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, May 2021

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, but may not be heavily trained in probability and statistics.

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 students. 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!

1 The Bayesics

TODO! write proper intro text.

Learning probability models

Throughout this book we will exclusively work with probability models. Probability models have the advantage of giving a precise quantification of uncertainty that can be directly used for decision making in the real world.

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

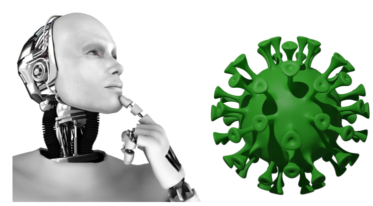

Figure 1.1: Artificial intelligence and infectious disease models are examples where Bayesian learning is often used for quantifying uncertainty.

While the initial chapters focus 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; authorities need to learn the basic reproduction number R_0 in probabilistic models to predict disease spreading and for making decisions about interventions. The Bayesian approach to predictions and decisions will be presented in the chapter Prediction and Decision making, and used in many places throughout the book.

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.

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

$$\Pr(X = x|\theta) = \theta^x(1-\theta)^{1-x}. \tag{1.1}$$

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 1.2.

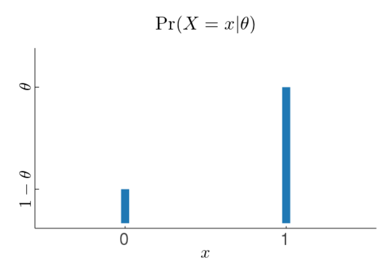

Figure 1.2: Bernoulli distribution with success probability $\theta = 0.8$.

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. If the data are iid we can express the likelihood in terms of the univariate distributions $p(X|\theta)$ as

likelihood function

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta). \tag{1.2}$$

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

$$p(x_1, \ldots, x_n|\theta) = \theta^s(1-\theta)^f, \tag{1.3}$$

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.

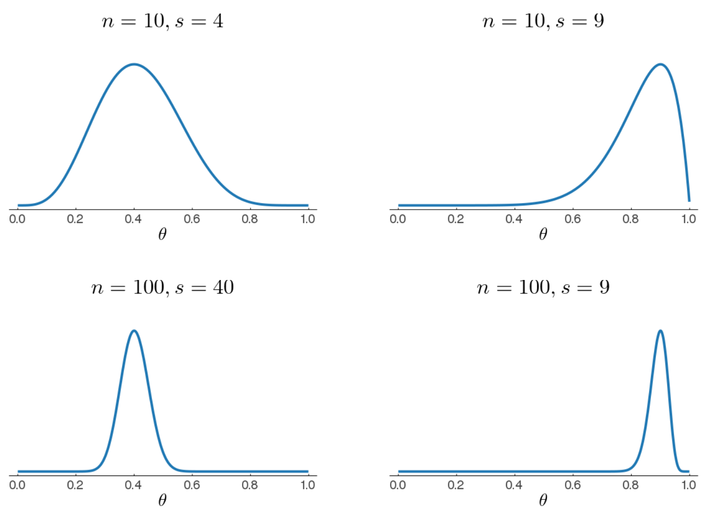

Figure 1.3: Bernoulli likelihood function for $n = 10$ and $s = 4$.

It is essential to have mental image of the likelihood function when thinking about statistical modeling. Figure 1.3 illustrates the likelihood function for Bernoulli model when $s = 4$ successes was obtained in $n = 10$ trials (top left) and when $s = 9$ successes was obtained in $n = 10$ trials (top right). The lower part of Figure 1.3 show results for $n = 100$ trials with the same success ratio s/n as in the upper part of the figure; the larger datasets make the likelihood more concentrated, i.e. more informative regarding the plausibility of different θ values.

Figure 1.3 nicely illustrates how the likelihood function can inform us about the plausibility of any given θ for any given dataset. If we want to select a single value, an **estimate** of θ, a natural candidate is the **maximum likelihood estimator**

estimate

maximum likelihood estimator

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} p(x_1, \ldots, x_n|\theta). \tag{1.4}$$

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 $p(x_1, \dots, x_n|\theta)$ directly with respect to θ it is often easier to maximize the *log-likelihood function*

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

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-likelhood function with respect to θ, setting that derivative to zero and solving for θ. Solving

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

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 $\theta = \hat{\theta}_{\text{MLE}}$.

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

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

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 $X_1, \dots, X_n|\theta \overset{iid}{\sim}$ Bernoulli then $S|\theta \sim \text{Binomial}(n, \theta)$ with mean $E(S) = n\theta$; see Figure 1.5 for an example of a **Binomial distribution**.

Binomial distribution

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

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

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

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

Figure 1.4: The binomial distribution.

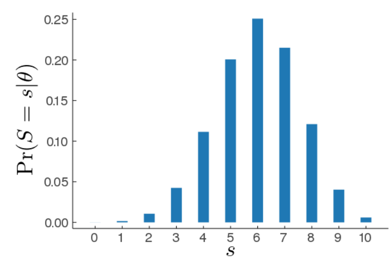

Figure 1.5: Binomial distribution with $n = 10$ and $\theta = 0.7$.

unbiased

Binomial distribution

sampling variance

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

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

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

sampling distribution

long run properties

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 sampling properties such as $\mathbb{E}(\hat{\theta}_{MLE})$ and $\mathbb{V}(\hat{\theta}_{MLE})$ are not used in the Bayesian approach, the likelihood *function* is at the core of Bayesian learning.

The likelihood functions in Figure 1.3 *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 1.6 reminds us of this error.

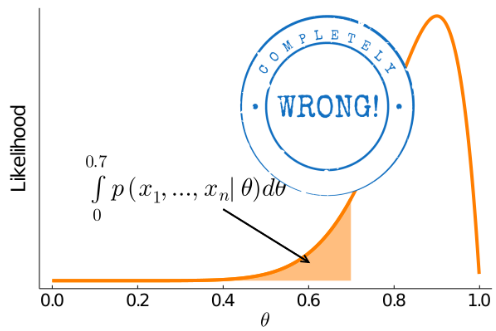

Figure 1.6: Areas under the likelihood function are **not** probabilities.

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.

Figure 1.7: Bruno de Finetti, 1906-1985, a founder of subjective probability.

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

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,

dutch book 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.

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

Bayes' theorem

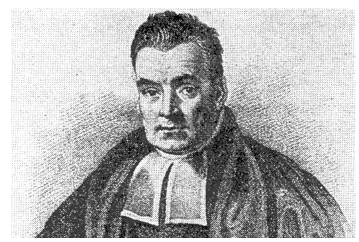

Figure 1.8: Reverend Thomas Bayes, ca 1701-1761, whose famous theorem was published posthumously. Interestingly, we are not quite sure that the man in the photo actually is Thomas Bayes. Probably not.

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

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 $A = \{\text{'Have disease'}\}$ and $B = \{\text{'Positive test'}\}$ and also $\Pr(\mathrm{B}|\mathrm{A}^c) = 0.05$, 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. $\Pr(\mathrm{A}|\mathrm{B})$. 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 $\Pr(\mathrm{A}) = 0.0001$. Given no other information (e.g. that you feel sick or have other symptoms) Bayes' theorem gives

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

where we have expressed $\Pr(\mathrm{B})$ 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

law of total probability

the initial $\Pr(\mathrm{A}) = 0.0001$, the probability of actually having the disease is still tiny. The lesson here is that prior probabilities matter.

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

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

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

$$\Pr(\theta_k|x_1, \ldots, x_n) = \frac{\Pr(x_1, \ldots, x_n|\theta_k)\Pr(\theta_k)}{\sum_{j=1}^{K}\Pr(x_1, \ldots, x_n|\theta_j)\Pr(\theta_j)}. \qquad (1.5)$$

Note how we again used the law of total probability in the denominator to express $\Pr(\mathrm{B}) = \Pr(x_1, \ldots, x_n)$. 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 (1.5) and we will therefore often write Bayes' theorem in proportional form

$$\Pr(\theta_k|x_1, \ldots, x_n) \propto \Pr(x_1, \ldots, x_n|\theta_k)\Pr(\theta_k), \qquad (1.6)$$

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 (1.6) is the *likelihood function* evaluated in the point θ_k. Equation (1.6) therefore expresses the fundamental idea in Bayesian learning:

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

Figure 1.10 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).

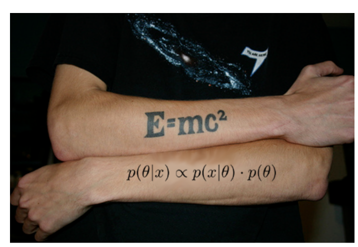

Figure 1.9: Great theorems make great tattoos.

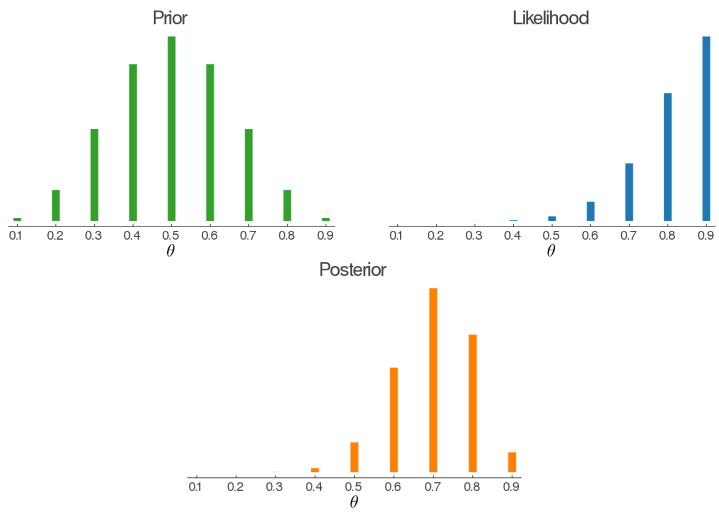

Figure 1.10: Prior, likelihood and posterior for Bernoulli model with $n = 10$ and $s = 9$.

Finally, taking a finer and finer grid in Equation 1.5 we get the following Bayes' theorem for a continuous parameter θ in the limit

$$p(\theta|x_1,\ldots,x_n) = \frac{p(x_1,\ldots,x_n|\theta)p(\theta)}{\int p(x_1,\ldots,x_n|\theta)p(\theta)d\theta}, \tag{1.7}$$

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

prior density

posterior density

$$p(\theta|x_1,\ldots,x_n) \propto p(x_1,\ldots,x_n|\theta)p(\theta). \tag{1.8}$$

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 omelette
> without breaking the Bayesian eggs.
>
> **Leonard Jimmy Savage**

Figure 1.11: Making a Bayesian omelette.

