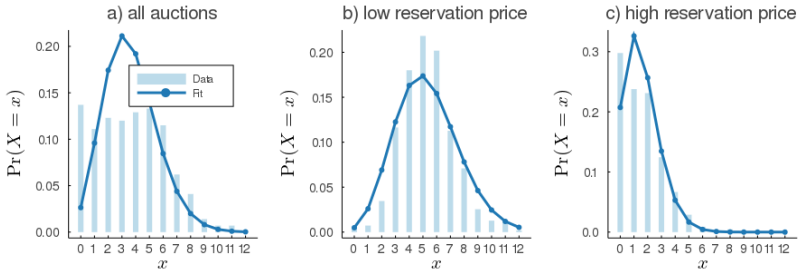


Figure 37: Assessing the fit of the Poisson model with the posterior mean estimate of θ.

The poor fit can be attributed to the heterogeneity of the auctions. For example, some of the auctions had a high so called reservation price, i.e. the lowest price that the seller is willing sell for, while other auctions had a very low reservation price. It is expected that a high reservation price discourages bidders from entering the auction.

To explore the effect of the reservation price we split the data into low and high reservation price auctions, and analyze the two auction types separately. The prior-to-posterior updating is shown in Figure 38; the priors now reflect that θ is likely to be larger for the auctions with low reservation prices. The posteriors are clearly different in the two subpopulations. The Poisson model fits better on the two subpopulations as shown in Figure 37 b) and c), but it is not perfect. We will return to this dataset in Chapter Regression using a Poisson regression with the reservation price as covariate as well as other auction specific covariates.

We have now seen that:

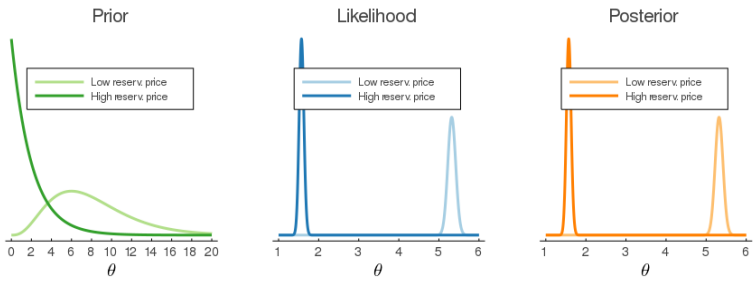


Figure 38: eBay auctions. Bayesian analysis of the numbers of bidders in $n = 550$ auctions with a low reservation price and $n = 450$ auctions with a high reservation price.

- the beta prior is conjugate to the Bernoulli likelihood
- the normal prior is conjugate to the normal likelihood
- the gamma prior is conjugate to the Poisson likelihood.

Here is a formal definition of a conjugate prior.

Definition (Conjugate prior). *A family of prior distributions $\mathcal{P}$ is **conjugate** to a family of likelihoods $\mathcal{L} = \{p(\mathbf{x}|\theta), \theta \in \Theta\}$ if*

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|\mathbf{x}) \in \mathcal{P} \qquad \text{for all } p(\mathbf{x}|\theta) \in \mathcal{L}.$$

Summarizing a posterior distribution

The posterior distribution for models with a single parameter are easily plotted and gives a complete visual quantification of uncertainty. Starting from the next chapter, our models will typically contain more than one parameter, and not seldom quite many. It is then impractical to plot the whole posterior distribution and we will now explore some commonly used numerical summaries of the posterior, for example a point estimate and posterior probability intervals.

A point estimate of θ summarizes the posterior with a single point. The three most commonly used Bayesian point estimates are:

- The posterior mean $\hat{\theta}_{\text{mean}} \equiv \mathbb{E}(\theta|x_1, \ldots, x_n)$.
- The posterior median $\hat{\theta}_{\text{med}}$, i.e. the 50th quantile of $p(\theta|x_1, \ldots, x_n)$.
- The posterior mode $\hat{\theta}_{\text{mode}} \equiv \arg\max_{\theta \in \Theta} p(\theta|x_1, \ldots, x_n)$.

We will see in chapter Prediction and Decision making that the choice of point estimate can be formalized as a decision problem.

A point estimate says nothing about the variability in the posterior. One way to quantify the uncertainty is the posterior standard deviation $\mathbb{S}(\theta|x_1, \ldots, x_n) = \sqrt{\mathbb{V}(\theta|x_1, \ldots, x_n)}$.

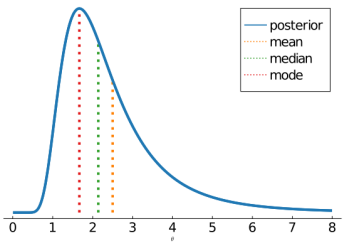


Figure 39: Three common point estimates for summarizing a posterior.

EXAMPLE: INTERNET AUCTION DATA. As we saw earlier the posterior for the mean θ of a Poisson distribution with a $\theta \sim \text{Gamma}(\alpha, \beta)$

prior is $\theta|x_1,\ldots,x_n \sim \text{Gamma}(\alpha+\sum_{i=1}^n x_i, \beta+n)$. From properties of the Gamma distribution, the posterior mean estimate is hence $(\alpha+\sum_{i=1}^n x_i)/(\beta+n)$ and the posterior variance is $(\alpha+\sum_{i=1}^n x_i)/(\beta+n)^2$. For the eBay data Poisson data we have $\mathbb{E}(\theta|x_1,\ldots,x_n) = \frac{2+3635}{0.5+1000} \approx 3.635$ bidders and $\mathbb{S}(\theta|x_1,\ldots,x_n) = \sqrt{\frac{2+3635}{(0.5+1000)^2}} \approx 0.060$.

Before presenting how to summarize a posterior by an interval, let us first informally recall the definition of a frequentist confidence interval. A 95% *confidence interval* for a parameter θ is a random interval $[l(X_1,\ldots,X_n), u(X_1,\ldots,X_n)]$ that contains the true θ in 95% of *all possible datasets* $X_1,\ldots,X_n$ from the data generating process. As usual with frequentist methods we are guaranteed a long run performance over all possible datasets, but the realized interval $[l(x_1,\ldots,x_n), u(x_1,\ldots,x_n)]$ either does or does not cover the true θ.

A Bayesian interval is defined in a much more direct way, and is conditional on the actually observed dataset. This simpler definition is possible since the posterior is a probability distribution; we have broken the Bayesian eggs and can enjoy the omelet. A 95% posterior **credibility interval** for $\theta \in \Theta \subset \mathbb{R}$ is an interval $[l,u] \subset \Theta$ such that $\Pr(\theta \in [l,u] \,|x_1,\ldots,x_n) = 0.95$, i.e. an interval that contains 95% of the posterior probability mass. We can generalize this to a more general region than an interval, for example a union of disjoint intervals, and of course to other probability coverages than 95%.

credibility interval

There are many ways to construct an credibility interval with a certain coverage probability. An **equal tail credibility interval** is an interval that cuts off equal probability in the left and right tail; for example, a 95% interval sets l and u to the 2.5% and 97.5% posterior quantile, respectively. Another popular interval construction is the highest posterior density (HPD) region which, as the name suggest, is made up of the θ values with highest posterior density. We use the word *region* instead of interval here since HPD regions need not be intervals. Here is the definition.

equal tail credibility interval

Definition (HPD region). *A **Highest Posterior Density (HPD) region** for $\theta \in \Theta$ with coverage probability γ is a region $R \subset \Theta$ such that:*

Highest Posterior Density (HPD) region

- $\Pr(\theta \in R|x_1,\ldots,x_n) = \gamma$ *and*
- $p(\theta_{\text{in}}|x_1,\ldots,x_n) \geq p(\theta_{\text{out}}|x_1,\ldots,x_n)$ *for all* $\theta_{\text{in}} \in R$ *and* $\theta_{\text{out}} \notin R$.

Figure 40 illustrates the difference between equal tail intervals and HPD regions for some example densities. Note how the equal tail interval construction can exclude θ values that actually have highest posterior density (middle graph) and how HPD regions can be disconnected (righthand graph).

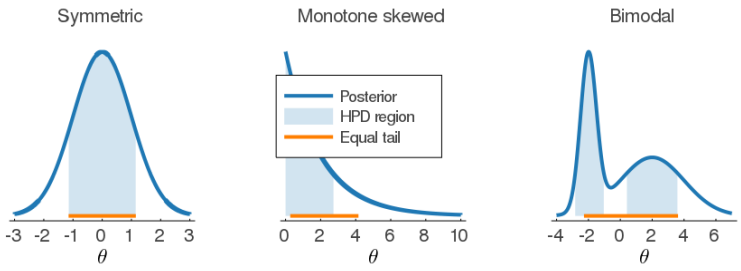


Figure 40: HPD regions vs equal tail intervals.

A disadvantage of HPD regions is that they are not invariant to reparametrization; the HPD for a transformed parameter $\eta(\theta)$ is typically not a direct map of the HPD for θ.

EXAMPLE: INTERNET AUCTION DATA The 95% equal tail interval for the mean number of bidders in the iid Poisson model is $[3.518, 3.754]$ which is virtually indistinguishable from the HPD interval $[3.517, 3.754]$ since the posterior is essentially symmetric, see Figure 41.

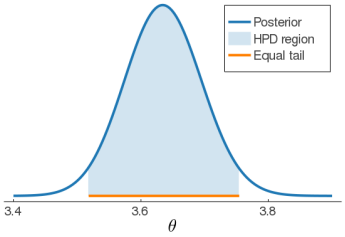


Figure 41: 95% credibility intervals for the Gamma posterior in the eBay auction data.

*Exponential Family and Sufficiency**

This section presents the concept of sufficient statistics and the exponential family of distributions, with particular emphasis on their role in Bayesian learning. While these concepts are very important in statistics, this starred section can be skipped at first reading, but should be read before the generalized linear models in Chapter Classification, where the exponential family plays a prominent role.

Sufficient statistics

In all models covered so far in this book, the dataset, $(x_1, \ldots, x_n)$, has only entered the likelihood through some low-dimensional summary statistic; for example the number of successes $s = \sum_{i=1}^n x_i$ in the Bernoulli model, the sample mean $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ in the Gaussian model, and the sum of counts, $\sum_{i=1}^n x_i$, in the Poisson model. Note that we did not choose this data reduction, it just turned out that the likelihood only depended on the summarizing statistic; the statistic captured all the relevant information in the sample. In all of the above examples, the statistic was one-dimensional. In other models more than a single dimension is needed to compress the dataset, and we let the vector-valued function $\mathbf{t}(x_1, \ldots, x_n) \to \mathbb{R}^k$ denote the statistic in general, where k is the dimension of reduction.

The following definition captures the idea that a statistic may contain *all* relevant information in the data about a parameter θ.

Definition. ***Sufficient statistic.*** *A statistic* $\mathbf{t}(X_1, \ldots, X_n)$ *is sufficient for* θ *if the conditional distribution of the sample* $X_1, \ldots, X_n$ *given the value of the statistic* $\mathbf{t}(X_1, \ldots, X_n)$ *does not depend on* θ.

Sufficient statistic

The sufficiency of a statistic can be checked by the following lemma; see Casella and Berger [2002] for a proof.

Lemma 1. ***Factorization criterion.*** *A statistic* $t(x_1, \ldots, x_n)$ *is sufficient for a parameter* θ *if and only if the likelihood can be factorized as*

Factorization criterion

$$p(x_1, \ldots, x_n|\theta) = h(x_1, \ldots, x_n) f(\mathbf{t}(x_1, \ldots, x_n); \theta), \tag{22}$$

where $h(x_1, \ldots, x_n)$ *does not depend on* θ *and* $f(\mathbf{t}; \theta)$ *is a function of the data only through the sufficient statistic* $\mathbf{t}(x_1, \ldots, x_n)$.

The idea behind sufficient statistics is so appealing that it is often formulated as a desired inference principle similar to the likelihood principle presented in the section Bayesian learning and the likelihood principle.

Definition. ***Sufficiency principle.*** *If* $\mathbf{t}(X_1, \ldots, X_n)$ *is a sufficient statistic for* θ *then any inference about* θ *should depend on the sample* $x_1, \ldots, x_n$ *only through the value* $\mathbf{t}(x_1, \ldots, x_n)$.