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 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 Noninformative priors where priors are intentionally chosen to be neutral or minimally informative. Section Invariant priors gives an alternative approach to so called objective priors using invariance arguments.

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.

2 *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.

Beta distribution

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

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

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

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

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

Figure 2.1: The beta distribution.

Bernoulli data

Let us return to iid Bernoulli data:

$$x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{Bern}(\theta). \tag{2.1}$$

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 2.2 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.

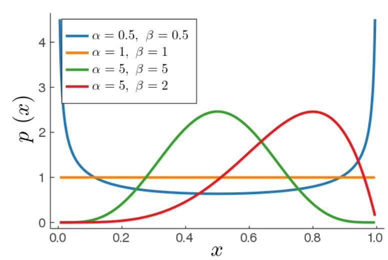

Figure 2.2: Some Beta distributions.

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 2.3: The uniform distribution.

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 (1.8):

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

where $p(x_1, \ldots, x_n | \theta) = \theta^s (1-\theta)^f$ is the likelihood for iid Bernoulli data and $p(\theta)$ is the $\theta \sim \text{Beta}(\alpha, \beta)$ prior. So,

$$p(\theta|x_1, \ldots, x_n) \propto \theta^s (1-\theta)^f \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\text{B}(\alpha,\beta)} \quad (2.2)$$

$$\propto \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}, \quad (2.3)$$

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 (2.2) 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 (2.2) 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 2.4. 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 2.4: 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 $\hat{\theta}_{\mathrm{MLE}} = s/n$. 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 $\mathrm{Beta}(1,1)$ 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 Sections Noninformative priors and Invariant priors.

Bayes estimator

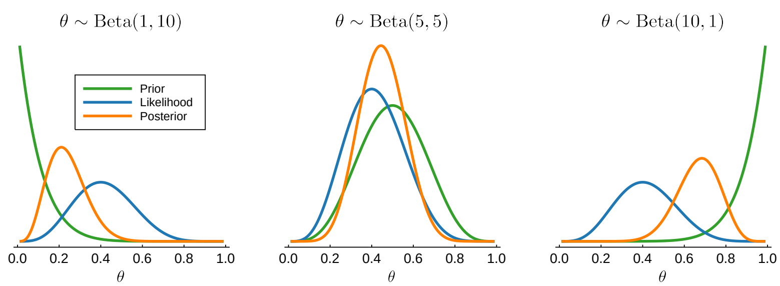

Figure 2.5: Bayesian analysis of $n = 10$ randomly chosen emails from the SpamBase data using three different priors. The likelihood is normalized.

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 $x_1, \ldots, x_n|\theta \overset{\mathrm{iid}}{\sim} \mathrm{Bern}(\theta)$ 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/

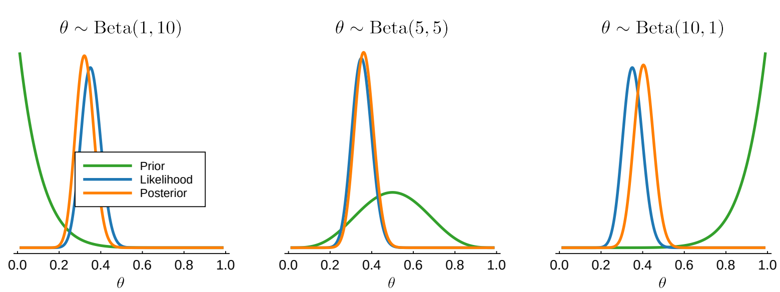

Figure 2.6: Bayesian analysis of $n = 100$ randomly chosen emails from the SpamBase data using three different priors. The likelihood is normalized.

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 2.5 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 2.6 are based on $n = 100$ randomly sampled emails, including the 10 emails used in Figure 2.5. The posteriors are now in rather close but not perfect agreement. Finally, Figure 2.7 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.

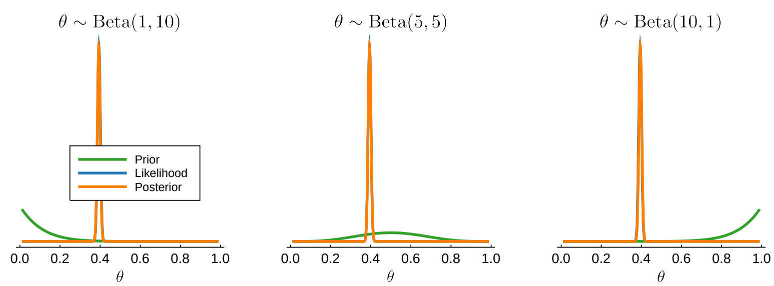

Figure 2.7: Bayesian analysis of all $n = 4601$ emails from the SpamBase data using three different priors. The likelihood is normalized.

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 $X_1, \ldots, X_n | \theta \sim \mathrm{Bern}(\theta)$, where n is a predetermined number of trials.
  Data: the outcome in each trial: $x_1, \ldots, x_n$.
- **Experiment 2**: sample data from $X_1, \ldots, X_n | \theta \sim \mathrm{Bern}(\theta)$, 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 $X_i | \theta \sim \mathrm{Bern}(\theta)$ 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

$$p(x_1, \dots, x_n|\theta) = \theta^s(1-\theta)^{n-s}, \tag{2.4}$$

In the second experiment we only get to observe that there was s successes in n trials, but the exact sequence $x_1, \dots, x_n$ 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

$$p(s) = \binom{n}{s}\theta^s(1-\theta)^{n-s}. \tag{2.5}$$

This is different from the likelihood in Experiment 1 since the outcome $S = s$ can be obtained from several different observed data sequences $x_1, \dots, x_n$, 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

$$p(n) = \binom{n-1}{s-1}\theta^s(1-\theta)^{n-s}. \tag{2.6}$$

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 (2.4)-(2.6) 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 (1.7)

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

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 2.8: The Gaussian distribution.

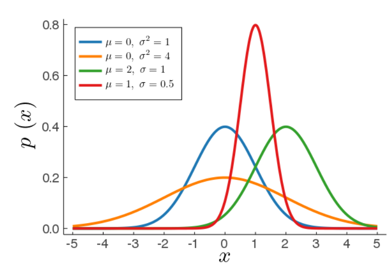

Figure 2.9: Some Normal distributions.

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

$$p(\theta|x_1,\ldots,x_n) \propto \exp\Big(-\frac{n}{2\sigma^2}(\theta-\bar{x})^2\Big), \tag{2.7}$$

and hence that the posterior for θ can be recognized as

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

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 (2.7) we have

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

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 2.10.

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 2.10 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

Conjugate analysis - Gaussian model with known variance

Model: $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$, σ^2 known
Prior: $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$
Posterior: $\theta | x_1, \ldots, x_n \sim \mathrm{N}\left(\mu_n, \tau_n^2\right)$.
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 2.10: Prior-to-Posterior updating for normal data with known variance and normal prior for the mean.

$n/\sigma^2 = 1/(\sigma^2/n)$ 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 2.10 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, $\mathbf{x} = (15.77, 20.5, 8.26, 14.37, 21.09)$, 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 $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$, 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.

Internet speed dataset

Figure 2.11 (left) 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 $\Pr(\theta \geq 20 | x_1, \ldots, x_n) \approx 0.275$

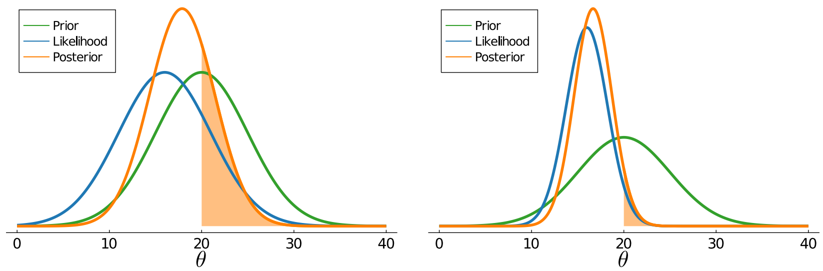

Figure 2.11: Internet speed data. Posterior updating based on $n = 1$ observation (left) and $n = 5$ observations (right). The orange shaded region marks out $\Pr(\theta > 20|x_1, \ldots, x_n)$.

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 2.11 (left) shows the updated posterior using all $n = 5$ data points with $\bar{x} = 16.001$; 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.

Figure 2.11 illustrates a situation where the posterior is computed by combining the prior at day 0, $N(\mu_0, \tau_0^2)$, with the likelihood for all $x_1, \ldots, x_n$ data points; hence the posterior on day n is computed as

$$p(\theta|x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n|\theta)p(\theta). \tag{2.8}$$

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

$$p(\theta|x_1, \ldots, x_n) \propto p(x_n|\theta)p(\theta|x_1, \ldots, x_{n-1}). \tag{2.9}$$

The updating in (2.8) and (2.9) give the same result, but (2.9) 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 2.12 for the internet speed data.

online learning

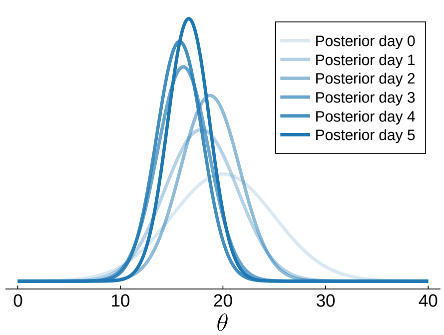

Figure 2.12: Internet speed data. Bayesian online learning.

The same online learning holds also for dependent data, e.g. time series, as is easily proved as follows

$$\begin{aligned} p(\theta|x_1,\dots,x_n) &\propto p(x_1,\dots,x_n|\theta)p(\theta) \\ &= p(x_n|\theta,x_1,\dots,x_{n-1})p(x_1,\dots,x_{n-1}|\theta)p(\theta) \\ &\propto p(x_n|\theta,x_1,\dots,x_{n-1})p(\theta|x_1,\dots,x_{n-1}), \end{aligned} \tag{2.10}$$

where the second line follows from the decomposition results in Figure 2.13. For iid data we have the additional simplication $p(x_n|\theta, x_1, \dots, x_{n-1}) = p(x_n|\theta)$, hence showing the equivalence of (2.8) and (2.9).

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 $1 \le m \le n-1$

batch learning

$$p(\theta|x_1,\dots,x_n) \propto p(x_{m+1},\dots,x_n|\theta)p(\theta|x_1,\dots,x_m). \tag{2.11}$$

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

- any intermediate posterior $p(\theta|x_1,\dots,x_m)$ 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 $p(x_{m+1},\dots,x_n|\theta)$.

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

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

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.

Decomposing distributions

For two random variables X, Y

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

For n random variables

$$\begin{aligned} p(x_1,\dots,x_n) = p(x_1)p(x_2|x_1)\times \\ \dots \times p(x_n|x_1,\dots,x_{n-1}) \end{aligned}$$

and conditional on θ

$$\begin{aligned} p(x_1,\dots,x_n|\theta) = p(x_1|\theta)\times \\ \dots \times p(x_n|x_1,\dots,x_{n-1},\theta) \end{aligned}$$

Figure 2.13: Marginal-Conditional decomposition of a joint distribution.

Poisson distribution

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

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

Figure 2.14: The Poisson distribution.

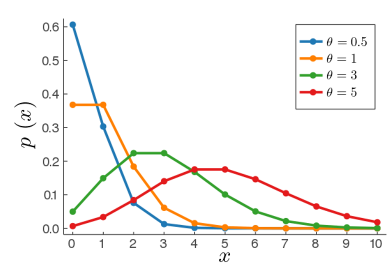

Figure 2.15: Some Poisson distributions.

Poisson data

Poisson distribution

Count data $X \in 0, 1, 2, \dots$ 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 restrictive in some application, but the model often fits many real datasets surprisingly well or can be extended to do so.

Figure 2.16.

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

$$p(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \propto \theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}. \tag{2.12}$$

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

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

Figure 2.16: Gamma distribution.

Comparing the functional form of the likelihood in (2.12) 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:

Gamma distribution

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

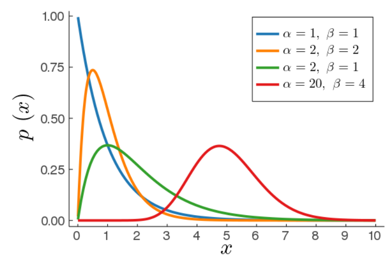

Figure 2.17: Some Gamma distributions.

Figure 2.18: Prior-to-Posterior updating for the Poisson data with a Gamma prior.

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.

eBayCoin dataset

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

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 $\bar{x} = 3.635$

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 2.19. 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 $\theta \in [3.4, 3.9]$.

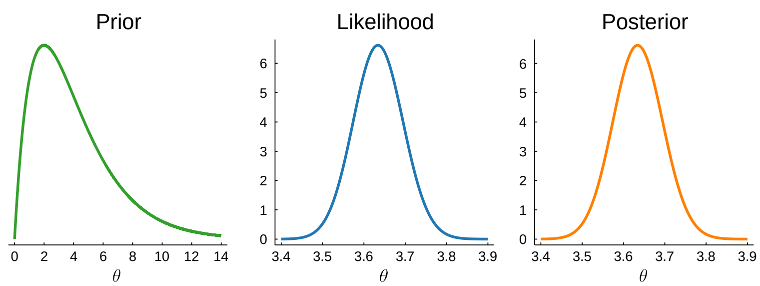

Figure 2.19: Bayesian analysis of the numbers of bidders in $n = 1000$ eBay coin auctions.

Figure 2.20 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.

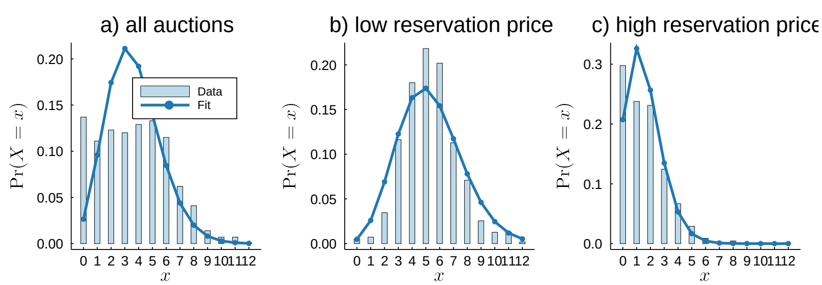

Figure 2.20: Assessing the fit of the Poisson model with the posterior mean estimate of θ.

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 for the auction with low reservation prices is set to $\text{Gamma}(4, 1/2)$ to reflect a belief that belief that such auctions are likely to attract more bids. The prior for the auction where the reservation prices are high is set to $\text{Gamma}(1, 1/2)$. The prior-to-posterior updating is shown in Figure 2.21. The posteriors are clearly different in the two subpopulations. The Poisson model fits better on the two subpopulations as shown in Figure 2.20 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.

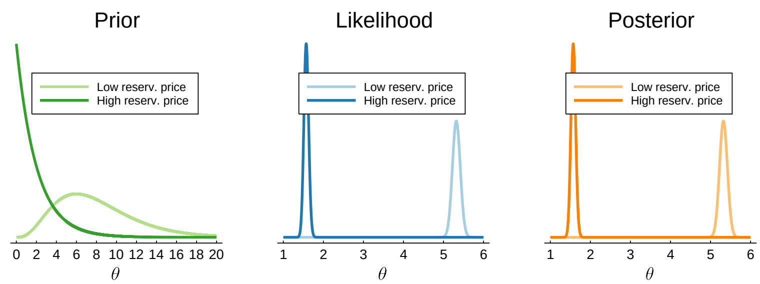

Figure 2.21: eBay auctions. Bayesian analysis of the numbers of bidders in $n = 550$ auctions with a low reservation price and $n = 450$ auctions with a high reservation price.

We have now seen that:

- 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*

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

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

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 $\mathbb{S}(\theta|x_1, \dots, x_n) = \sqrt{\mathbb{V}(\theta|x_1, \dots, x_n)}$.

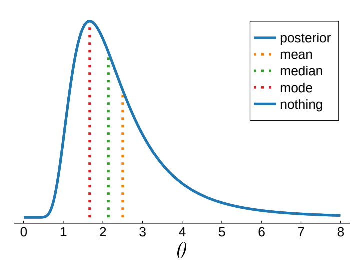

Figure 2.22: Three common point estimates for summarizing a posterior.

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 omelette. 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 2.23 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).

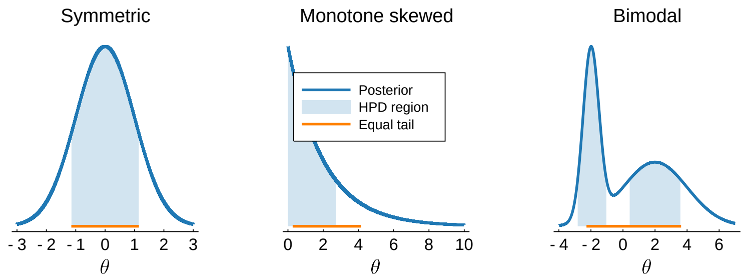

Figure 2.23: Illustration of HPD regions (shaded areas) and equal tail intervals (orange line).

A disadvantage of HPD regions is that they are not invariant to reparametrization: if $[a, b]$ is an HPD region for θ, then $[f(a), f(b)]$ is typically not an HPD region for a transformed parameter $\phi = f(\theta)$ for a non-linear transformation $f(\cdot)$.

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 2.24.

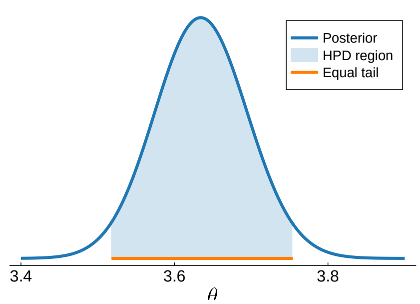

Figure 2.24: 95% credibility intervals for the Gamma posterior in the eBay auction data.

*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.

Sufficient statistic

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.*

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

Factorization criterion

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*

$$p(x_1, \ldots, x_n | \theta) = h(x_1, \ldots, x_n) f(\mathbf{t}(x_1, \ldots, x_n); \theta), \qquad (2.13)$$

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.

Sufficiency 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 $\mathbf{t}(x_1, \ldots, x_n)$.*

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 $\mathbf{t}(x_1, \ldots, x_n)$. $\square$

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

$$p(x|\theta) = h(x) \exp\Big(\eta(\theta)t(x) - A(\theta)\Big), \text{ for } x \in \mathcal{X}, \tag{2.14}$$

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, $\ln\left(\frac{\theta}{1-\theta}\right)$.

Student-t distribution

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

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

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

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

Figure 2.25: The student-t distributions.

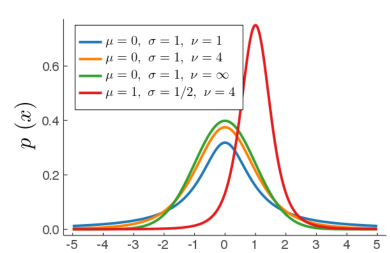

Figure 2.26: Some Student-t distributions.

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.

student-t distribution

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

$$p(x_1, ..., x_n|\theta) = \left[\prod_{i=1}^n h(x_i)\right] \exp\left(\eta(\theta)\sum_{i=1}^n t(x_i) - nA(\theta)\right). \tag{2.15}$$

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 (2.15)

$$p(\theta) = H(\tau_0, \nu_0) \exp\Big(\eta(\theta)\tau_0 - \nu_0 A(\theta)\Big), \tag{2.16}$$

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. 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 $\eta(\theta) = \ln\big(\frac{\theta}{1-\theta}\big)$ and $A(\theta) = \ln\big(\frac{1}{1-\theta}\big)$, for Bernoulli data. The prior in (2.16) 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 (2.16) 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 2.27: 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 (2.15) with prior (2.16)

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

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 2.27.

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.

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 2.28: The exponential distribution.

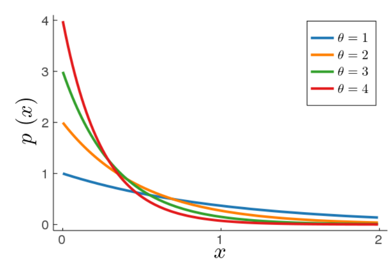

Figure 2.29: Some Exponential distributions.

4. Derive the posterior distribution for the normal model with a normal prior in Figure 2.10. *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.

NOTEBOOKS

1. Analyzing spam data with an iid Bernoulli model.
2. Analysis of internet download speed data using a Gaussian model with known variance.
3. Analyzing number of eBay bidders with a Poisson model.

3 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 3.1 gives an illustration of a bivariate ($K = 2$) normal distribution.

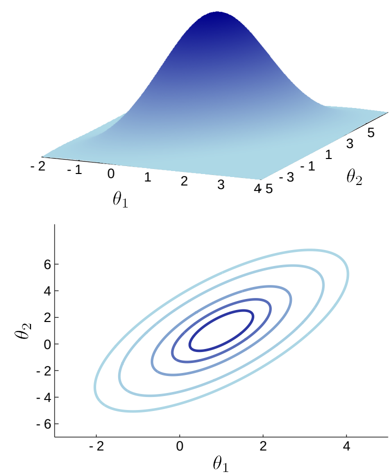

Figure 3.1: Surface and contour plot of the bivariate normal distribution. The contour levels contain 25, 50, 75, 95 and 99% of the probability mass, respectively.