Sufficiency principle

Theorem 1. *Bayesian learning satisfies the sufficiency principle.*

Proof. If $\mathbf{t}(x_1, \ldots, x_n)$ is a sufficient statistic for θ then by Lemma 1

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &= \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{\int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta} \\ &= \frac{h(x_1, \ldots, x_n) f(\mathbf{t}(x_1, \ldots, x_n); \theta)p(\theta)}{\int h(x_1, \ldots, x_n) f(\mathbf{t}(x_1, \ldots, x_n); \theta)p(\theta)d\theta} \\ &= \frac{f(\mathbf{t}(x_1, \ldots, x_n); \theta)p(\theta)}{\int f(\mathbf{t}(x_1, \ldots, x_n); \theta)p(\theta)d\theta}, \end{aligned}$$

which only depends on the data through the sufficient statistic $\mathbf{t}(x_1, \ldots, x_n)$. □

Exponential family

All models considered so far are part of the large and important exponential family of distributions. A random variable X follows a distribution in the (one-parameter) **exponential family** if its density can be written in the form

exponential family

$$p(x|\theta) = h(x) \exp\Big(\eta(\theta)t(x) - A(\theta)\Big), \text{ for } x \in \mathcal{X}, \tag{23}$$

where $h(x)$ is a function of only x and $A(\theta)$ is a function of only θ. The support $\mathcal{X}$ is not allowed to depend on θ, so that for example the Uniform$(0, \theta)$ distribution does not belong to the exponential family. The function $\eta(\theta)$ is called the **natural parameter** and is an invertible transformation of the parameter θ. Here are some examples.

natural parameter

EXAMPLE: POISSON DISTRIBUTION. The Pois(θ) distribution can be rewritten as follows

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} = \frac{e^{x \ln \theta} e^{-\theta}}{x!} = \frac{1}{x!} \exp(x \ln \theta - \theta),$$

which is in the exponential family with $h(x) = (x!)^{-1}$, $A(\theta) = \theta$, $\eta(\theta) = \ln \theta$ and $t(x) = x$. Note in particular that the natural parameter is the logarithm of the Poisson mean, $\eta(\theta) = \ln \theta$.

EXAMPLE: BERNOULLI DISTRIBUTION. The Bern(θ) distribution can also be written as an exponential family:

$$p(x|\theta) = \theta^x (1-\theta)^{1-x} = \left(\frac{\theta}{1-\theta}\right)^x (1-\theta) = \exp\Big(\eta(\theta) x - A(\theta)\Big),$$

where $\eta(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$, $A(\theta) = \ln\left(\frac{1}{1-\theta}\right)$, $t(x) = x$ and $h(x) = 1$. The natural parameter for the Bernoulli distribution is therefore the log-odds, $\ln\left(\frac{\theta}{1-\theta}\right)$.

The normal distribution and many other distributions can similarly be shown to belong to the exponential family; but not all do, for example the **student-t distribution**. We will use ExpFam(θ) as a generic notation for a distribution in the exponential family, leaving the specific $h(x)$, $A(\theta)$, $\eta(\theta)$ and $t(x)$ functions implicit.

The likelihood function for iid data from an ExpFam(θ) distribution is

$$p(x_1, ..., x_n|\theta) = \left[\prod_{i=1}^n h(x_i)\right] \exp\left(\eta(\theta) \sum_{i=1}^n t(x_i) - nA(\theta)\right). \tag{24}$$

Lemma 1 can be direcly used to show that $\sum_{i=1}^n t(x_i)$ is a sufficient statistic for θ. In the next chapter well will see a multiparameter version of the exponential family with a vector of k sufficient statistics. The Pitman–Koopman–Darmois theorem [Bernardo and Smith, 2009] proves that among distributions whose support does not depend on θ, only the exponential family have sufficient statistics of fixed dimension, i.e the dimension k does not depend on the size of the data, n (or at least is bounded).

The exponential family has several other attractive properties [Sundberg, 2019]. One property of particular interest here is that a conjugate prior always exists for models in the exponential family. In fact, the following family of priors is conjugate to the exponential family likelihood in (24)

$$p(\theta) = H(\tau_0, \nu_0) \exp\Big(\eta(\theta)\tau_0 - \nu_0 A(\theta)\Big), \tag{25}$$

where $H(\tau_0, \nu_0)$ is the normalizing constant. Note that this prior has two hyperparameter τ_0 and ν_0 that needs to be set by the user.

student-t distribution

Student-t distribution

$X \sim \mathrm{t}(\mu, \sigma, \nu)$ for $X \in (-\infty, \infty)$

$$p(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu\sigma^2}} \times \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

$$\mathbb{E}(X) = \mu \text{ if } \nu > 1$$

$$\mathbb{V}(X) = \sigma^2 \frac{\nu}{\nu - 2} \text{ if } \nu > 2$$

Figure 42: The student-t distributions.

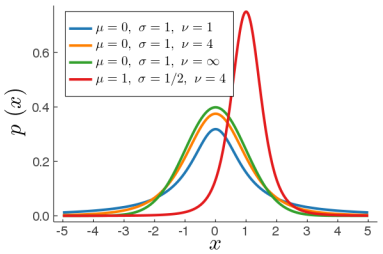


Figure 43: Some Student-t distributions.

We will use the symbol $\theta \sim \text{ExpFamConj}(\tau_0, \nu_0)$ for this prior distribution, where it must be remembered that the form of the prior depends on which specific exponential family member the prior is conjugate to, i.e. it depends on $\eta(\theta)$ and $A(\theta)$.

EXAMPLE: BERNOULLI MODEL. It was shown above that $\eta(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$ and $A(\theta) = \ln\left(\frac{1}{1-\theta}\right)$, for Bernoulli data. The prior in (25) is therefore

$$\begin{aligned} p(\theta) &\propto \exp\Big(\eta(\theta)\tau_0 - \nu_0 A(\theta)\Big) \\ &= \exp\Big(\ln\big(\frac{\theta}{1-\theta}\big)\tau_0 - \nu_0 \ln\big(\frac{1}{1-\theta}\big)\Big) \\ &\propto \theta^{\tau_0}(1-\theta)^{\nu_0-\tau_0}, \end{aligned}$$

which is proportional to the $\text{Beta}(\tau_0, \nu_0 - \tau_0)$ distribution. The parametrization in (25) is hence interpreted as the information from a (imaginary) prior sample of τ_0 success in ν_0 trials. The $\text{Beta}(\alpha, \beta)$ prior from before expresses instead the prior information as a sample of α success and β failures.

Conjugate analysis from iid exponential family data

Model: $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{ExpFam}(\theta)$
Prior: $\theta \sim \text{ExpFamConj}(\tau_0, \nu_0)$
Posterior: $\theta | x_1, \ldots, x_n \sim \text{ExpFamConj}\big(\tau_0 + \sum_{i=1}^n t(x_i), \nu_0 + n\big)$

Figure 44: Prior-to-Posterior updating for iid exponential family data with a conjugate prior.

The posterior distribution for θ in the exponential family with a conjugate prior is obtained by multplying the likelihood in (24) with prior (25)

$$p(\theta | x_1, ..., x_n) \propto \exp\left[\eta(\theta)\Big(\tau_0 + \sum_{i=1}^n t(x_i)\Big) - (\nu_0 + n)A(\theta)\right],$$

which is of the form ExpFamConj, but with updated hyperparameters: $\tau_0 \Rightarrow \tau_0 + \sum_{i=1}^n t(x_i)$ and $\nu_0 \Rightarrow \nu_0 + n$. We summarize this in Figure 44.

This result shows that we can think quite generally about ν_0 as the (imaginary) prior sample size and τ_0 as the prior data compressed by the sufficient statistic. For example, in the Poisson model the information in the conjugate prior equals a prior sample of ν_0 data points with a mean count of τ_0/ν_0.

EXERCISES

1. Let $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{Expon}(\theta)$ be exponentially distributed data. Show that the Gamma distribution is the conjugate prior for this model.
2. I determined my normal prior in the internet speed data example by specifying the prior mean θ_0 and standard deviation τ_0. Assume that another person instead specified a 95% prior probability interval for θ as $[20, 30]$. Use this information to determine that persons normal prior, i.e. compute θ_0 and τ_0 for this person.
3. (a) Let $x_1, ..., x_{10}$ be a sample with $\bar{x} = 1.873$. Assume the model $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} N(\theta, 1)$ and the prior $\theta \sim N(0, 5)$. Compute the posterior distribution of θ.

   (b) You now get hold of a second sample $y_1, \ldots, y_{10}|\theta \overset{\text{iid}}{\sim} N(\theta, 2)$, where θ is the same quantity as in (a) but the measurements have a larger variance. The sample mean in this second sample is $\bar{y} = 0.582$. Compute the posterior distribution of θ using both samples (the x's and the y's) under the assumption that the two samples are independent.

   (c) You finally obtain a third sample $z_1, \ldots, z_{10}|\theta \overset{\text{iid}}{\sim} N(\theta, 3)$, with mean $\bar{z} = 1.221$. Unfortunately, the measuring device for this latter sample was defective and any measurement above 3 was recorded as exactly 3. There were two such measurements. Give an expression for the unnormalized posterior distribution (likelihood x prior) for θ based on all three samples (x, y and z). If you have computer available you may plot this unnormalized posterior over a grid of θ values. *Hint: the posterior distribution is not normal anymore when the measurements are truncated at* 3.
4. Derive the posterior distribution for the normal model with a normal prior in Figure 25. *Hint: complete the square.*
5. (a) Let $x_1, \ldots, x_n|\theta \sim \text{Uniform}(\theta - 1/2, \theta + 1/2)$. Let $\hat{\theta} = \bar{x}$ be an estimator of θ. Derive an expression for the sampling variance of $\hat{\theta}$.

   (b) Derive the posterior distribution for θ assuming a uniform prior distribution. *Hint: once you have observed some data, some values for θ are no longer possible.*

   (c) Assume that you have observed three data observations: $x_1 = 1.1, x_2 = 2.09, x_3 = 1.4$. What would a frequentist conclude about θ? What would a Bayesian conclude? Discuss.
6. Show that the $N(\mu, 1)$ distribution belongs to the exponential family.

Exponential distribution

$X \sim \text{Expon}(\theta)$ for $X \in (0, \infty)$

$$p(x) = \theta e^{-\theta x}$$
$$\mathbb{E}(X) = 1/\theta$$
$$\mathbb{V}(X) = 1/\theta^2$$

Figure 45: The exponential distribution.

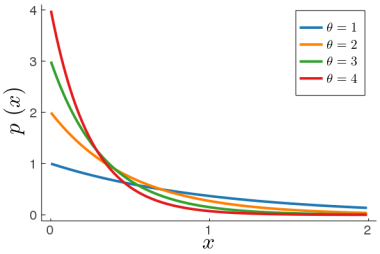


Figure 46: Some Exponential distributions.

NOTEBOOKS

1. See the notebook single-parameter models.

Multi-parameter models

Joint posterior distributions

Most models have more than one parameter, and many models are incredibly rich on parameters. Datasets are increasingly rapidly in size and makes it possible to estimate increasingly more complex models. To explore how Bayesian methods can be used in multiparameter models we first return in this chapter to the iid $N(\theta, \sigma^2)$, but now in the more realistic setting where both θ and σ^2 are unknown parameters. In later chapters we will tackle regression and classification models where each covariate (input) x_k affects the response (output) y through a regression coefficient β_k; hence in a regression with K covariates we have K regression coefficients $\beta_1, \ldots, \beta_K$.

Consider a general probability model $p(x_1, \ldots, x_n|\theta_1, \ldots, \theta_K)$ with K parameters for a dataset $x_1, \ldots, x_n$; for example the iid normal model where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. Bayesian learning proceeds exactly as with a single parameter, except that the prior and posterior distribution are now both multidimensional joint distributions. Figure 47 gives an illustration of a bivariate ($K = 2$) normal distribution.

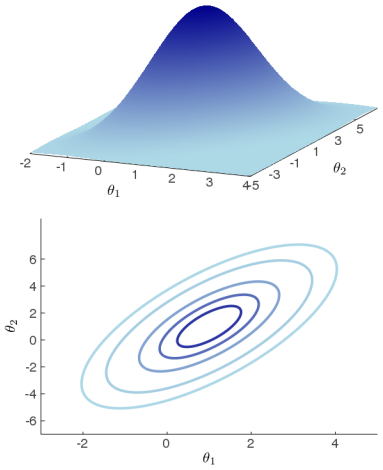


Figure 47: Surface and contour plot of the bivariate normal distribution. The contour levels contain 25, 50, 75, 95 and 99% of the probability mass, respectively.