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

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3.1}$$

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 3.2 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

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

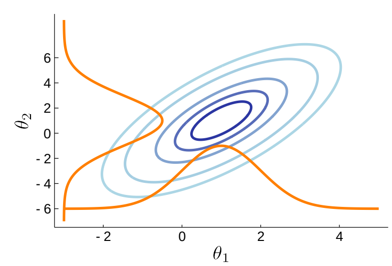

Figure 3.2: Contour plot of the bivariate normal distribution in Figure 3.1 along with the marginal distibutions.

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

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

$$p(\boldsymbol{\theta}_a) = \int \cdots \int p(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b) d\boldsymbol{\theta}_b. \tag{3.2}$$

We will see examples of marginalization in the following sections.

Gaussian data with unknown variance

The previous chapter analyzed iid normal data $x_1, \ldots, x_n|\theta, \sigma^2 \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$ 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

$$\theta|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0) \tag{3.3}$$
$$\sigma^2 \sim \text{Inv}-\chi^2(\nu_0, \sigma_0^2). \tag{3.4}$$

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 3.3. 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".

Inv-χ^2 distribution

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

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

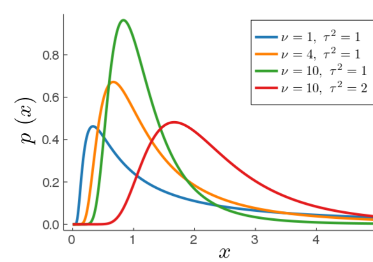

Figure 3.3: Some Scaled-Inv-Gamma distributions.

The conjugate prior in (3.3) 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 3.4 shows that the posterior is indeed in the same form as the prior in (3.3), as required for a conjugate prior. There is a lot of greek letters in Figure 3.4, 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)

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: $\theta | \sigma^2, \mathbf{x} \sim N(\mu_n, \sigma^2/\kappa_n)$
$\sigma^2 | \mathbf{x} \sim \text{Inv}-\chi^2(\nu_n, \sigma_n^2)$
$\mu_n = w\bar{x} + (1-w)\mu_0$
$w = \frac{n}{\kappa_0 + n}$
$\kappa_n = \kappa_0 + n$
$\nu_n = \nu_0 + n$
$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \mu_0)^2$
where $\bar{x} = \sum_{i=1}^n x_i$ and $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$
Marginal: $\theta | \mathbf{x} \sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$

Figure 3.4: 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 $\kappa_n = \kappa_0 + n$ is the sum of the number of prior observations κ_0 and 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 3.4. In Exercise 1 you are ask to show that the marginal posterior of θ is a student-t distribution; see Figure 2.25 and 2.26 for a definition and properties. Specifically, we have the following result

$$\theta | x_1, \ldots, x_n \sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n), \tag{3.5}$$

where μ_n, σ_n^2, κ_n and ν_n are all defined as in Figure 3.4. Note that also the marginal prior for θ follows a student-t distribution of the form (3.5), 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 that also σ^2, the variability of the measurements from the speed testing service, is 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 3.6, which I find sensible.

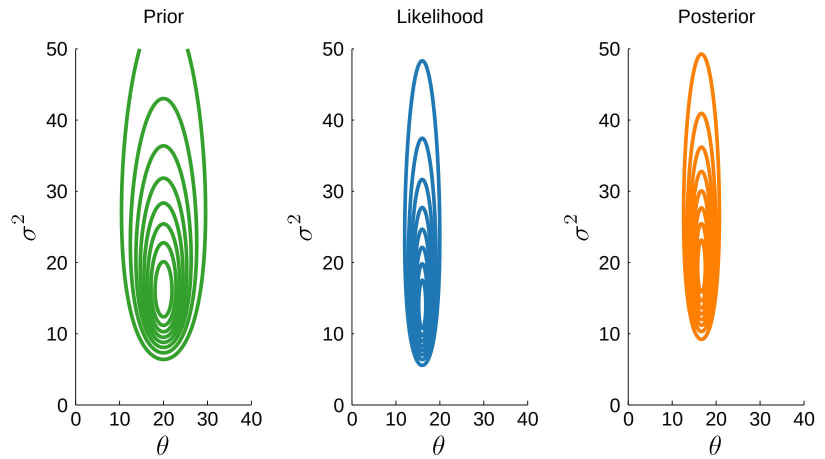

Figure 3.5: Prior-to-Posterior updating for the internet speed data in the iid Normal model. Contours of joint distributions of θ and σ^2.

Figure 3.5 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 3.6. 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.

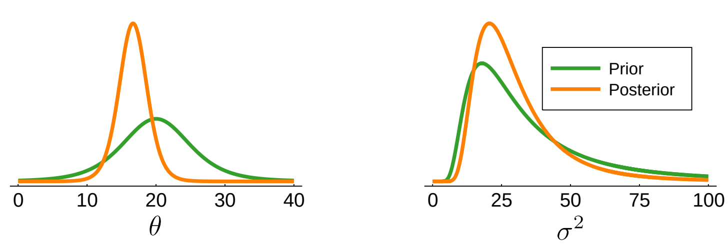

Figure 3.6: Marginal posteriors for the internet speed data in the iid Normal model.

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 posteriors in analytical form. This is rarely 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

Posterior simulation - iid Gaussian with conjugate prior.

```
Input: data x = (x_1,...,x_n)
       number of posterior draws m.
compute μ_n, σ_n^2, κ_n and ν_n using Figure 3.4.
for i in 1:m do
    σ^2 ← rInvChi2(ν_n,σ_n^2)
    θ ← 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 = ντ^2/x
    return y
```

Figure 3.7: 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.

The algorithm in Figure 3.7 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.

EXAMPLE: INTERNET SPEED DATA

Let us now simulate from the posterior of θ and σ^2 in the Internet speed data. The second and third columns in Table 3.1 show the output from generating $m = 10{,}000$ joint posterior draws with the algorithm in Figure 3.7.

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

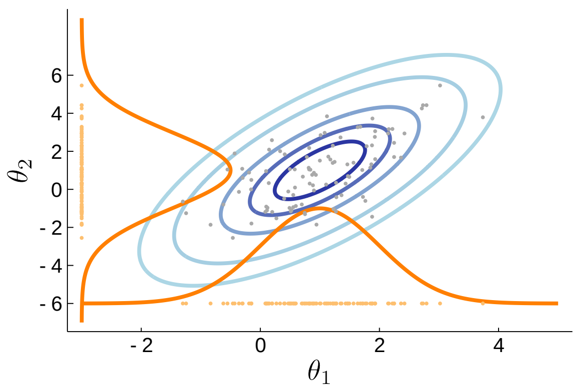

Figure 3.8: Illustrating marginalization by selection. The figure plots the contours of a joint distribution with the marginal distributions overlayed as orange curves. The gray points are 100 draws from joint distribution and the orange points are projections of the gray points on the two axes. The orange points correspond to the draws obtained by selecting out each parameter from the joint simulation and clearly represent the marginal posteriors.

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 integration is replaced by plotting a histogram of the selected column. This is illustrated in Figure 3.8.

Figure 3.11 shows the marginals for the internet speed data example obtained from simulation; the figure also plots the analytical marginal posteriors, which happen to be known in this simple example.

The histograms of the simulated draws in Figure 3.11 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

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

where $\xrightarrow{p}$ denotes **convergence in probability**, see Figure 3.9. 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. This is version of the **law of large numbers**, see Figure 3.10. The left side of Figure 3.12 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

Convergence in probability

A sequence of random variables $X_1, \ldots, X_n$ **converges in probability to a constant** c, if and only if for any $\epsilon > 0$

$\Pr(|X_n - c| > \epsilon) \to 0$ as $n \to \infty$.

We then write $X_n \xrightarrow{p} c$.

$X_1, \ldots, X_n$ **converges in probability to a random variable** X if and only if for any $\epsilon > 0$

$\Pr(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

We write $X_n \xrightarrow{p} X$.

Figure 3.9: Convergence in probability.

Law of large numbers

Let $X_1, X_2, \ldots$ be iid random variables with finite mean μ. Then

$$\bar{X}_n \xrightarrow{p} \mu \text{ as } n \to \infty,$$

where $\xrightarrow{p}$ denotes convergence in probability.

There is also a strong law of large numbers based on an alternative notion of probabilistic convergence called **almost sure convergence**, as well as laws for variables that are not iid.

Figure 3.10: Weak law of large numbers.

simulation consistent

$m = 1000$.

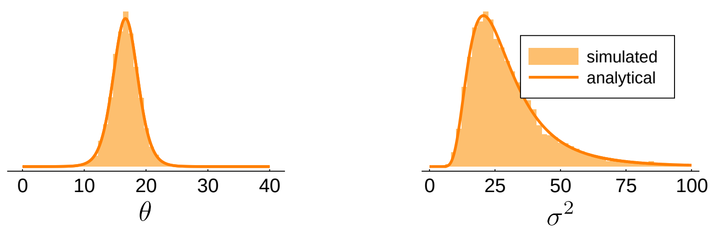

Figure 3.11: Histogram of simulated marginal posteriors for the internet speed data with analytical marginal posterior densities overlayed.

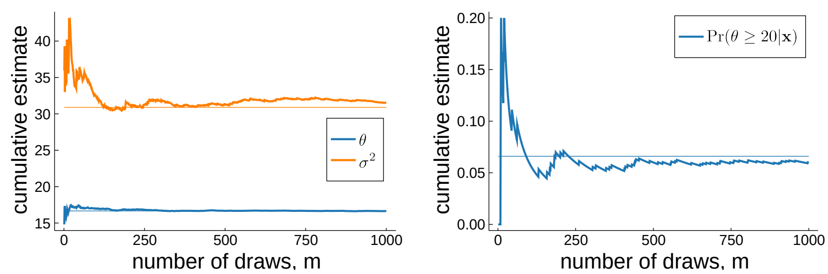

Figure 3.12: Convergence of the Monte Carlo estimate of the posterior expectation of θ and σ^2 (left) and $\Pr(\theta \geq 20|\mathbf{x})$ (right). The analytical posterior results are displayed as thin horizontal lines.