Using Bayes' theorem in proportional form, the **joint posterior distribution** $p(\theta_1, \ldots, \theta_K|x_1, \ldots, x_n)$ is given by

joint posterior distribution

$$p(\theta_1, \ldots, \theta_K|x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n|\theta_1, \ldots, \theta_K)p(\theta_1, \ldots, \theta_K),$$

where $p(\theta_1, \ldots, \theta_K)$ is a multidimensional prior distribution and $p(x_1, \ldots, x_n|\theta_1, \ldots, \theta_K)$ is the likelihood function; Note that the likelihood function is now a **likelihood surface** in the sense that it is a function of several parameters, $\theta_1, \ldots, \theta_K$.

likelihood surface

To keep the notation simpler we often use vector notation and write $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_n)$ and $\mathbf{x} \equiv (x_1, \ldots, x_n)$. The multivariate Bayes' theorem can then be expressed as

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{26}$$

Marginalization

The joint posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ contains all posterior information about $\boldsymbol{\theta}$, but is obviously hard to visualize in the same way as we did for single-parameter models. In many cases we are also most interested in a subset of parameters, and the other parameters are only needed to model the data well but are of no real interest. Such parameters are just a nuisance when presenting inferences and are therefore often called **nuisance parameters**. Getting rid of nuisance parameters is very difficult in a non-Bayesian setting, for example when using maximum likelihood estimation. So what is the Bayesian solution to this dilemma?

nuisance parameters

Nuisance parameters can be handled in a very natural way in a Bayesian approach since the posterior distribution is a probability distribution for $\boldsymbol{\theta}$. We can therefore just integrate out, or marginalize out, the nuisance parameters just as in ordinary probability calculus. Take a simple example where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and assume that the parameter of interest is θ_1 whereas θ_2 is considered a nuisance parameter; θ_1 could for example be the mean of iid Gaussian model and θ_2 the variance. The marginal posterior of θ_1 is then

$$p(\theta_1) = \int p(\theta_1, \theta_2) d\theta_2,$$

where the integration is over the full support of θ_2. Figure 48 illustrates the marginalization concept. Using the decomposition $p(\theta_1, \theta_2) = p(\theta_1|\theta_2)p(\theta_2)$ we can alternatively express this as

$$p(\theta_1) = \int p(\theta_1|\theta_2)p(\theta_2) d\theta_2,$$

which shows that marginalization is acheived by averaging over the values of θ_2 with weights given by $p(\theta_2)$.

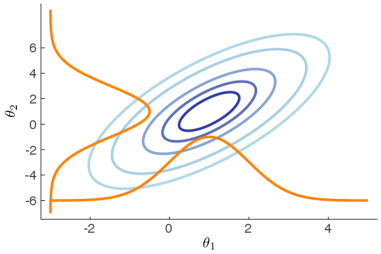


Figure 48: Contour plot of the bivariate normal distribution in Figure 47 along with the marginal distibutions.

More generally, with more than two parameters, partition the elements of $\boldsymbol{\theta}$ into two vectors, $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$. The marginal posterior of $\boldsymbol{\theta}_a$ is the obtained by marginalizing out $\boldsymbol{\theta}_b$ from the joint posterior

$$p(\boldsymbol{\theta}_a) = \int \cdots \int p(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b) d\boldsymbol{\theta}_b. \tag{27}$$

We will see examples of marginalization in the following sections.

Gaussian data with unknown variance

The previous chapter analyzed iid normal data $x_1, \ldots, x_n|\theta, \sigma^2 \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$ under the usually unrealistic assumption that σ^2 is known. Let us now tackle the case where both parameters are unknown. It

turns our that the conjugate prior for this model has dependence between θ and σ, so we will describe the prior using the decomposition $p(\theta|\sigma^2)p(\sigma^2)$ as follows

$$\theta|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0) \tag{28}$$
$$\sigma^2 \sim \text{Inv}{-}\chi^2(\nu_0, \sigma_0^2). \tag{29}$$

The marginal conjugate prior for σ^2 involves a new distribution, the **scaled inverse chi-squared distribution**, denoted by $\text{Inv}{-}\chi^2(\nu_0, \sigma_0^2)$; see Figure 49. This distribution is a specific parametrization of the **inverse Gamma distribution**. The name comes from the characterization

scaled inverse chi-squared distribution

inverse Gamma distribution

$$X \sim \chi_\nu \quad \Rightarrow \quad Y = \nu\tau^2 \frac{1}{X} \sim \text{Inv}{-}\chi^2(\nu, \tau^2),$$

so that a $\text{Inv}{-}\chi^2(\nu, \tau^2)$ variable really is an inverted χ^2_ν variable scaled by $\nu\tau^2$. Note that the parameter τ^2 is close to the mean when ν is large. The mode is $\nu\tau^2/(\nu+2)$, so τ^2 is somewhere between the mode and the mean. We will therefore call τ^2 the location of $\text{Inv}{-}\chi^2(\nu, \tau^2)$, or sometimes just sloppily as "our best guess".

The conjugate prior in (28) is specified via the four prior hyperparameters:

- μ_0 - the prior mean for θ
- κ_0 - the number of prior data observations for θ
- σ_0^2 - the prior location of σ^2
- ν_0 - the prior degrees of freedom for σ^2.

Note that, similar to the conjugate prior for the exponential family, we are only *interpreting* κ_0 as the number of prior observations. The prior may not actually be based on previous data, but the information in the prior $\theta|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0)$ has the equivalent strength of an imaginary prior sample of κ_0 observations from a data generating process with variance σ^2.

Figure 50 shows that the posterior is indeed in the same form as the prior in (28), as required for a conjugate prior. There is a lot of greek letters in Figure 50, but note that the same sort of intuition applies here as in the case with a known variance in Chapter Single-parameter models:

- the posterior mean μ_n is a weighted average of the data mean $\bar{x}$ and the prior mean μ_0
- the weight on the data $w = n/(\kappa_0 + n)$ is close to one when either the data is informative (large n) or the prior is weak (small κ_0)

Inv-χ^2 distribution

$X \sim \text{Inv}{-}\chi^2(\nu, \tau^2)$, $X \in (0, \infty)$

$$p(x) = \frac{(\tau^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \frac{\exp\left(\frac{-\nu\tau^2}{2x}\right)}{x^{1+\nu/2}}$$
$$\mathbb{E}(X) = \frac{\nu}{\nu-2}\tau^2$$
$$\mathbb{V}(X) = \frac{2\nu^2\tau^4}{(\nu-2)^2(\nu-4)}$$

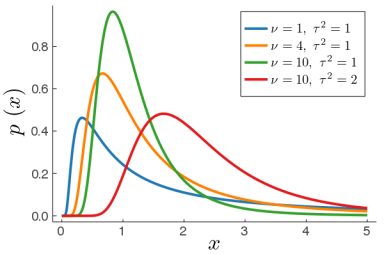


Figure 49: Some Scaled-Inv-Gamma distributions.

Gaussian iid data with conjugate prior

Model: $x_1, \ldots, x_n | \theta, \sigma^2 \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$
Prior: $\theta | \sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0)$
$\sigma^2 \sim \text{Inv}-\chi^2(\nu_0, \sigma_0^2)$
Posterior: $\theta | \sigma^2, x_1, \ldots, x_n \sim N(\mu_n, \sigma^2/\kappa_n)$
$\sigma^2 | x_1, \ldots, x_n \sim \text{Inv}-\chi^2(\nu_n, \sigma_n^2)$
$\mu_n = w\bar{x} + (1-w)\mu_0$
$w = \frac{n}{\kappa_0+n}$
$\kappa_n = \kappa_0 + n$
$\nu_n = \nu_0 + n$
$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{x}-\mu_0)^2$
where $\bar{x} = \sum_{i=1}^n x_i$ and $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$

Figure 50: Prior-to-Posterior updating for the iid Gaussian model with unknown mean and variance using the conjugate prior.

- the reason why σ^2 does not appear in w is that the prior variance for θ is scaled by σ^2 in the conjugate prior, and σ^2 therefore cancels out in w.
- the posterior sample size κ_n is the number of prior observations κ_0 plus the sample size n.

Interest centers mainly on the average download speed, so we would like to obtain the marginal posterior distribution of θ. This distribution can be derived by marginalizing out the nuisance parameter σ^2 from the joint posterior

$$p(\theta | x_1, \ldots, x_n) = \int p(\theta | \sigma^2, x_1, \ldots, x_n) p(\sigma^2 | x_1, \ldots, x_n) d\sigma^2,$$

where $p(\theta | \sigma^2, x_1, \ldots, x_n)$ and $p(\sigma^2 | x_1, \ldots, x_n)$ are given in Figure 50. In Exercise 1 you are ask to show that the marginal posterior of θ is a student-t distribution; see Figure 42 and 43 for a definition and properties. Specifically, we have the following result

$$\theta | x_1, \ldots, x_n \sim t(\mu_n, \sigma_n^2/\kappa_n, \nu_n), \tag{30}$$

where μ_n, σ_n^2, κ_n and ν_n are all defined as in Figure 50. Note that also the marginal prior for θ follows a student-t distribution of the form (30), but with hyperparameters naturally subscripted by 0 instead of n.

EXAMPLE: INTERNET SPEED DATA. Let us return to the example with the $n = 5$ download speeds with a mean of $\bar{x} = 15.998$ Mbit/s from the chapter Single-parameter models. This time we assume also σ^2, the variability of the measurements from the speed testing service, to

be unknown. I will use the prior hyperparameters $\mu_0 = 20$, $\kappa_0 = 1$, $\nu_0 = 5$ and $\sigma_0^2 = 5^2$, which agrees in location with my previous prior when σ^2 was assumed known at $\sigma^2 = 5^2$; setting $\nu_0 = 5$ gives a prior equal to the green distribution in the right graph of Figure 52, which I find sensible.

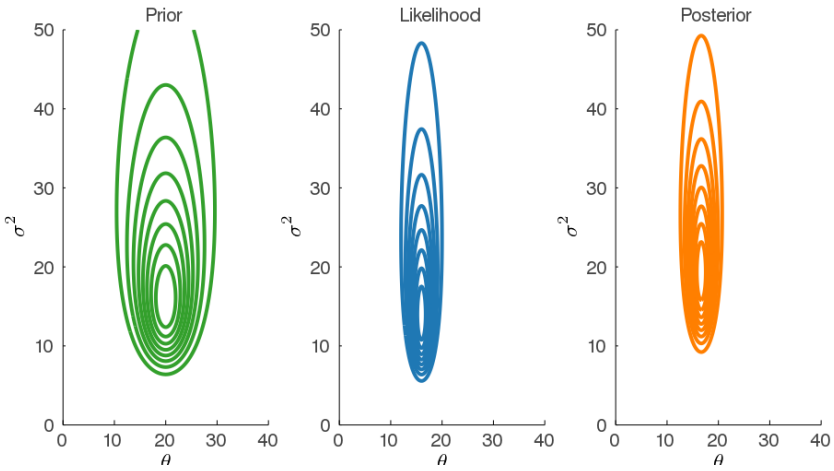


Figure 51: Prior-to-Posterior updating for the internet speed data in the iid Normal model. Contours of joint distributions of θ and σ^2.

Figure 51 displays contours of the joint prior, likelihood and posterior for θ and σ^2; the posterior is more concentrated than the prior, especially for θ. The marginal priors and posterior for the two parameters are shown in Figure 52. The data have made both marginal posteriors more concentrated, but less so for σ^2 since we do not learn so much about a variance from only $n = 5$ observations. The probability of at least 20 Mbit download speed has decreased from the prior probability of 0.5 to 0.066 in the posterior.

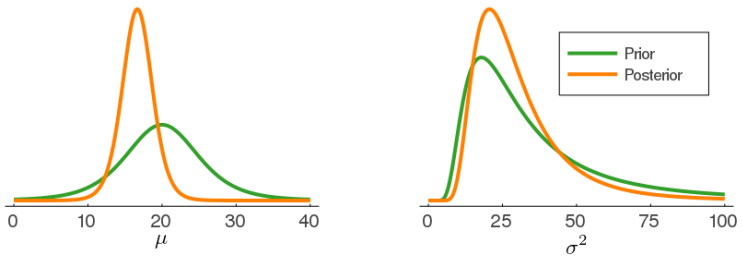


Figure 52: Marginal posteriors for the internet speed data in the iid Normal model.