The **central limit theorem** (Figure 3.14) can be used to prove that $\bar{\theta}_{1:m}$ **converges in distribution** (Figure 3.13) to a normal distribution. Hence, the following approximation of the posterior estimate $\bar{\theta}_{1:m}$ is accurate when m is large:

$$\bar{\theta}_{1:m} \sim N\left(\mathbb{E}(\theta|\mathbf{x}), \frac{\mathbb{V}(\theta|\mathbf{x})}{m}\right), \tag{3.6}$$

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 (3.6) can be used to determine the required number of draws m needed for a given estimation precision. A multivariate version of the central limit theorem can be used to prove a similar result to (3.6) when $\boldsymbol{\theta}$ is a vector; an interesting aspect is that $\mathrm{Cov}(\bar{\boldsymbol{\theta}}_{1:m})$ (a covariance matrix in the multiparameter case) still decreases at the rate $1/m$, regardless of the dimension of $\boldsymbol{\theta}$.

Convergence in distribution

A sequence of random variables $X_1, \ldots, X_n$ **converges in distribution** to the random variable X, if and only if

$$F_n(x) \to F(x) \text{ as } n \to \infty,$$

for all x where $F(\cdot)$ is continuous, where $F_n(x)$ and $F(x)$ are the cumulative distribution function (CDF) of X_n and X, respectively.

We then write $X_n \xrightarrow{d} X$.

Figure 3.13: Convergence in distribution.

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

coefficient of variation

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 (3.6) exists also in this case, with the expected value and variance replaced by those of $f(\boldsymbol{\theta})$.

To illustrate how simulation immediately provides inference for any function of the parameters, Table 3.1 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 3.15; the left part of the figure shows the results for the standard deviation $f(\theta, \sigma^2) = \sqrt{\sigma^2}$.

Central limit theorem (CLT)

Let $X_1, X_2, \ldots$ be iid random variables with finite mean μ and variance σ^2. Then

$$\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow{d} \mathrm{N}(0,1),$$

as $n \to \infty$ where $\xrightarrow{d}$ denotes convergence in distribution.

The CLT is often informally written as

$$\bar{X}_n \xrightarrow{d} \mathrm{N}(\mu, \sigma^2) \text{ as } n \to \infty.$$

Figure 3.14: The central limit theorem.

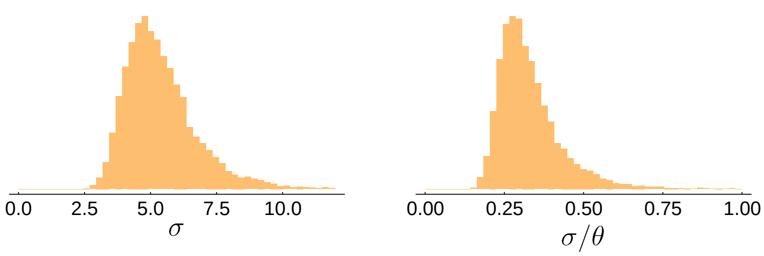

Figure 3.15: Histogram of simulated marginal posteriors for σ (left) and the coefficient of variation σ/θ (right) for the internet speed data.

The final column of Table 3.1 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; the right side of Figure 3.12 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 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.

Categorical data

ordinal data

multi-class

A multi-class random variable X is often written in **one-hot encoding** as $\mathbf{x} = (x_1, \ldots x_C)$ where $X = c$ is encoded as $x_c = 1$ and

one-hot encoding

$x_j = 0$ 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 $X|\boldsymbol{\theta} \sim \text{Cat}(\theta_1, \ldots, \theta_C)$ has probability distribution

$$p(x) = \theta_1^{x_1} \cdots \theta_C^{x_C}, \tag{3.7}$$

where $(x_1, \ldots x_C)$ 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 $\text{Cat}(\theta_1, \ldots, \theta_C)$ distribution generalizes the Bernoulli distribution to the case $C > 2$. Figure 3.16 is an example of $\text{Cat}(\theta_1, \ldots, \theta_C)$ for $C = 4$.

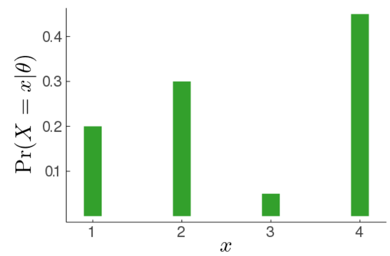

Figure 3.16: Categorical distribution with probabilities $\boldsymbol{\theta} = (0.20, 0.30, 0.05, 0.45)$.

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 $c = 1, \ldots, C$ in multi-class data. This gives data as a count vector $\mathbf{y} = (y_1, \ldots, y_C)$ where y_c is the number of observations in category c in $n = \sum_{c=1}^{C} y_c$ 'trials'. Here is an example:

Mobile phone survey data. A survey was conducted among $n = 513$ mobile phone users. Among other questions, the participants were asked: 'What kind of mobile phone do you mainly use?' with the four options:

1. iPhone
2. Android
3. Windows
4. 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 3.17. The Binomial distribution with x successes in n trials with probability θ in Figure 1.4 is the special case with $C = 2$ categories, which is seen by defining $\theta_1 = \theta$, $\theta_2 = 1 - \theta$, $y_1 = x$, $y_2 = n - x$, and noting that

multinomial distribution

$$\frac{n!}{y_1! y_2!} = \frac{n!}{x!(n-x)!} = \binom{n}{x}. \tag{3.8}$$

Multinomial distribution

$(Y_1, \ldots, Y_C) \sim \text{MultiNom}(n, \boldsymbol{\theta})$ where $\sum_{c=1}^{C} Y_c = n$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$ and $\sum_c \theta_c = 1$.

$$p(\mathbf{y}) = \frac{n!}{y_1! \cdots y_C!} \theta_1^{y_1} \cdots \theta_C^{y_C}$$

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

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

Figure 3.17: The multinomial distribution.

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 mobile phone 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

$$(y_1, y_2, y_3 + y_4) \sim \text{Multinomial}(\theta_1, \theta_2, \theta_3 + \theta_4).$$

In particular, merging to only two categories - for example 'iPhone' and 'not iPhone' - gives a binomial distribution where the probability of success (iPhone) is θ_1 and 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 3.18.

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 $\alpha_c > 0$..

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

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

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

$$\mathbb{V}(\theta_c) = \frac{\mathbb{E}(\theta_c)(1 - \mathbb{E}(\theta_c))}{1 + \alpha_+}$$

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

Marginal distributions:

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

Figure 3.18: The Dirichlet distribution.

unit simplex

The Dirichlet distribution is specified with the prior hyperparameters $\alpha_c > 0$, see Figure 3.19 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 left graph of Figure 3.19, 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 3.18; hence, the prior hyperparameters $\boldsymbol{\alpha} = (1.5, \ldots, 1.5)$ and $\boldsymbol{\alpha} = (5, \ldots, 5)$ in the upper part of Figure 3.19 have the same mean, but the latter has smaller variance. Finally, the bottom part of Figure 3.19 shows examples where the prior mean is different over the categories.

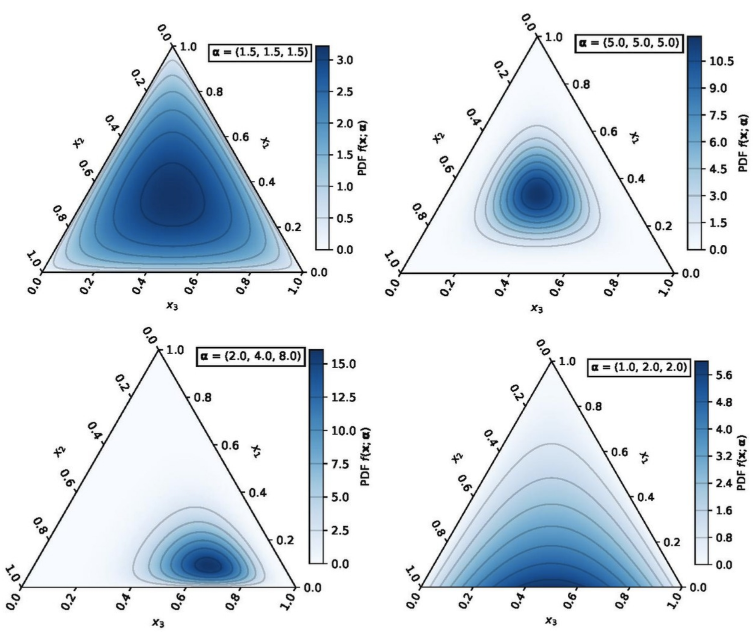

Figure 3.19: Examples of Dirichlet distributions for $\mathbf{x} = (x_1, x_2, x_3)$. Source: Wikipedia.

The Dirichlet$(1, \ldots, 1)$ has constant density and is therefore the

uniform distribution on the unit simplex; this generalizes the result that $\text{Beta}(1,1)$ 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.

uniform distribution on the unit simplex

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

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{3.9}$$

$$= \frac{n!}{y_1! \cdots y_C!}\theta_1^{y_1} \cdots \theta_C^{y_C} \cdot \frac{\Gamma(\sum_{c=1}^{C} \alpha_c)}{\prod_{c=1}^{C} \Gamma(\alpha_c - 1)}\theta_1^{\alpha_1 - 1} \cdots \theta_C^{\alpha_C - 1} \tag{3.10}$$

$$= \theta_1^{\alpha_1 + y_1 - 1} \cdots \theta_C^{\alpha_C + y_C - 1}, \tag{3.11}$$

which is proportional to the $\text{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_C + y_C)$ density. This is a convenient result: the posterior is simply obtained by adding the data count y_c to the prior hyperparameter α_c in each category. This parallels and generalizes the binary case where a $\text{Beta}(\alpha, \beta)$ prior was updated to a posterior by adding the number of successes s to α and the number of failures f to β. Figure 3.20 summarizes the prior-to-posterior updating for multinomial data with a Dirichlet prior.

Multinomial data with Dirichlet prior

Model:	$\mathbf{y}\vert\boldsymbol{\theta} \sim \text{Multinomial}(\boldsymbol{\theta})$, where
	$\mathbf{y} = (y_1, \ldots, y_C)$ are counts in C categories
	$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$ are category probabilities.
Prior:	$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, for $\alpha = (\alpha_1, \ldots, \alpha_C)$
Posterior:	$\boldsymbol{\theta}\vert\mathbf{y} \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{y})$

Figure 3.20: Prior-to-Posterior updating for multinomial data with the Dirichlet prior.

Mobile phone survey data We are now ready to analyze the four market shares $\theta_1, \ldots, \theta_4$ in the mobile phone 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.

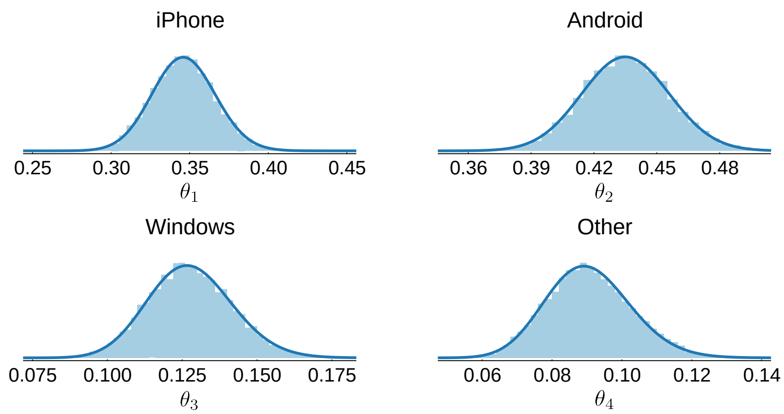

Figure 3.21: Marginal posteriors of the market shares for the mobile phone survey data. Simulated (histogram) draws and analytical density functions (solid curves).

The joint posterior distribution of all four shares is by Figure 3.20 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 3.21 as histograms from Monte Carlo simulation (see the algorithm in Figure 3.22); the analytical posteriors from Figure 3.18 are overlayed.

draw	θ_1	θ_2	θ_3	θ_4	θ_2 largest
1	0.338	0.446	0.130	0.086	1
2	0.332	0.457	0.124	0.086	1
3	0.325	0.442	0.136	0.094	1
⋮	⋮	⋮	⋮	⋮	⋮
10,000	0.343	0.443	0.132	0.081	1
Mean	0.346	0.435	0.127	0.090	0.991

Table 3.2: Posterior simulation output for the multinomial model applied to the mobile phone survey data. The last column is a computed binary indicator for the event that Android has the largest market share, i.e. if $\theta_2 > \max(\theta_1, \theta_3, \theta_4)$.

Figure 3.21 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 $\boldsymbol{\theta}|\mathbf{y} \sim \text{Dirichlet}(\boldsymbol{\alpha}+\mathbf{y})$ over the region $\{\boldsymbol{\theta} : \theta_2 > \max(\theta_1, \theta_3, \theta_4)\}$, a tedious calculation. The probability is however easily computed by simulation by recording for each posterior $\boldsymbol{\theta}$ draw if the condition $\theta_2 > \max(\theta_1, \theta_3, \theta_4)$ is satisfied; see Table 3.2, which shows that

$$\Pr(\text{Andriod has largest market share}|\mathbf{y}) \approx 0.991.$$

We are almost certain that Android is the most popular mobile phone in the population targeted by the survey.

Multivariate normal data with known covariance

This section considers the iid **multivariate normal** model for a p-

multivariate normal

Posterior simulation - Multinomial data, Dirichlet prior.

Input: data $\mathbf{y} = (y_1, \ldots, y_C)$
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$(\boldsymbol{\alpha} + \mathbf{y})$
end
Output: m posterior draws of $\boldsymbol{\theta} = (\theta_1, ..., \theta_C)$.

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

Figure 3.22: 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.

dimensional data vector $\mathbf{x}$:

$$\mathbf{x}_1, \ldots \mathbf{x}_n | \boldsymbol{\theta}, \boldsymbol{\Sigma} \overset{\text{iid}}{\sim} N(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \tag{3.12}$$

where $\boldsymbol{\theta}$ is the p-dimensional mean vector and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite covariance matrix. We will here take $\boldsymbol{\Sigma}$ to be known and derive the posterior for $\boldsymbol{\theta}$.

Presenting a Bayesian analysis of this model here gives us a chance to meet the important multivariate normal distribution and its properties relatively early in the book; see Figure 3.23 for the density and properties, and Figure 3.24 for contour plots of some example densities.

The likelihood for the multivariate model in (3.12) is the product of the individual densities for each vector observation $\mathbf{x}_i$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n | \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\theta})\right),$$

A vector version of the argument leading up to (2.7) in the univariate case can be used to show that the likelihood can be written as the exponential of a quadratic (form):

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n | \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \exp\left(-\frac{n}{2}(\boldsymbol{\theta} - \bar{x})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \bar{x})\right), \tag{3.13}$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ is the usual sample mean vector.

Not too surprisingly, the multivariate normal prior

$$\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0),$$

Multivariate normal

$\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathbf{x} \in \mathbb{R}^p$, $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite covariance matrix.

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

$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$
$\mathbb{V}(\mathbf{x}) = \boldsymbol{\Sigma}$

Define the decomposition

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$$

and similarly for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

Marginal distributions:

$x_k \sim \mathrm{N}(\mu_k, \sigma_k^2)$
$\mathbf{x}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$

Conditional distributions:

$\mathbf{x}_1 | \mathbf{x}_2 \sim \mathrm{N}(\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$

where

$\tilde{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$
$\tilde{\boldsymbol{\Sigma}}_1 = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$

Figure 3.23: The multivariate normal distribution.

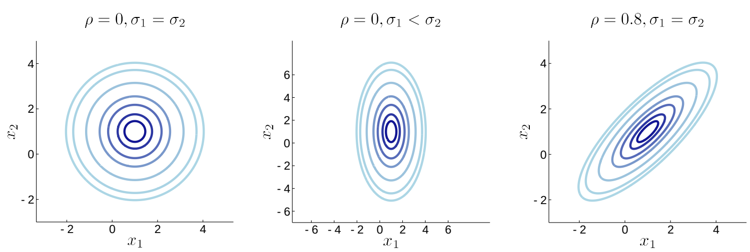

Figure 3.24: Contour plots of some bivariate normal distributions with correlation ρ.

turns out to be conjugate for this model. The posterior can be derived by multiplying together the likelihood in (3.13) with the prior and completing the quadratic forms in the exponentials; see Figure 3.25 for a general result on quadratic form completion. The posterior can then be shown to indeed be a multivariate normal:

$$\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathrm{N}(\boldsymbol{\theta}_n, \boldsymbol{\Lambda}_n),$$

where

$$\begin{aligned} \boldsymbol{\theta}_n &= (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\theta}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}) \\ \boldsymbol{\Lambda}_n^{-1} &= \boldsymbol{\Lambda}_0{}^{-1} + n\boldsymbol{\Sigma}^{-1}. \end{aligned}$$

Letting $\boldsymbol{\Lambda}_0^{-1} \to \mathbf{0}$ (in a matrix sense) we obtain a noninformative (uniform) prior and the posterior

$$\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathrm{N}\big(\bar{\mathbf{x}}, n^{-1}\boldsymbol{\Sigma}\big).$$

Completing quadratic forms

This formula shows how to combine two quadratic forms in a vector of interest $\mathbf{x}$, to a single quadratic form in $\mathbf{x}$ plus constant terms:

$$\begin{aligned} &(\mathbf{x}-\mathbf{a})^\top \mathbf{A}(\mathbf{x}-\mathbf{a}) + (\mathbf{x}-\mathbf{b})^\top \mathbf{B}(\mathbf{x}-\mathbf{b}) \\ &\quad = (\mathbf{x}-\mathbf{d})^\top \mathbf{D}(\mathbf{x}-\mathbf{d}) \\ &\quad + (\mathbf{d}-\mathbf{a})^\top \mathbf{A}(\mathbf{d}-\mathbf{a}) \\ &\quad + (\mathbf{d}-\mathbf{b})^\top \mathbf{B}(\mathbf{d}-\mathbf{b}), \end{aligned}$$

where

$\mathbf{D} = \mathbf{A} + \mathbf{B}$ and $\mathbf{d} = \mathbf{D}^{-1}(\mathbf{A}\mathbf{a} + \mathbf{B}\mathbf{b})$.

Figure 3.25: Completing quadratic forms.

Likelihood and Information

We will end this chapter by defining some useful measures of how much information a dataset carries about the parameters in a model. Recall from the spam data example in Chapter Single-parameter models that the likelihood became more and more peaked around the maximum likelihood estimate (MLE) as the sample size increased. This suggest that the information in a dataset can be measured by how peaked the likelihood is around its mode. The following formalizes this idea.

A Taylor expansion of the log likelihood around the MLE $\hat{\theta}$ gives

$$\begin{aligned} \ln p(\mathbf{x}|\theta) = \ln p(\mathbf{x}|\hat{\theta}) + \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}}(\theta - \hat{\theta}) \\ + \frac{1}{2!}\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}}(\theta - \hat{\theta})^2 + \ldots \end{aligned}$$

The high order terms indicated by $\ldots$ can be shown to be small in large samples. From the definition of the MLE we know that

$$\frac{\partial \ln p(\theta|\mathbf{x})}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0$$

We therefore have the following approximation of the likelihood in large samples

$$p(\mathbf{x}|\theta) \approx p(\mathbf{x}|\hat{\theta}) \exp\left(-\frac{1}{2} J_{\mathbf{x}}(\hat{\theta})(\theta - \hat{\theta})^2\right)$$

where

$$J_{\mathbf{x}}(\tilde{\theta}) = -\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}}.$$

Hence, the likelihood function will be proportional to the $\mathrm{N}[\hat{\theta}, J_{\mathbf{x}}^{-1}(\hat{\theta})]$ density in large samples. The quantity $J_{\mathbf{x}}(\tilde{\theta})$ is clearly the precision in the likelihood and is a natural measure of the information in the data $\mathbf{x}$ about the parameter θ:

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

observed information

$$J_{\theta,\mathbf{x}} = -\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}_{\mathrm{MLE}}} \tag{3.14}$$

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

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

Fisher information

$$I(\theta) = \mathbb{E}_{\mathbf{x}|\theta}\left(J_{\theta,\mathbf{x}}\right). \tag{3.15}$$

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

Definition (Observed information - multiparameter 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

$$J_{\boldsymbol{\theta},\mathbf{x}} = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}}, \tag{3.16}$$

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

Definition (Fisher information - multiparameter case)**.** *The **Fisher information matrix** is the expected information matrix over all possible samples from the model*

Fisher information matrix

$$I(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}}\left(J_{\boldsymbol{\theta},\mathbf{x}}\right). \tag{3.17}$$

The matrix $\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$ in (3.16) may be a little intimidating. Writing out its elements explicitly in the case of two parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2)$,

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

we see that calculating $\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$ 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.

EXERCISES

1. Derive the marginal posterior of θ in (3.5) 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. Analyzing mobile phone survey data with a multinomial model.

4 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 fuse information from a variety of different sources. This chapter discusses different types of prior information and how they can be combined in a given model. We will return to the issue of prior elicitation in later chapters when we perform more serious modelling.

There are 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. This can be the case when there is not enough time or effort to carefully determine a prior; we therefore want to make sure that the prior is not greatly affecting the results. Another situation where a noninformative prior may be desired is when reporting 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 are only weakly informative. The 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 this topic 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

time series

stochastic process

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 4.1 shows a time series of Swedish inflation, Figure 4.2 display the daily number of rides with a bike sharing company, and Figure 4.3 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 4.4, or meteorological data collected at different geographical locations.