A first look at Monte Carlo simulation

The iid Gaussian model with conjugate prior is an example of a model where we can obtain both the joint and the marginal pos-

teriors in analytical form. This will seldom be the case in more complex models or when non-conjugate priors are used. The idea with Monte Carlo methods is to simulate **posterior draws** of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|x_1,\ldots,x_n)$ and approximate the posterior by for example a histogram. We will have much more to say about this in Chapter Posterior simulation where powerful simulation algorithms are presented, but we will already here introduce the most basic Monte Carlo simulation method.

posterior draws

The algorithm in Figure 53 gives pseudo-code for simulating from the $p(\theta,\sigma^2|\mathbf{x})$ in the iid normal model by iteratively simulating from $p(\sigma^2|\mathbf{x})$ followed by simulation from $p(\theta|\sigma^2,\mathbf{x})$. Note how this involves using the most recently simulated value of σ^2 when simulating θ. The algorithm includes the subfunction rInvChi2(ν_n,σ_n^2) to draw from the Inv$-\chi^2$ distribution. The algorithm implicitly assumes that the standard library of your programming language includes random number generators rChi2(ν) and rNormal($\mu_n,\sigma^2/\kappa_n$) for the χ^2 and normal distributions, respectively.

Posterior simulation - iid Gaussian with conjugate prior.

Input: data $\mathbf{x} = (x_1,\ldots,x_n)$
number of posterior draws m.
compute μ_n, σ_n^2, κ_n and ν_n using Figure 50.
for i in $1{:}m$ **do**
  $\sigma^2 \leftarrow$ rInvChi2(ν_n,σ_n^2)
  $\theta \leftarrow$ rNormal(μ_n,σ^2/κ_n)
end
Output: m draws for θ and σ^2 from joint posterior.

Function rInvChi2(ν,τ^2)
  x = rChi2(ν)
  y = $\nu\tau^2/x$
  return y

Figure 53: Algorithm for posterior simulation for the iid Normal model with conjugate prior. The rNormal and rChi2 random number generators are assumed to be part of the standard library. The variable σ^2 is highlighted in orange to indicate that the most recent draw of σ^2 is used in the call to the rNormal function.

Example: Internet speed data

Let us now use the algorithm in Figure 53 to simulate from the posterior of θ and σ^2 in the Internet speed datas. The second and third columns in Table 1 shows the output from generating $m = 10{,}000$ joint posterior draws with the algorithm in Figure 53. One attractive feature of simulating from the joint posterior distribution is that all marginal posterior distributions are directly obtained by just selecting the column for the parameter in question; tedious in-

tegration is replaced by plotting a histogram of the selected column. Figure 54 shows the marginals obtained from simulation with the analytical marginal posteriors, which happen to be known in this simple example.

The histograms of the simulated draws in Figure 54 are clearly approximating the posteriors extremely well. Monte Carlo simulation is theoretically known to be **simulation consistent** in the sense that we are guaranteed to get arbitrary close to the true posterior if we simulate a large number of draws. For example, if we let $\theta^{(i)}$ denote the ith posterior draw of any of the parameters in a model, then the law of large numbers implies

$$\bar{\theta}_{1:m} \equiv \frac{1}{m}\sum_{i=1}^{m}\theta^{(i)} \overset{a.s.}{\to} \mathbb{E}(\theta|\mathbf{x}) \text{ as } m \to \infty,$$

where $\overset{a.s.}{\to}$ denotes almost sure convergence, the strongest form of probabilistic convergence. The result says that the mean of the posterior draws will get closer and closer to the theoretical posterior mean $\mathbb{E}(\theta|\mathbf{x})$ as we increase the number of simulations, m. Figure 55 illustrates this convergence by plotting the posterior mean estimates $\bar{\theta}_{1:m}$ for increasing m; note that the figure shows the cumulative estimates only up to $m = 1000$.

draw	θ	σ^2
1	18.165	18.451
2	20.431	29.943
3	15.565	29.094
⋮	⋮	⋮
10,000	16.400	21.668
Mean	16.645	30.813

Table 1: Posterior simulation output for the Internet speed dataset.

simulation consistent

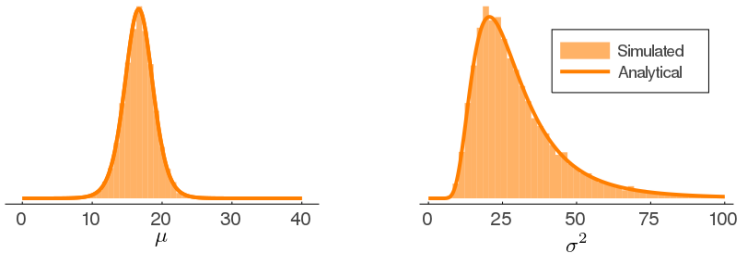


Figure 54: Histogram of simulated marginal posteriors for the internet speed data with analytical marginal posterior densities overlayed.

The central limit theorem (CLT) can be used to prove that $\bar{\theta}_{1:m}$ converges in distribution to a normal distribution. Hence, the following approximation of the posterior estimate $\bar{\theta}_{1:m}$ is accurate when m is large:

$$\bar{\theta}_{1:m} \sim N\left(\mathbb{E}(\theta|\mathbf{x}), \frac{\mathbb{V}(\theta|\mathbf{x})}{m}\right), \tag{31}$$

where $\mathbb{V}(\theta|\mathbf{x})$ is the posterior variance of θ; note that we get the usual reduction in variance that comes from taking averages of m draws, i.e. the variance of $\bar{\theta}_{1:m}$ decreases with m. The result in (31) can be used to determine the required number of draws m needed for a given estimation precision. A multivariate version of the CLT can be used to prove a similar result to (31) when $\boldsymbol{\theta}$ is a vector; an interest-

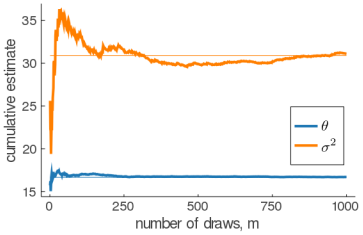


Figure 55: Convergence of the Monte Carlo estimate of the posterior expectation of θ and σ^2. The analytical posterior expectations are displayed as thin horizontal lines.

ing aspect is that $\text{Cov}(\bar{\boldsymbol{\theta}}_{1:m})$ (a covariance matrix in the multiparameter case) still decreases at the rate $1/m$, regardless of the dimension of $\boldsymbol{\theta}$.

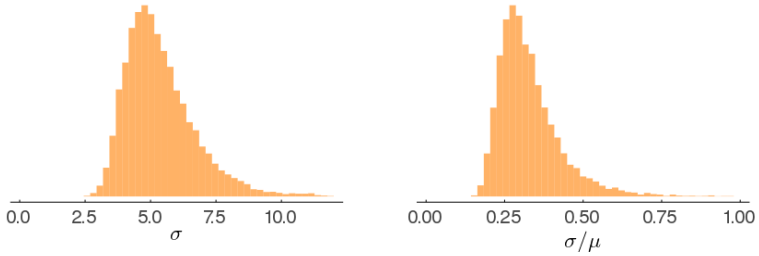


Figure 56: Histogram of simulated marginal posteriors for σ (left) and the coefficient of variation σ/θ (right) for the internet speed data.

It is often the case that the quantities of interest are functions $f(\boldsymbol{\theta})$ of the parameters; for example the **coefficient of variation** σ/θ in the iid normal model. Even when the posterior for the model parameters $\boldsymbol{\theta}$ is available analytically, deriving the posterior for $f(\boldsymbol{\theta})$ involves tedious multidimensional change-of-variables calculations. Here is a second attractive property of simulation: the posterior for $f(\boldsymbol{\theta})$ can be directly obtained from a posterior sample of $\boldsymbol{\theta}$ by simply computing the function $f(\boldsymbol{\theta})$ for each posterior draw. Provided the posterior variance of $f(\boldsymbol{\theta})$ exists, a central limit theorem of the form (31) exists also in this case, with the expected value and variance replaced by those of $f(\boldsymbol{\theta})$.

draw	θ	σ^2	σ/θ	$\theta \geq 20$
1	18.165	18.451	0.236	0
2	20.431	29.943	0.267	1
3	15.565	29.094	0.346	0
⋮	⋮	⋮	⋮	⋮
10,000	16.400	21.668	0.283	0
Mean	16.645	30.813	0.330	0.066

Table 2: Posterior simulation output for the Internet speed dataset with computed functions of the parameters.

coefficient of variation

To illustrate how simulation immediately provides inference for any function of the parameters, Table 2 contains a fourth column named σ/θ with the computed coefficient of variation for each draw. We can now just plot a histogram of this new column to approximate the marginal posterior of the function $f(\theta, \sigma^2) = \sigma/\theta$. The results are presented in the right part of Figure 56; the left part of the figure shows the results for the standard deviation $f(\theta, \sigma^2) = \sqrt{\sigma^2}$.

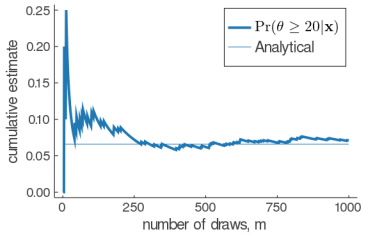


Figure 57: Convergence of the Monte Carlo estimate of $\Pr(\theta \geq 20|\mathbf{x})$. The analytical posterior probability is displayed as a thin horizontal line.

The final column of Table 2 is a binary variable that records if θ was at least 20, i.e. it computes the indicator function $f(\theta, \sigma^2) = I(\theta \geq 20)$. The marginal posterior probability $\Pr(\theta \geq 20|\mathbf{x})$ is then easily approximated by the mean of the final column; Figure 57 illustrates the Monte Carlo convergence of this estimate.

Multinomial data

Categorical data have observations that belong to one of C discrete classes. A computer bug can for example be allocated to C developing teams; an items sold in an auction may reported as: 'defective', 'normal quality', or 'new'; a continuous variable like age can recorded in age intervals: 0-18, 19–28, 29–49, 50–64 and 65+, which

Categorical data

would then also be a categorical variable. The categories in the latter two situations are examples of **ordinal data** where the categogies have a natural order. There are special models for ordinal data which we will not cover in this chapter; here we will consider categorical data without natural order. Categorical variables are often called **multi-class** in the machine learning literature.

ordinal data

multi-class

A multi-class random variable X is often written in **one-hot encoding** as $\mathbf{x} = (x_1, \dots x_C)$ where $X = c$ is encoded as $x_c = 1$ and $x_j = 0$ for $j \neq c$; hence when $C = 3$, $\mathbf{x} = (0, 1, 0)$ means that the observation belongs to the second class. The categorical random variable $X|\boldsymbol{\theta} \sim \text{Cat}(\theta_1, \dots, \theta_C)$ has probability distribution

one-hot encoding

$$p(x) = \theta_1^{x_1} \cdots \theta_C^{x_C}, \tag{32}$$

where $(x_1, \dots x_C)$ is the one-hot encoding of x, $0 < \theta_c < 1$ is the probability of class c and $\sum_{c=1}^{C} \theta_c = 1$. Note how Bernoulli data is the special case with $C = 2$ categories 'success' and 'failure', so that the $\text{Cat}(\theta_1, \dots, \theta_C)$ distribution generalizes the Bernoulli distribution to the case $C > 2$. Figure 58 is an example of $\text{Cat}(\theta_1, \dots, \theta_C)$ for $C = 4$.

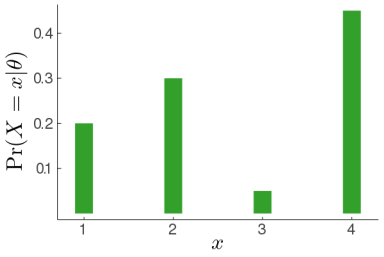


Figure 58: Categorical distribution with probabilities $\boldsymbol{\theta} = (0.20, 0.30, 0.05, 0.45)$.

We saw in Section The likelihood function and maximum likelihood estimation that counting the number of successes s in n binary Bernoulli trials gave rise to $S \sim \text{Binomial}(n, \theta)$ data. In the same way we can count the number of observations in category c for $c = 1, \dots, C$ in multi-class data. This gives data as a count vector $\mathbf{y} = (y_1, \dots, y_C)$ where y_c is the number of observations in category c in $n = \sum_{c=1}^{C} y_c$ 'trials'. Here is an example:

SMARTPHONE SURVEY DATA. A company conducted a survey among $n = 513$ smartphone users. Among other questions, the participants were asked: 'What kind of mobile phone do you mainly use?' with the four options: i) iPhone, ii) Android, iii) Windows and iv) Other/Don't know. The number of responses in the four categories were: $\mathbf{y} = (180, 230, 62, 41)$.