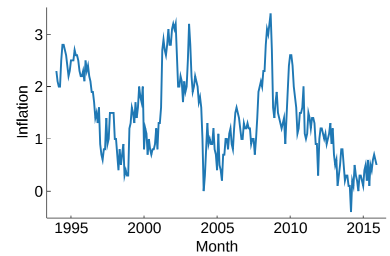

Figure 4.1: Swedish inflation 1995-2016 - annualized monthly observations.

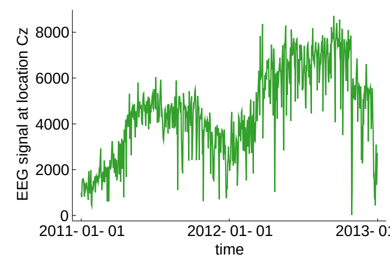

Figure 4.2: Daily number of rides with a bike sharing company.

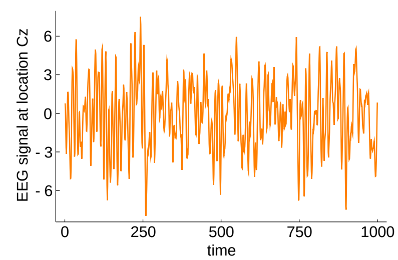

Figure 4.3: EEG recordings of electrical activity at one brain scalp location.

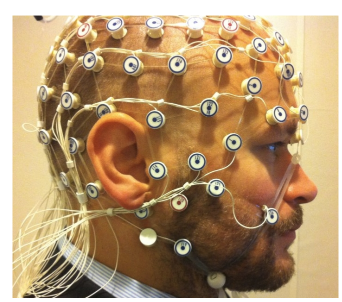
Figure 4.4: Positioning of EEG electrodes on a subject's brain scalp.

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

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), \quad (4.1)$$

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 $\mathrm{AR}(p)$ process models today's value y_t as a linear function of the values 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 $\mathrm{AR}(p)$ process in (4.1) is in **steady-state form** where the parameter μ is the unconditional mean $\mathbb{E}(y_t)$ of the process. We assume that the $\mathrm{AR}(p)$ process is **stationary**, meaning that the mean $\mathbb{E}(y_t)$ and variance $\mathbb{V}(y_t)$ remain unchanged over time. Moreover, the covariance between any two time points $\mathrm{Cov}(y_t, y_s)$ in a stationary process is fully determined by the time distance $|t - s|$ between the observations. The assumption of a constant mean may seem restrictive, but this 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 4.5 where an $\mathrm{AR}(1)$ 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 joint prior distribution for all model parameters $p(\mu, \phi_1, \ldots, \phi_p, \sigma^2)$. 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

autoregressive model

lagged value

steady-state form

stationary

used in applications. We will walk through a number of methods for prior elicitation and use different methods for different parameters.

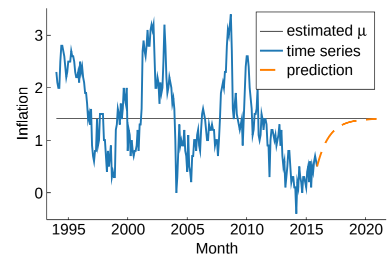

Figure 4.5: Swedish inflation 1995-2016 with 60 months ahead mean prediction in dashed orange.

Past or other data

past 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 using 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 to reflect that the estimates from older data are not necessarily close to the estimates on new data; this is similar to how an older survey was used for the Dirichlet prior in the mobile phone survey data.

other 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.

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 4.5. 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 unfortunately 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 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 on $\phi_1, \ldots, \phi_p$ 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:

Regularization priors

lag length

$$\phi_k \sim \mathrm{N}\left(\mu_k, \frac{\tau^2}{k^2}\right), \tag{4.2}$$

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 $\phi_1 = 0.8$, a reasonable prior guess in the case of 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 means; 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, that is for larger k. Since the prior means for ϕ_k are zero for $k > 1$, this means that longer lags are more heavily shrunk toward zero. The idea here is that 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

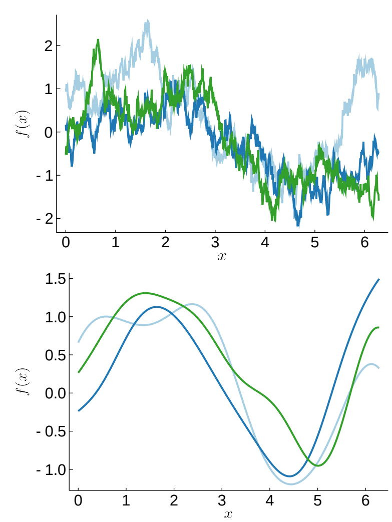

Figure 4.6: Three simulated draws from a prior over functions without smoothness beliefs (top) and with smooth beliefs (bottom).

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 function $f(x)$. Rather than assuming a restrictive functional form, most commonly linear, 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 4.6 shows examples of priors for function with wiggly and smooth beliefs. 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. Note however that a regularization prior still represents subjective

smoothness beliefs

beliefs; my prior beliefs regarding the function $f(\cdot)$ puts higher prior probability on the smooth functions in the bottom part of Figure 4.6 than on the wiggly functions shown in the top part of the figure. This then *implies* a posterior that favors smoother functions, unless the data strongly suggest otherwise.

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. The joint posterior $p(\mu, \phi_1, \ldots, \phi_p, \sigma^2, \tau^2|\mathbf{y})$ involves the now unknown 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, as will be demonstrated in a similar context in the chapter Regularization.

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 2.27 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 4.7 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. The next section presents 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 4.7: 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*

Jeffreys' prior

$$p(\boldsymbol{\theta}) = |I(\boldsymbol{\theta})|^{1/2}. \tag{4.3}$$

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

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 4.7

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

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

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

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

$$p(\theta) = I(\theta)^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto \text{Beta}(1/2, 1/2). \tag{4.4}$$

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 $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$. 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 4.7 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 $\text{Beta}(1/2, 1/2)$ 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

reference prior

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.

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 $\text{Beta}(1/2, 1/2)$ 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.
3. Let $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{Expon}(\theta)$.
   a) Show that Jeffreys prior is $p(\theta) \propto 1/\theta$. Is it proper?
   b) Derive the posterior of θ for Jeffreys prior. Is it proper?
   c) Motivate the particular form of the Jeffreys prior as non-informative.

NOTEBOOKS

1. See the notebook priors.

5 *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

Gaussian linear regression model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2) \qquad \text{for } i = 1, \ldots, n, \tag{5.1}$$

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 first element of each $\mathbf{x}_i$ is typically 1 so that β_1 is the **intercept** term; the intercept β_1 is, rather confusingly, called the **bias** in machine learning. Finally, the model is said to be **homoscedastic** since the error variance σ^2 is the same (homo means same or identical in Greek) for all observations. The case with heteroscedastic errors, $\mathbb{V}(\varepsilon_i) = \sigma_i^2$, will be presented later in the book.

regression coefficients

weights

intercept

bias

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 I_n), \tag{5.2}$$

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 distribution

$$\mathbf{y}|\boldsymbol{\beta},\sigma^2,\mathbf{X} \sim \mathrm{N}(\mathbf{X}\beta,\sigma^2 I_n), \tag{5.3}$$

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**

least squares estimator

residuals

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

When the errors are homoscedastic Gaussian, $\hat{\boldsymbol{\beta}}$ is also the MLE since the log-likelihood from (5.3) is a constant plus $-(1/2\sigma^2)Q(\boldsymbol{\beta})$; hence, minimizing the sum of squared residuals $Q(\boldsymbol{\beta})$ is the same as maximizing the likelihood $p(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\mathbf{X})$.

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 5.1 with $\boldsymbol{\mu} = \mathbf{X}\beta$, $\Sigma = \sigma^2 I_n$ and $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}$ to obtain

$$\hat{\boldsymbol{\beta}}|\boldsymbol{\beta},\sigma^2,\mathbf{X} \sim \mathrm{N}\big(\boldsymbol{\beta},\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big). \tag{5.4}$$

Linear transformation of Gaussians

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

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

Particularly, for $m = 1$ and $\mathbf{A} = (a_1,\ldots,a_p)^\top$, a row vector, we get that linear combinations $\sum_{j=1}^{p} a_j x_j$ of Gaussian variables are Gaussian.

Figure 5.1: Linear transformation of Gaussians.

The result in (5.4) shows that the MLE is unbiased for $\boldsymbol{\beta}$, that is, $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. 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

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) \propto N(\mathbf{y}|\mathbf{X}\beta, \sigma^2 I_n) \cdot \frac{1}{\sigma^2} \\ &= |2\pi\sigma^2 I_n|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \cdot \frac{1}{\sigma^2}, \end{aligned} \tag{5.5}$$

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

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), \tag{5.6}$$

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 (5.6) in (5.5) we obtain the posterior

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-(n+2)/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right\} \tag{5.7}$$

$$\cdot \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\} \tag{5.8}$$

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 (5.7) 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 (5.7)

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

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

$$p(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-[1+(n-p)/2]} \exp\left\{-\frac{1}{2\sigma^2}(n-p)s^2\right\}, \qquad (5.9)$$

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

Gaussian linear regression with non-informative prior

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \varepsilon \overset{iid}{\sim} \text{N}(0, \sigma^2 I_n)$

Prior: $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$

Posterior: $\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim \text{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$
$\sigma^2|\mathbf{y}, \mathbf{X} \sim \text{Inv}-\chi^2(n-p, s^2)$

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

We summarize the prior-to-posterior updating in Gaussian linear regression with a noninformative prior in Figure 5.2.

Conjugate prior

Let us now turn to the more interesting case with a conjugate prior for the Gaussian linear regression. Recall that the conjugate prior for the iid Normal model $x_1, \ldots, x_n|\theta, \sigma^2 \sim \text{N}(\theta, \sigma^2)$ was of the form $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$ where

$$\begin{aligned} \theta|\sigma^2 &\sim \text{N}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \text{Inv}-\chi^2(\nu_0, \sigma_0^2). \end{aligned}$$

The conjugate prior in linear regression is very similar

$$\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\mu}_0, \sigma^2\Omega_0^{-1}) \qquad (5.10)$$

$$\sigma^2 \sim \text{Inv}-\chi^2(\nu_0, \sigma_0^2), \qquad (5.11)$$

with the prior sample size κ_0 replaced by the $p \times p$ precision matrix Ω_0.