The **multinomial distribution** generalizes the binomial distribution to $C > 2$ categories; its main properties are summarized in Figure 59. The Binomial distribution in Figure 6 is the special case with $C = 2$ categories, which is seen by defining $\theta = \theta_1$, $\theta_2 = 1 - \theta$, $x = x_1$, $x_2 = n - x$, and noting that

Multinomial distribution

$(X_1, \dots, X_C) \sim \text{MultiNom}(n, \boldsymbol{\theta})$
where $\sum_{c=1}^{C} X_c = n$,
$\boldsymbol{\theta} = (\theta_1, \dots, \theta_C)$ and $\sum_c \theta_c = 1$.

$$p(\mathbf{x}) = \frac{n!}{x_1! \cdots x_C!} \theta_1^{x_1} \cdots \theta_C^{x_C}$$

$$\mathbb{E}(X_c) = n\theta_c$$

$$\mathbb{V}(X_c) = n\theta_c(1 - \theta_c)$$

Figure 59: The multinomial distribution.

multinomial distribution

$$\frac{n!}{x_1! x_2!} = \frac{n!}{x!(n-x)!} = \binom{n}{x}. \tag{33}$$

The multinomial distribution is a multivariate distribution with convenient marginalization properties. For example, if we group the

counts in one or more categories - for example turning the smartphone dataset into three categories by merging 'Windows' and 'Other' - the distribution remains multinomial. The probability of a merged category is simply the sum of the probabilities of the merged categories. Hence

$$(x_1, x_2, x_3 + x_4) \sim \text{Multinomial}(\theta_1, \theta_2, \theta_3 + \theta_4).$$

In particular, merging to only two categories - for example 'iPhone' and 'not iPhone' - gives a binomial distribution where the probability of failure (not iPhone) is $\theta_2 + \theta_3 + \theta_4$.

A Bayesian analysis of multinomial data requires a prior distribution for the model parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$. Since each θ_c is a probability, the first distribution that comes to mind may be a Beta distribution; the Beta distribution is not appropriate here however since it does not enforce the constraint that the probabilities sum to one. Hence, the parameter space of the multinomial distribution is the **unit simplex**, i.e. the set $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C) : 0 < \theta_c < 1$ and $\sum_c \theta_c = 1$. Luckily, there is a very nice distribution on the unit simplex, the Dirichlet distribution, summarized in Figure 60.

The Dirichlet distribution is specified with the prior hyperparameters $\alpha_c > 0$, see Figure 61 for some examples. The *relative* sizes of the elements in $\boldsymbol{\alpha}$ determine the prior means for elements of $\boldsymbol{\theta}$. For example, setting $\alpha_1 = \cdots = \alpha_C = 1.5$, as in the upper graph of Figure 61, gives equal prior mean for all categories: $\mathbb{E}(\theta_c) = 1/C$ for all c. The *absolute* size of $\boldsymbol{\alpha}$, measured by $\alpha_+ = \sum_{c=1}^C \alpha_c$, is inversely related to the variance, see Figure 60; hence, the prior hyperparameters $\boldsymbol{\alpha} = (1.5, \ldots, 1.5)$ and $\boldsymbol{\alpha} = (5, \ldots, 5)$ in the upper and middle part of Figure 61 have the same mean, but the latter has smaller variance. Finally, the bottom part of Figure 61 shows an example where the prior mean is different over the categories.

The Dirichlet$(1, \ldots, 1)$ has constant density and is therefore the **uniform distribution on the unit simplex**; this generalizes the result that Beta$(1, 1)$ is uniform on the unit interval $[0, 1]$. Finally, when $\alpha_c < 1$, the Dirichlet density becomes 'bathtub shaped' with probability mass piling up against the edges of the unit simplex.

The Dirichlet distribution is conjugate to the multinomial likelihood which is easily seen by computing the posterior

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{34}$$

$$= \frac{n!}{x_1! \cdots x_C!} \theta_1^{x_1} \cdots \theta_C^{x_C} \cdot \frac{\Gamma(\sum_{c=1}^C \alpha_c)}{\prod_{c=1}^C \Gamma(\alpha_c - 1)} \theta_1^{\alpha_1 - 1} \cdots \theta_C^{\alpha_C - 1} \tag{35}$$

$$= \theta_1^{\alpha_1 + x_1 - 1} \cdots \theta_C^{\alpha_C + x_C - 1}, \tag{36}$$

which is proportional to the Dirichlet$(\alpha_1 + x_1, \ldots, \alpha_C + x_C)$ density. This is a convenient result: the posterior is simply obtained by

Dirichlet distribution

$\boldsymbol{\theta}|\boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ where
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$, $\sum_c \theta_c = 1$,
$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)$ and $\alpha_c > 0$..

$$p(\boldsymbol{\theta}) = k \cdot \theta_1^{\alpha_1 - 1} \cdots \theta_C^{\alpha_C - 1}$$

$$k = \frac{\Gamma(\sum_{c=1}^C \alpha_c)}{\prod_{c=1}^C \Gamma(\alpha_c - 1)}.$$

$$\mathbb{E}(\theta_c) = \frac{\alpha}{\sum_{c=1}^C \alpha_c}$$

$$\mathbb{V}(\theta_c) = \frac{\tilde{\alpha}_c(1 - \tilde{\alpha}_c)}{1 + \alpha_+}$$

$$\alpha_+ = \sum_{c=1}^C \alpha_c.$$

Marginal distributions:

$$\theta_c \sim \text{Beta}\left(\alpha_c, \alpha_+ - \alpha_c\right).$$

Figure 60: The Dirichlet distribution.

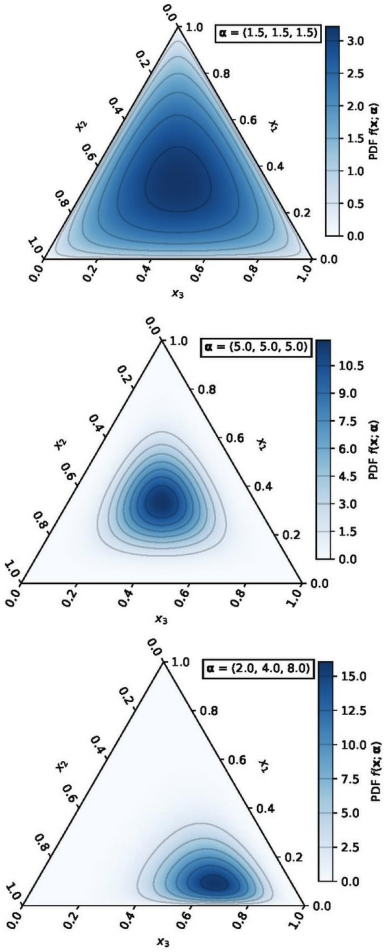


Figure 61: Examples of Dirichlet distributions for $\mathbf{x} = (x_1, x_2, x_3)$.
Source: Wikipedia.

unit simplex

uniform distribution on the unit simplex

adding the data count x_c to the prior hyperparameter α_c in each category. This parallels and generalizes the binary case where a Beta(α, β) prior was updated to a posterior by adding the number of successes s to α and the number of failures f to β. Figure 62 summarizes the prior-to-posterior updating for multinomial data with a Dirichlet prior.

Multinomial data with Dirichlet prior

Model: $\mathbf{n}|\boldsymbol{\theta} \sim$ Multinomial$(\boldsymbol{\theta})$, where
$\mathbf{n} = (n_1, \ldots, n_C)$ are counts in C categories
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$ are category probabilities.
Prior: $\boldsymbol{\theta} \sim$ Dirichlet$(\boldsymbol{\alpha})$, for $\alpha = (\alpha_1, \ldots, \alpha_C)$
Posterior: $\boldsymbol{\theta} \sim$ Dirichlet$(\boldsymbol{\alpha} + \mathbf{n})$

Figure 62: Prior-to-Posterior updating for multinomial data with the Dirichlet prior.

SMARTPHONE SURVEY DATA We are now ready to analyze the four market shares $\theta_1, \ldots, \theta_4$ in the smartphone data. We will determine the prior hyperparameters in the Dirichlet prior using data from a similar survey from four year ago. The proportions in the four categories back then were: 30%, 30%, 20% and 20%. This was a large survey, but since time has passed and user patterns most likely have changed, I value the information in this older survey as being equivalent to a survey with only 50 participants. This gives us the prior:

$$(\theta_1, \ldots, \theta_4) \sim \text{Dirichlet}(\alpha_1 = 15, \alpha_2 = 15, \alpha_3 = 10, \alpha_4 = 10)$$

Note that $\mathbb{E}(\theta_1) = 15/50 = 0.3$ and so on, so the prior mean is set equal to the proportions from the older survey. Also, $\sum_{k=1}^{4} \alpha_k = 50$, so the prior information is equivalent to a survey based on 50 respondents, as required.

The joint posterior distribution of all four shares is by Figure 62 equal to

$$(\theta_1, \ldots, \theta_4)|\mathbf{y} \sim \text{Dirichlet}(15 + 180, 15 + 230, 10 + 62, 10 + 41)$$

The marginal posteriors are plotted in Figure 63 as histograms from Monte Carlo simulation (see the algorithm in Figure 64); the analytical posteriors from Figure 60 are overlayed.

Figure 63 indicates that Android may have the largest market share with a posterior mean around 0.44 versus iPhones posterior mean of 0.35. Computing the probability that Android has the largest market share involves integrating the joint posterior $\boldsymbol{\theta}|\mathbf{y} \sim$ Dirichlet$(\boldsymbol{\alpha} + \mathbf{y})$ over the region $\{\boldsymbol{\theta} : \theta_2 > \max(\theta_1, \theta_3, \theta_4)\}$, a tedious calculation. The probability is however easily computed by

draw	θ_1	θ_2	θ_3	θ_4	I
1	0.33	0.47	0.10	0.09	1
2	0.34	0.44	0.11	0.09	1
3	0.36	0.41	0.13	0.08	1
⋮	⋮	⋮	⋮	⋮	⋮
10,000	0.35	0.43	0.14	0.08	1
Mean	0.34	0.43	0.13	0.09	0.99

Table 3: Posterior simulation output for the multinomial model applied to the Smartphone survey data. The last column is a computed binary indicator for the event that Android has the largest market share, i.e. if $\theta_2 > \max(\theta_1, \theta_3, \theta_4)$.

simulation by recording for each posterior $\boldsymbol{\theta}$ draw if the condition $\theta_2 > \max(\theta_1, \theta_3, \theta_4)$ is satisfied; Table 3 shows that

$$\Pr(\text{Andriod has largest market share}|\mathbf{y}) \approx 0.99,$$

so we can be almost certain that Android is the most popular smartphone operating system.

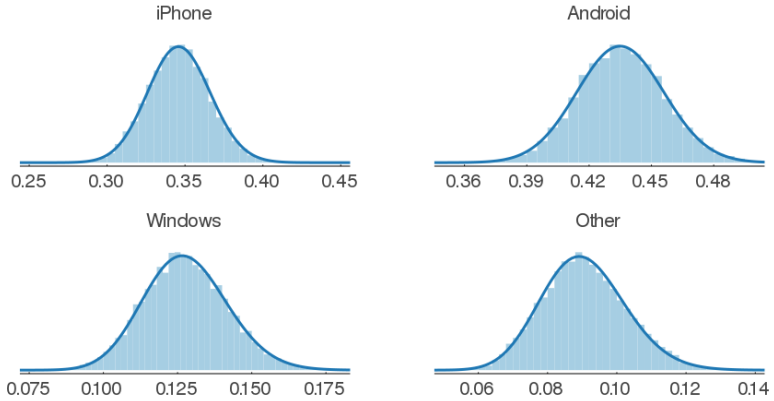


Figure 63: Marginal posteriors of the market shares for the smartphone survey data. Simulated (histogram) draws and analytical density functions (solid curves).

Posterior simulation - Multinomial data, Dirichlet prior.

Input: data $\mathbf{n} = (n_1, \ldots, n_C)$
prior hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)$
the number of posterior draws m.

for i in $1{:}m$ **do**
  $\boldsymbol{\theta} \leftarrow$ rDirichlet$(\boldsymbol{\alpha} + \mathbf{n})$
end
Output: m posterior draws of $\boldsymbol{\theta} = (\theta_1, ..., \theta_C)$.

Function rDirichlet$(\boldsymbol{\alpha})$
  for c in $1{:}C$ **do**
    $\mathbf{y}[c] \leftarrow$ rGamma$(\boldsymbol{\alpha}[c], 1)$
  end
  return $\mathbf{y}/$Sum$(\mathbf{y})$

Figure 64: Algorithm for posterior simulation for the multinomial model with the conjugate Dirichlet prior. The rGamma random number generator is assumed to be part of the standard library.

Likelihood and Information

We will here introduce some notions of information, which will also be important at many other places in this book. The first idea is to

measure the amount of information in a log-likelihood function by the second derivative at the maximum likelihood estimate (MLE):

Definition (Observed information - one-parameter case)**.** *The **observed information** in a sample* $\mathbf{x} = (x_1, \ldots, x_n)$ *is defined as* observed information

$$J_{\theta,\mathbf{x}} = -\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}_{\text{MLE}}} \tag{37}$$

To see why this makes sense, recall for calculus that the second derivative measures how fast the first derivative changes, i.e. $J_{\theta,\mathbf{x}}$ measures how peaked the log-likelihood is around the maximum. The negative sign in the definition makes sure the information is always positive, since we know from calculus that the second derivative is negative at the maximum.

The observed information can be extended to the multiparameter case as follows.

Definition (Observed information - multi-parameter case)**.** *The **observed information matrix** in a sample* $\mathbf{x} = (x_1, \ldots, x_n)$ *from the model* $p(\mathbf{x}|\boldsymbol{\theta})$ *with a p-dimensional parameter vector* $\boldsymbol{\theta}$ *is defined as* observed information matrix

$$J_{\boldsymbol{\theta},\mathbf{x}} = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{MLE}}}, \tag{38}$$

where $\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}$ *is the* $p \times p$ *matrix of second derivatives.*

The matrix $\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}$ in (38) may be a little intimidating. Writing out its elements explicitly in the case of two parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2)$,

$$\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \begin{pmatrix} \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2^2} \end{pmatrix},$$

we see that calculating $\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}$ is no harder than calculating a single second derivative, there are just more of them. Luckily, we will learn in the Chapter Classification that we can often let the computer do this job for us.

The observed information $J_{\boldsymbol{\theta},\mathbf{x}}$ varies from sample to sample. The average, or expected, information is called the Fisher information:

Definition (Observed information)**.** ***Fisher information** is the expected information over all possible samples from the model* Fisher information

$$I(\theta) = \mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}}\left(J_{\theta,\mathbf{x}}\right). \tag{39}$$

EXERCISES

1. Derive the marginal posterior of θ in (30) for the iid Gaussian model $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$.
2. Let $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$, where θ is assumed known. Show that the Inv$-\chi^2$ distribution is a conjugate prior for σ^2.

NOTEBOOKS

1. See the notebook multi-parameter models.

Priors

The secret sauce of Bayesian learning is the prior. Only with a prior can we turn a likelihood function into a probability distribution for the unknown parameters, and subsequenly use this posterior distribution for decision making. Priors make it possible to use fuse information from a variety of different sources. This chapter discusses some of these different types of prior information, and we will return to this issue in later chapters when we perform more serious modelling.

Eliciting a prior takes effort and there many be situations where one may want to use as little prior information as possible, or at least use a prior where the information added is transparent to everyone involved. One example is the reporting of scientific results to an unknown audience with potentially rather different prior opinions. The ideal would be to present the posterior distribution for a variety of different priors to contrast the different views and to examine the possibility of a subjective consensus. This is challenging however, particularly when the model contains many parameters and data is weak; Sections Noninformative priors and Invariant priors presents several 'non-informative' priors that may be appealing in such circumstances.

Time series

A time series model will be used to illustrate some ways in which priors can be specified. Time series data have **dependent observations**, and models for such data are therefore necessarily more complex; it is however worthwhile to spend a little time on it in this chapter as the particular model presented here will be used many times in this book.

dependent observations

A **time series** is a realization of a **stochastic process** observed over discrete number of time periods, here denoted by $t = 1, 2, \ldots, T$. Time series are one of the most commonly occuring data types and are destined to play a large role in the future as time-stamped data are now collected by many electronic devices and at a rapid pace. Figure

time series

stochastic process

65 shows a time series of Swedish inflation, Figure 66 display the daily number of rides with a bike sharing company, and Figure 67 illustrates a time series of electroencephalography (EEG) recordings of electrical activity at one brain location. Many timeseries consist of multivariate measurements at every time period, for example EEG recordings taken simultaneously at multiple locations, see Figure 68, or meteorological data collected at different spatial locations.

The **autoregressive model** of order p is a time series model of the form

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), \quad (40)$$

where y_{t-k} is the kth **lagged value** of time series and ε_t are the disturbances, or innovations, that drives the process. Hence, an AR(p) process models today's value y_t as a linear function of the measurements at the p most recent days $y_{t-1}, \ldots, y_{t-p}$ plus a random disturbance ε_t. The time series may equally well be observed on another frequency than daily, for example monthly, with lags being past months. The effect of the kth lags is captured by the AR coefficients ϕ_k.

The AR(p) process in (40) is in **steady-state form** where the parameter μ is the unconditional mean $\mathbb{E}(y_t)$ of the process. We assume that the AR(p) process is **stationary**, meaning that the mean μ and variance $\mathbb{V}(y_t)$ remain unchanged over time, and also that the covariance between any two time points $\mathrm{Cov}(y_t, y_s)$ is only a function of the time distance $|t - s|$. The assumption of a constant mean may seem restrictive, but often means stationary around a deterministic time trend. The unconditional mean μ is important since long horizon forecasts are guaranteed to end up at μ when the process is stationary, i.e.

$$\mathbb{E}(y_{T+h}|y_{1:T}) \to \mu \text{ as } h \to \infty,$$

where $y_{1:T}$ are all historical data available at the time of the forecast $t = T$. The convergence usually happens rather fast in applications; see Figure 69 where an AR(1) model estimated by maximum likelihood is used to predict Swedish inflation for the coming 60 months.

In later chapters we will learn how to obtain the joint posterior of all parameter $p(\mu, \phi_1, \ldots, \phi_p, \sigma^2|\mathbf{y})$ by approximation or simulation. In this chapter will only worry about how to elicit a prior distribution for all model parameters, i.e. the joint prior $p(\mu, \phi_1, \ldots, \phi_p, \sigma^2)$. We make the simplifying assumption that all parameters are independent a priori; this is most likely not our true beliefs since properties like stationarity involves all ϕ parameters, but it is nevertheless what is most often used in applications. We will walk through a number of methods for prior elicitation and use different methods for different parameters.

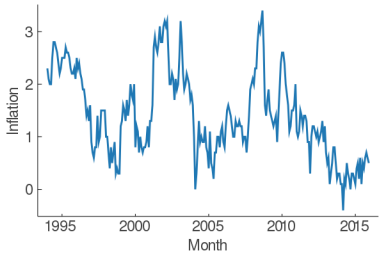


Figure 65: Swedish inflation 1995-2016 - annualized monthly observations.

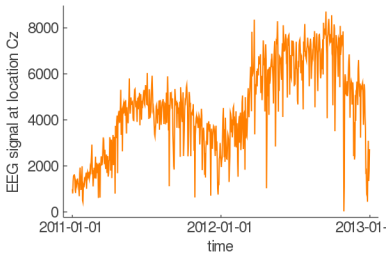


Figure 66: Daily number of rides with a bike sharing company.

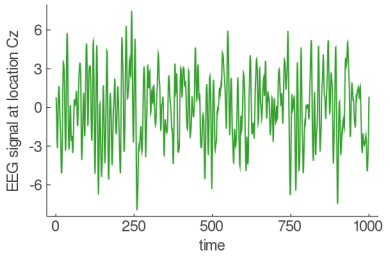


Figure 67: EEG recordings of electrical activity at one brain scalp location.

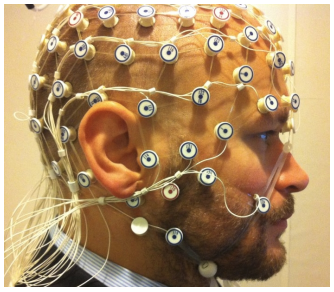


Figure 68: Positioning of EEG electrodes on a subject's brain scalp.

autoregressive model

lagged value

steady-state form

stationary

Past or other data

Bayes' theorem dictates that we are not allowed to use the same data in the likelihood and in the prior, i.e. no double dipping of the data if you want the posterior to correctly quantify uncertainty. It is however allowed to use **past data** for specifying the prior as long as that data are not used in the likelihood; for example, fitting the time series model to data on Swedish inflation data before 1985 and use those estimates as the prior mean. Since older data can be from a different economic regime, one would probably use a fairly large prior variance; this is similar to how we used an older survey for the Dirichlet prior in the Smartphone survey data.

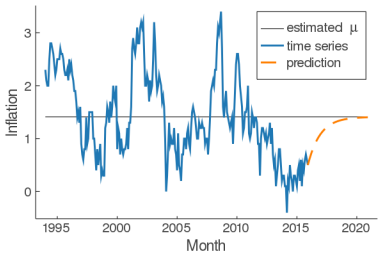


Figure 69: Swedish inflation 1995-2016 with 60 months ahead mean prediction in dashed orange.

past data

We may base our prior on estimates of the model's parameters from **other data**, e.g. inflation data from other countries during the same time period 1985 – 2016. Other countries are certainly different from Sweden, but still relevant, especially data from similar countries.

other data

Expert opinion

The ML estimate of the mean of the time series is $\hat{\mu}_{MLE} = 1.409$, which constrains the mean forecasts at longer horizon to end up at 1.409; see Figure 69. This is lower than the Central Bank of Sweden's inflation target at 2%. We can use this form of expert opinion as a $\mu \sim N(2, \tau_0^2)$ prior with a small prior variance τ_0^2, if we trust the central bank experts. Prior information on the steady-state has been shown to improve forecasting performance for a number of economic variables; see Villani [2009].

Prior elication of the experts were made on a quantity that was well understood by central bank economists, the long run behavior of inflation. The challenge is to elicit prior beliefs from experts on quantities that the expert understands well. This will often involve observable quantities, like inflation, rather than abstract parameters in statistical models. The process is often iterative where model consequences from the initially given expert opinion are presented to the expert, who then adjusts the initial opinion. Eliciting expert opinions is large area in itself, with help from cognitive science to account for the biases and shortcomings that are part of being a human.

Structured regularization priors

An important type of prior beliefs are priors that regularize, or shrink, parameter-rich models. **Regularization priors** are particularly popular in machine learning for probabilistically restricting

Regularization priors

complex models that would otherwise easily overfit the data. There will be many examples of regularization priors later in the book, but we can get an first understanding of the concept from a commonly used prior for the autoregressive parameters $\phi_1, \ldots, \phi_p$ in the AR process. A regularization prior $\phi_1, \ldots, \phi_p$ makes it possible to use a large **lag length** p even on shorter time series. The prior embodies the idea that the magnitude of the ϕ_k are likely to be smaller for larger k, as in the following prior:

lag length

$$\phi_k \sim \mathrm{N}\left(\mu_k, \frac{\tau^2}{k^2}\right), \tag{41}$$

where $\mu_k = 0$ for all k except for the first lag where $\mu_1 = 0.8$, for example. This centers the prior on the AR(1) process with coefficient $\phi_1 = 0.8$, which is a reasonable prior guess for Swedish inflation.

The hyperparameter τ is the prior standard deviation of ϕ_1. The hyperparameter τ is called the **global shrinkage** since it has the effect of shrinking all ϕ_k toward their prior mean; this is the same effect as the prior standard deviation τ_0 had in the iid normal model in Chapter Single-parameter models where the posterior mean μ_n was shrunk toward the prior mean μ_0 via the weight w. Finally, the regularization part of the prior is that the factor $1/k^2$ reduces the prior variance of ϕ_k for longer lags; longer lags are more likely to be redundant a priori, and their ϕ_k will only be sizeable in the posterior if the data strongly suggest so.

global shrinkage

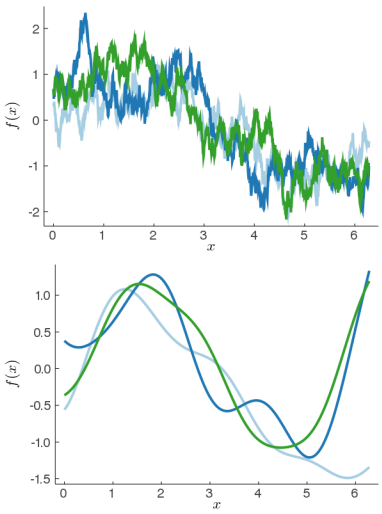


Figure 70: Three simulated draws from a prior over functions without smoothness beliefs (top) and with smooth beliefs (bottom).

Priors can more generally be used to incorporate **smoothness beliefs**. For example, we will later analyze nonlinear regression models where a response variable y is functionally related to an explanatory variable x via some nonlinear function $f(x)$. Rather than assuming a restrictive functional form we often want $f(\cdot)$ to be flexible enough to adapt to almost any shape. However, our prior beliefs may still be that $f(\cdot)$ is smooth; Figure 70 shows examples of priors for function with wiggly and smooth beliefs. Note that the parameter space here is the abstract space of functions, as will be explained in Chapter Gaussian processes. We will in later chapters see many examples of quite elegant use of priors to impose smoothness without loosing desired flexibility. A well designed smoothness prior tames the flexibility in the right way and thereby helps to avoid overfitting the data.

smoothness beliefs

Hierarchical priors

The structure of the presented regularization prior for the AR(p) process is attractive, but it may be hard to specify an exact value for the global shrinkage τ. The solution is simple: if something is unknown to you, put a prior on it. This gives rise to the following

hierarchical prior on the AR coefficients

hierarchical prior

$$p(\phi_1, \ldots, \phi_p, \tau^2) = p(\phi_1|\tau^2) \cdots p(\phi_p|\tau^2)p(\tau^2),$$

where each $p(\phi_k|\tau^2)$ is the previous $\mathrm{N}\left(\mu_k, \frac{\tau^2}{k^2}\right)$, with independence now only conditionally on τ^2, and $p(\tau^2)$ is the marginal prior for the unknown prior hyperparameter τ^2. Note that the joint posterior $p(\mu, \phi_1, \ldots, \phi_p, \sigma^2, \tau^2)$ also involves τ^2, so data will also inform us about τ^2. Since τ^2 is a variance parameter, the prior $\tau^2 \sim \mathrm{Inv}{-}\chi^2(\nu_0, \tau_0^2)$ is a natural choice. We still need to specify τ_0^2 our 'best guess' for τ^2 and the uncertainty via ν_0, but the posterior is often considerably less sensitive to these prior hyperparameters further down the hierarchy.

Noninformative priors

It is often convenient to use a prior with relatively little information, at least for some model parameters. Eliciting priors takes effort and we sometimes prefer to specify priors for some parameters with a little less care than other key parameters. The data may also be known to be highly informative on some model parameters and the prior will therefore anyway be overruled by the likelihood. In short, it can be convenient to give some parameters a noninformative prior. A noninformative prior is a bit of a misnomer since any prior carries some information; see Irony and Singpurwalla [1997] for transcribed car dialogue among Bayesian statisticians about this topic. Consider for example the iid Bernoulli(θ) where $\theta \in [0, 1]$. The Uniform$(0, 1)$ distribution is a candidate for a noninformative prior since it assigns the same density to every possible value of θ. There are at least two arguments against this seemingly natural idea.

First, recall that the posterior from a $\theta \sim \mathrm{Beta}(\alpha, \beta)$ prior is $\theta|\mathbf{x} \sim \mathrm{Beta}(\alpha + s, \beta + f)$. This means that the prior carries the information equivalent to a prior sample of α successes and β failures. Since the Uniform$(0, 1)$ distribution is the Beta$(1, 1)$ distribution, the uniform prior is equivalent to a prior sample of $n = 2$ trials with one success and one failure; this is clearly *some* information. An alternative definition of a noninformative prior is the **zero sample prior** Beta(ϵ, ϵ) where $\epsilon \downarrow 0$, i.e. ϵ is a tiny number; the posterior is then Beta(s, f). The idea of the zero sample prior carries directly over the conjugate analysis for exponential family models presented in Figure 44 by letting ν_0 and τ_0 go to zero.

zero sample prior

A second argument against a uniform density as noninformative is that uniformity is typically not preserved when θ is transformed to an alternative parametrization $\phi = g(\theta)$, where $g(\cdot)$ is a one-to-one

transformation; for example $g(\theta) = \log(\theta/(1-\theta))$, the log-odds transformation of the Bernoulli success probability θ. To see this we use the results on transformations of random variable in Figure 71 to obtain

$$p_\phi(\phi) = p_\theta\left(g^{-1}(\phi)\right)\left|\frac{\partial g^{-1}(\phi)}{\partial \phi}\right| = 1 \cdot \frac{e^\phi}{(1+e^\phi)^2},$$

since $p_\theta(\theta)$ is uniform and the inverse transformation is $g^{-1}(\phi) = e^\phi/(1+e^\phi)$. Hence, a uniform distribution for θ does not imply a uniform distribution on the log-odds. In the next section we will encounter rules for constructing priors that are guaranteed to be invariant to one-to-one transformations of the model parameter.

Transforming variables

Let $X \sim p_X(x)$ and $Y = g(X)$, where $g(\cdot)$ is a one-to-one continuously differentiable transformation with inverse $X = g^{-1}(Y)$.
The density of Y is then

$$p_y(y) = p_x(g^{-1}(y))\left|\frac{\partial}{\partial y} g^{-1}(y)\right|$$

Figure 71: Transformation of random variables.

Invariant priors

As we saw in the previous section, a prior which is uniform in one parametrization is usually not uniform in another parametrization; the uniform distribution is not an **invariant prior** for θ in the Bernoulli model. Jeffreys' rule is a method for constructing priors that are guaranteed to be invariant to any one-to-one transformation of the parameter.

invariant prior

Definition (Jeffreys' rule)**.** ***Jeffreys' prior*** *for a parameter vector $\boldsymbol{\theta}$ in a model $p(\mathbf{x}|\boldsymbol{\theta})$ is of the form*

$$p(\boldsymbol{\theta}) = |I(\boldsymbol{\theta})|^{1/2}. \tag{42}$$

where $I(\boldsymbol{\theta})$ is the Fisher information matrix and $|\cdot|$ denotes the matrix determinant.

Jeffreys' prior

We will for simplicity concentrate on the one-parameter version $p(\theta) = I(\theta)^{1/2}$ in this section. It can be proved that Jeffreys' prior is invariant to reparametrization [Migon et al., 2014], which was physicist Harold Jeffreys' original motivation for the rule [Jeffreys, 1998]. Invariance means that the following two ways to obtain a prior for θ give identical results:

(A) apply Jeffreys' rule directly in the θ-parametrization to obtain

$$p_\theta(\theta) = I(\theta)^{1/2}.$$

(B) apply Jeffreys' rule in the ϕ-parametrization to first obtain

$$p_\phi(\phi) = I(\phi)^{1/2},$$

and then transform to $p_\theta(\theta)$ by the variable transformation formula in Fig 71

$$p_\theta(\theta) = p_\phi(\phi(\theta))\left|\frac{d\phi(\theta)}{d\theta}\right| = I(\phi(\theta))^{1/2}\left|\frac{d\phi(\theta)}{d\theta}\right|.$$

EXAMPLE: JEFFREYS' PRIOR FOR BERNOULLI TRIALS. Consider once again the iid Bernoulli model

$$x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta),$$

with likelihood $\ln p(\mathbf{x}|\theta) = s \ln \theta + f \ln(1-\theta)$. The first and second derivative of the log-likelihood are

$$\frac{d \log p(\mathbf{x}|\theta)}{d\theta} = \frac{s}{\theta} - \frac{f}{(1-\theta)}$$
$$\frac{d^2 \log p(\mathbf{x}|\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{f}{(1-\theta)^2}$$

so that the Fisher information is (using lowercase letter for the random variable s and f)

$$I(\theta) = \frac{E_{\mathbf{x}|\theta}(s)}{\theta^2} + \frac{E_{\mathbf{x}|\theta}(f)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Thus, the Jeffreys' prior is

$$p(\theta) = I(\theta)^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto \text{Beta}(1/2, 1/2). \tag{43}$$

Hence Jeffreys' prior lies between the zero imaginary sample prior $\text{Beta}(\epsilon, \epsilon)$ and the uniform $\text{Beta}(1,1)$. This derivation corresponds to Route A above. Exercise 1 shows that the same $\theta \sim \text{Beta}(1/2, 1/2)$ prior is obtained by taking Route B.

EXAMPLE: JEFFREYS' PRIOR FOR A GAUSSIAN VARIANCE. Consider the model $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$. Let us also assume that θ is known and we use Jeffreys' rule to obtain the invariant prior for σ^2. The log-likelihood is

$$\log p(\mathbf{x}|\sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}$$

with first and second derivative

$$\frac{\partial}{\partial \sigma^2} \log p(\mathbf{x}|\sigma^2) = -\frac{1}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \theta)^2}{2(\sigma^2)^2}$$
$$\frac{\partial^2}{\partial (\sigma^2)^2} \log p(\mathbf{x}|\sigma^2) = \frac{1}{2(\sigma^2)^2} - \frac{\sum_{i=1}^n (x_i - \theta)^2}{(\sigma^2)^3}.$$

Since $\mathbb{E}_{\mathbf{x}} \sum_{i=1}^n (x_i - \theta)^2 = \sum_{i=1}^n \mathbb{E}_{x_i}(x_i - \theta)^2 = n\sigma^2$ we have

$$I(\sigma^2) = -\frac{1}{2(\sigma^2)^2} + \frac{n\sigma^2}{(\sigma^2)^3} = -\frac{1}{2(\sigma^2)^2} + \frac{n}{(\sigma^2)^2} = \frac{n - 1/2}{(\sigma^2)^2},$$

so Jeffreys' prior for the variance is

$$p(\sigma^2) = I(\sigma^2)^{1/2} \propto \frac{1}{\sigma^2},$$

which also implies that Jeffreys' prior for standard deviation is $p(\sigma) \propto \frac{1}{\sigma}$ by the variable transformation formula in Figure 71 and the invariance of the Jeffreys' prior. Since

$$\int_0^\infty \frac{1}{\sigma} d\sigma = \infty$$

Jeffreys' rule gives an **improper prior** in this case, i.e. a not a proper density since its integral diverges. Improper priors are somewhat strange, but can be successfully used in practice if the posterior density is known to be proper, i.e. has a finite integral over the whole parameter space. The $1/\sigma$ form of Jeffreys' prior may seem peculiar as it seemingly favors small values for σ. One way of understanding this prior is that it corresponds to a uniform distribution on $\log \sigma \in \mathbb{R}$. In the case where θ and σ^2 are unknown, the multiparameter version of Jeffreys' rule shows that Jeffreys prior for σ is still $1/\sigma$ and the prior for θ is uniform.

improper prior

Jeffreys' rule has a serious drawback: it violates the likelihood principle; see Section Bayesian learning and the likelihood principle. The reason is that Jeffreys' rule is based on the Fisher information, which is an expectation with respect to the sampling distribution $p(\mathbf{x}|\boldsymbol{\theta})$. Exercise 2 asks you to derive Jeffreys' prior for binary data obtained by negative binomial sampling, instead of Bernoulli trials. This exercise shows that Jeffreys' prior for the success probability θ is not the Beta$(1/2, 1/2)$ that we obtained for Bernoulli trials.

Probably the most promising so called Objective Bayes approach is the **reference prior** proposed by José Bernardo based on information arguments. It is motivated as a non-informative prior useful for scientific reporting where one wants to present posterior results to a wide audience using a single well understood prior. The reference prior is invariant to one-to-one transformations and is in fact equal to Jeffreys' prior when the usual regularity conditions for likelihood inference apply. The reference prior is more general however, and avoids some of problems that have been found with Jeffreys' rule; see Bernardo and Smith [2009] for a comprehensive introduction to reference priors.

reference prior

EXERCISES

1. Show that using Jeffreys rule to obtain a prior for the log odds $\phi \equiv \log \theta/(1-\theta)$ in Bernoulli trials implies the same Beta$(1/2, 1/2)$ prior for θ (i.e. that Route A and B in the text give the same prior).
2. Derive Jeffreys' prior for the success probability θ in the negative binomial model for a dataset where n trials were needed to obtain a predetermined s number of successes. Compare with the

Jeffreys' prior derived for the Bernoulli model in the text. Discuss the implication for the likelihood principle.

NOTEBOOKS

1. See the notebook priors.

Regression

Regression models are the most important of all statistical models as they appear as a component in nearly any situation where an output variable y is modeled as function of a set of input variables $\mathbf{x} = (x_1, \ldots, x_p)^\top$, where $\top$ denotes vector tranpose. The input variables are often called **covariates**, predictors or **features**, and the output variable is most commonly termed the **response variable** or target variable. In the chapter Classification we will see regression models for a binary response variable and also for response variables of other data types, for example counts. Regression is also the basis for deep neural networks where a linear combination of covariates are passed through several nonlinear activation functions before finally being linked to the response.

covariates

features

response variable

The basic **Gaussian linear regression model** is

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2) \qquad \text{for } i = 1, \ldots, n,$$

Gaussian linear regression model

where $\mathbf{x}_i$ is a vector with observations on the p covariates and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the vector of **regression coefficients**. The β_j are called **weights** in the machine learning literature and are therefore frequently denoted by w_j. The model is said to be **homoscedastic** since the error variance σ^2 is the same for all observations.

regression coefficients

weights

homoscedastic

It is convenient to stack all n response observations in a vector $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and the covariate observations vectors as rows in the $n \times p$ covariate matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$. The Gaussian linear regression model can then be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 I_n), \tag{44}$$

where $\boldsymbol{\varepsilon}$ is vector with all the ε_i and $\mathrm{N}(0, \sigma^2 I_n)$ is the multivariate normal distribution with diagonal covariance matrix $\sigma^2 I_n$ and I_n is the identity matrix; the simple diagonal structure of $\mathrm{Cov}(\varepsilon)$ reflects the assumption that the ε_i are independent with the same variance.

Likelihood and MLE

The likelihood for the linear regression model with homoscedastic Gaussian errors is given by the following multivariate normal distri-

bution

$$\mathbf{y}|\beta, \sigma^2, \mathbf{X} \sim \mathrm{N}(\mathbf{X}\beta, \sigma^2 I_n), \tag{45}$$

where we note that the covariates $\mathbf{X}$ are assumed fixed so the likelihood is the distribution of only the response $\mathbf{y}$.

The **least squares estimator** $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is well known to minimize the sum of squared **residuals**

$$Q(\boldsymbol{\beta}) \equiv (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

least squares estimator

residuals

When the errors are homoscedastic Gaussian, $\hat{\boldsymbol{\beta}}$ is also the MLE since the log-likelihood from (45) is a constant plus $-(1/2\sigma^2)Q(\boldsymbol{\beta})$.

The sampling distribution of the MLE is easily obtained since $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is a linear function of $\mathbf{y}$ and $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is a constant matrix. Since $\mathbf{y}|\beta, \sigma^2, \mathbf{X} \sim \mathrm{N}(\mathbf{X}\beta, \sigma^2 I_n)$, the frequentist sampling distribution of $\hat{\boldsymbol{\beta}}$ is obtained by applying the result in Figure 72 with $\boldsymbol{\mu} = \mathbf{X}\beta$, $\Sigma = \sigma^2 I_n$ and $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}$ to obtain

$$\hat{\boldsymbol{\beta}}|\beta, \sigma^2, \mathbf{X} \sim \mathrm{N}\big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big),$$

from which we also see that the MLE is unbiased for $\boldsymbol{\beta}$.

Linear transformation of Gaussians

Let $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ be multivariate Gaussian and $\mathbf{A}$ a constant full rank $m \times p$ matrix. Then

$$\mathbf{A}\mathbf{x} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\Sigma\mathbf{A}^\top).$$

Figure 72: Linear transformation of Gaussians.

The MLE for σ^2 can be shown to be $\hat{\sigma}^2 \equiv (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n$. The estimator $\hat{\sigma}^2$ is biased for σ^2, and the following unbiased estimator is typically used instead

$$s^2 \equiv \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - p}.$$

Non-informative prior

We will start with the invariant Jeffreys' prior (see Section Invariant priors) which can be shown to be

$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2},$$

i.e. an improper uniform distribution for $\boldsymbol{\beta}$ independently of σ^2; note that σ^2 has the same $1/\sigma^2$ prior as in the iid normal model derived in Invariant priors.

The joint posterior for $\boldsymbol{\beta}$ and σ^2 is given by Bayes' theorem as

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2) \propto N(\mathbf{y}|\mathbf{X}\beta, \sigma^2 I_n) \cdot \frac{1}{\sigma^2} \\ &= |2\pi\sigma^2 I_n|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \cdot \frac{1}{\sigma^2}, \end{aligned} \tag{46}$$

where the conditioning on the fixed covariates $\mathbf{X}$ is suppressed to simplify the notation. Now, $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ can be rewritten using the MLE $\hat{\boldsymbol{\beta}}$ as

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \tag{47}$$

which can be directly verified by substituting the definition of $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Recall from linear algebra that the determinant of a diagonal matrix is the product of its diagonal elements, so $|2\pi\sigma^2 I_n| = (2\pi\sigma^2)^n \propto (\sigma^2)^n$. Using this result and (47) in (46) we obtain the posterior

$$p(\boldsymbol{\beta},\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-(n+2)/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right\} \tag{48}$$

$$\cdot\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\} \tag{49}$$

The posterior is most transparent if we use the decomposition of the joint posterior

$$p(\boldsymbol{\beta},\sigma^2|\mathbf{y}) = p(\boldsymbol{\beta}|\sigma^2,\mathbf{y})p(\sigma^2|\mathbf{y}).$$

Focusing first on $p(\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X})$ we only need to be concerned with the last factor in (48) as it is the only part that depends on $\boldsymbol{\beta}$; note that $\hat{\boldsymbol{\beta}}$ only depends on the data. We immediately recognize this last factor as proportional to the multivariate normal density, so

$$\boldsymbol{\beta}|\sigma^2,\mathbf{y} \sim N(\hat{\boldsymbol{\beta}},\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

The marginal posterior of σ^2 is obtained by integrating out $\boldsymbol{\beta}$ in (48)

$$\begin{aligned}
p(\sigma^2|\mathbf{y}) &= \int p(\boldsymbol{\beta},\sigma^2|\mathbf{y})d\boldsymbol{\beta} \\
&\propto (\sigma^2)^{-(n+2)/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right\} \\
&\cdot\int\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}d\boldsymbol{\beta} \\
&\propto (\sigma^2)^{-(n+2)/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right\}(\sigma^2)^{p/2},
\end{aligned}$$

where the last proportionality comes from the fact that

$$\int\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}d\mathbf{x} = |2\pi\Sigma|^{1/2}$$

for any p-vectors $\mathbf{x}$ and $\boldsymbol{\mu}$, and positive definite matrix Σ since we know that the $N(\mathbf{x}|\boldsymbol{\mu},\Sigma)$ density integrates to one over $\mathbb{R}^p$. The marginal posterior for σ^2 is therefore

$$p(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-[1+(n-p)/2]}\exp\left\{-\frac{1}{2\sigma^2}(n-p)s^2\right\}, \tag{50}$$

which can be recognized as proportional to the Inv$-\chi^2(n-p,s^2)$ density.

We summarize the prior-to-posterior updating in Gaussian linear regression with a noninformative prior in Figure 73.

Gaussian linear regression with non-informative prior

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 I_n)$
Prior: $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$
Posterior: $\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$
$\sigma^2|\mathbf{y}, \mathbf{X} \sim \mathrm{Inv}{-}\chi^2(n-p, s^2)$
where $\hat{\boldsymbol{\beta}} \equiv (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ and $s^2 \equiv (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p)$.

Figure 73: Prior-to-Posterior updating for the Gaussian linear regression with non-informative prior.

Conjugate prior

Let us now turn to the more interesting case with a conjugate prior for the Gaussian linear regression. Recall the form of the conjugate prior for the iid normal model

Gaussian linear regression with conjugate prior

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 I_n)$
Prior: $\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\mu}_0, \sigma^2\Omega_0^{-1})$
$\sigma^2|\mathbf{y}, \mathbf{X} \sim \mathrm{Inv}{-}\chi^2(\nu_0, \sigma_0^2)$
Posterior: $\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$
$\sigma^2|\mathbf{y}, \mathbf{X} \sim \mathrm{Inv}{-}\chi^2(\nu_n, \sigma_n^2)$
where $\hat{\boldsymbol{\beta}} \equiv (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ and $s^2 \equiv (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p)$.

Figure 74: Prior-to-Posterior updating for the Gaussian linear regression with non-informative prior.

EXERCISES

1. This is the first problem.
2. This is the second problem.

NOTEBOOKS

1. See the notebook regression.

Prediction and Decision making

EXERCISES

1. (a) Let $x_1, ..., x_n \overset{iid}{\sim} Bern(\theta)$, with a Beta$(\alpha, \beta)$ prior for θ. Derive the predictive distribution for x_{n+1}.

   (b) You need to decide if you bring your umbrella during your daily walk. It has rained on two days during the last ten days, and you assess those ten days to be representative also for the weather today, the 11th day. Your utility for the action-state combinations are given in the table below. Assume a Beta$(1, 1)$ prior for θ. Compute the Bayesian decision.

   (c) How sensitive is your decision in (b) to the changes in the prior hyperparameters, α and β?

2. (a) Let x_i be the number of sales of a product on month i. Let $x_1, ..., x_n \overset{iid}{\sim} N(\theta, \sigma^2)$ be the (approximate) distribution for the sales, and let $\theta \sim N(200, 50^2)$ a priori. Assume that $\sigma^2 = 25^2$ and that we have observed $n = 5$ and $\bar{x} = 320.4$. Compute the predictive distribution for x_6.

   (b) The company has the choice of performing a marketing campaign for their product. The marketing campaign costs 300 and is believed to increase sales by 20% compared to when no campaign is performed. The company sells the product for $p = 10$ dollar and the cost of producing the product is $q = 5$ dollar. There are no fixed production costs. Assume that the company's utility is described by $U(y) = 1 - \exp(-y/1000)$, where y is the total profit from sales in the next month. Should the company perform the marketing campaign? *Hint: the expected value of the exponential function of a normal random variable* $S \sim N(\mu, \sigma^2)$ *is* $\mathbb{E}(\exp(S)) = \exp(\mu + \sigma^2/2)$.

NOTEBOOKS

1. See the notebook Prediction and Decision.

Prediction in normal model with known variance

My streaming service becomes unreliable and buffers at speeds below 5Mbit/sec, and what I really like to know is the probability of this 'catastrophic' event. Finding the probability that a *single* measurement will be lower than 5MBit/sec is instead an exercise in prediction.

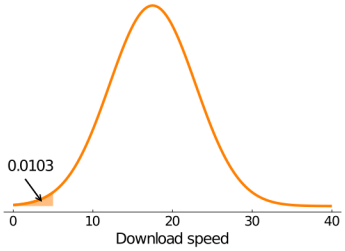


Figure 75: WiFi example. Predictive density.

whatever

- Exchangeability and De Finettis theorem

whatever

Classification

EXERCISES

1. This is the first problem.
2. This is the second problem.

NOTEBOOKS

1. See the notebook Classification.

Posterior simulation

Gibbs sampling

Markov Chain Monte Carlo

Hamiltonian Monte Carlo

Probabilistic programming frameworks

Variational inference

Regularization

Model comparison

Posterior model probabilities

Bayesian cross-validation

- M-completed? - Generalization performance. - WAIC - Cross-validation This is some text [3]

[3] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 3rd edition. CRC press, 2013

L2-regularization and Ridge

L1-regularization and Lasso

Global-local regularization and Horseshoe

Variable selection

Gaussian processes

Gaussian processes

Mixture models

Finite mixtures

Mixtures of regressions

Latent Dirichlet allocation

Infinite mixtures

Bibliography

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml/datasets/Spambase/.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 3rd edition. CRC press, 2013.

Telba Z Irony and Nozer D Singpurwalla. Non-informative priors do not exist - a dialogue with josé m. bernardo. *Journal of Statistical Planning and Inference*, 65(1):159–177, 1997.

Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

Helio S Migon, Dani Gamerman, and Francisco Louzada. *Statistical inference: an integrated approach*. CRC press, 2014.

Rolf Sundberg. *Statistical modelling by exponential families*, volume 12. Cambridge University Press, 2019.

Mattias Villani. Steady-state priors for vector autoregressions. *Journal of Applied Econometrics*, 24(4):630–650, 2009.

Bertil Wegmann and Mattias Villani. Bayesian inference in structural second-price common value auctions. *Journal of Business & Economic Statistics*, 29(3):382–396, 2011.

Index