A detailed elicitation of the matrix Ω_0 can be demanding. One simple choice is $\Omega_0 = \kappa_0 \boldsymbol{I}_p$ which corresponds to using the same

Gaussian linear regression with conjugate prior

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

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: $\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}_n, \sigma^2\Omega_n^{-1})$
$\sigma^2|\mathbf{y},\mathbf{X} \sim \mathrm{Inv}-\chi^2(\nu_n, \sigma_n^2)$
$\boldsymbol{\beta}|\mathbf{y} \sim t_{\nu_n}(\mu_n, \sigma_n^2\Omega_n^{-1})$

where
$\boldsymbol{\mu}_n = \Omega_n^{-1}(\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} + \Omega_0\boldsymbol{\mu}_0)$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, $\Omega_n = \mathbf{X}^\top\mathbf{X} + \Omega_0$,
$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-p)s^2 + (\boldsymbol{\mu}_n - \hat{\boldsymbol{\beta}})^\top\mathbf{X}^\top\mathbf{X}(\boldsymbol{\mu}_n - \hat{\boldsymbol{\beta}})$
$+ (\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)^\top\Omega_0(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)$ and $s^2 \equiv (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p)$.

Figure 5.3: Prior-to-Posterior updating for the Gaussian linear regression with conjugate prior.

prior variance of σ^2/κ_0 for each regression coefficient and assuming that the regression coefficients are a priori independent, since $\sigma^2\Omega_0^{-1}$ is diagonal. Another convenient choice is Zellner's prior where $\Omega_0 = \kappa_0(\mathbf{X}^\top\mathbf{X})$; the covariates are assumed to be known and so can be used when formulating a prior. One way to understand Zellner's prior is that its prior covariance matrix is $\kappa_0^{-1}\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, a scaled version of the sampling covariance matrix of the MLE, $\mathbb{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$. Zellner's prior therefore automatically adjusts to the potentially different scales of the covariates. The covariance matrix in Zellner's prior can more generally be defined as a scaled version of the Fisher information, i.e. the prior information is proportional to the expected information from a sample of size n. By setting $\kappa_0 = 1/n$, Zellner's prior therefore becomes a noninformative **unit information prior** with information content equal to the expected information from just a single observation. We will have more to say about Ω_0 in the chapter Regularization.

unit information prior

Figure 5.3 shows that the prior in (5.10) is indeed a conjugate prior. Figure 5.3 also gives the marginal posterior of $\boldsymbol{\beta}$ as **multivariate student-t distribution**, see Figure 5.4. The proof of these results are given at the end of this chapter.

Multivariate student-t

$\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu \sim \mathrm{t}_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathbf{x} \in \mathbb{R}^p$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix and $\nu > 0$ are the degrees of freedom.

$$p(\mathbf{x}) = \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)(\nu\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \times \left(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-(\nu+p)/2}$$

$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$ if $\nu > 1$

$\mathbb{V}(\mathbf{x}) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$ if $\nu > 2$

Marginal distributions:

$x_k \sim \mathrm{t}_\nu(\mu_k, \sigma_k^2)$
$\mathbf{x}_1 \sim \mathrm{t}_\nu(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$

Conditional distributions:

$\mathbf{x}_1|\mathbf{x}_2 \sim \mathrm{t}_{\nu+p_2}(\tilde{\boldsymbol{\mu}}_1, c(\mathbf{x}_2)\cdot\tilde{\boldsymbol{\Sigma}}_1)$

where

$\tilde{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$
$\tilde{\boldsymbol{\Sigma}}_1 = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$
$c(\mathbf{x}_2) = \frac{\nu + d(\mathbf{x}_2)}{\nu + p_2}$
$d(\mathbf{x}_2) = (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$.

University salaries data. The **salaries dataset**, described in the book Fox and Weisberg [2019] and made available in the R package `car`, contains salaries for $n = 397$ university professors. The professors have three different ranks (Assistant, Associate and Full professor) and works in two different disciplines (A and B). The number of years since the PhD degree (academic age) is thought to be an important determinant of salaries. Table 5.1 summarizes the data.

Since salaries are positive and often skewed, we follow the usual convention of taking the natural logarithm of salaries as the response variable to make them more normal. Figure 5.5 plots the response variable `logsalary` against `phdage`, the year since the PhD degree normalized so that `phdage`$= 0$ is a fresh PhD graduate and `phdage`$= 1$ for the professor with highest academic age in the dataset. The relationsship seems to be nonlinear with salaries first rapidly increasing with `phdage` and then possibly decreasing toward the end of the career; note however that the data are **cross-sectional** where each observation is a unique professor, not **longitudinal** where persons are measured at several points in time. The nonlinearity will be modelled by using also the square of `phdage` as a covariate. Some of the nonlinearities also seem to disappear when we control for the rank, see the graph on the top right in Figure 5.5).

cross-sectional

longitudinal

variable	description	data type	values	comment
`logsalary`	log(salary)	continuous	$(-\infty, \infty)$	
`phdage`	years since PhD	continuous	$[0, 1]$	normalized
`rank`	humidity	continuous	$\{A, B, C\}$	ass.Prof $\rightarrow$ Prof
`sex`	sex	binary	$[M, F]$	coded as $M = 1$
`discipline`	year	binary	$\{A, B\}$	coded as $A = 1$

Table 5.1: Summary of the university salaries data.

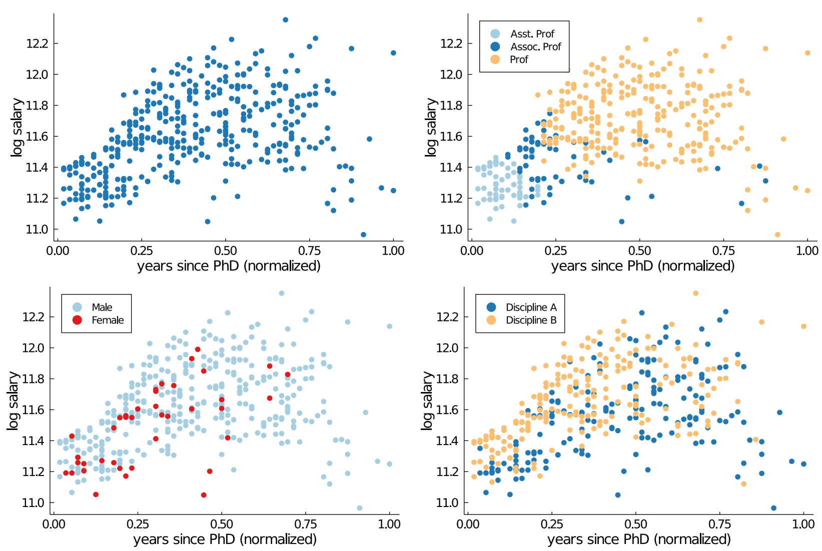

Figure 5.5: University salaries data. Scatterplot of `logsalary` against `phdage` (topleft), colorcoded by `rank` (top right), `sex` (bottom left) and `discipline` (bottom right). See Table 5.1 for variable definitions.

Datasets typically contains categorical covariates that needs to be recoded into several binary variables, often called **dummy variables** in statistics and **one-hot encoding** in machine learning. The usual practice is to code a categorical variable with K different values, or levels, into K binary variables, where an observation in category k is recorded as 1 in the kth binary variable and 0 in the other variables. For example, the variable `rank` is A for assistant professors, B for associate professors, and C for full professors. This variable is coded

dummy variables

one-hot encoding

into $K = 3$ new binary variables: `rank1`, `rank2`, and `rank3` where, for example, an observation for an associate professor is coded as 1 in `rank2` and 0 in `rank1` and `rank3`.

Using all K binary variables as covariates in a regression model introduces an exact linear dependence, or exact **multicollinearity**, between the covariates: the sum of the K covariates is exactly one for any observation. This causes problems in the estimation of the regression coefficients and standard practise is therefore to use only $K-1$ of the binary covariates. We will always drop the binary variable for the first category, which is then the **reference category**. The β coefficient for each of the $K-1$ included covariates now measures the additional effect of the category *over and above* the effect in the reference category. The effect of the reference category ends up in the intercept since all of the $K-1$ included covariates are zero for observations in the reference category.

multicollinearity

reference category

The model for $y = \texttt{logsalary}$ is then

$$\begin{aligned}\texttt{logsalary} = \beta_0 + \beta_1 \cdot \texttt{phdage} + \beta_2 \cdot \texttt{phdagesqr} + \beta_3 \cdot \texttt{rank2} \\ + \beta_4 \cdot \texttt{rank3} + \beta_5 \cdot \texttt{sex} + \beta_6 \cdot \texttt{discipline} + \varepsilon,\end{aligned}$$

where `phdagesqr` is the square of `phdage`, `sex` and `discipline` are each 0-1 coded variables where `sex=1` for males and `discipline=1` for discipline A, respectively. The errors ε are iid $\mathrm{N}(0, \sigma^2)$.

I will first elicit a prior σ^2 and then for $\boldsymbol{\beta}$. My prior for σ^2 is $\mathrm{Inv}\!-\!\chi^2(\nu_0 = 10, \sigma_0^2 = 0.3^2)$ and is plotted in Figure 5.8. I came up with this prior by first looking up online that the median salary for associate professors (middle rank) in the US is around $80,000. Since we will assume that `logsalary` is normally distributed, the salary on the original scale follows a **log-normal distribution**, see Figure 5.6. I plotted the implied log normal distribution of salaries, $\mathrm{LN}(80000, \sigma_0^2)$, for some different values of σ_0^2. The log normal distribution for salary given $\sigma_0^2 = 0.3^2$ is shown to the left in Figure 5.9, where the orange line marks out the salary spread as given by the difference between the 10% and 90% percentiles. This agrees rather well with my prior beliefs about the salary spread and $\sigma_0^2 = 0.3^2$ therefore seems reasonable. To determine ν_0, I compute the same measure of salary spread for 100,000 draws from the $\mathrm{Inv}\!-\!\chi^2(\nu_0, \sigma_0^2 = 0.3^2)$ prior for some different value of ν_0. The result for $\nu_0 = 10$ to the right in Figure 5.9 agrees with by prior beliefs: the spread could be as low as $50,000, but also as much as $150,000; I am not very familiar with US salaries.

Log-Normal distribution

$X \sim \mathrm{LN}(\mu, \sigma^2)$
Support: $X \in (0, \infty)$

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

$\mathbb{E}(X) = \exp\left(\mu + \sigma^2/2\right)$

$\mathbb{V}(X) = (\exp \sigma^2 - 1)\exp(2\mu + \sigma^2)$

and μ is the median of X.

If $Y \sim \mathrm{N}(\mu, \sigma^2)$ then $\log Y \sim \mathrm{LN}(\mu, \sigma^2)$.

Figure 5.6: The log-normal distribution.

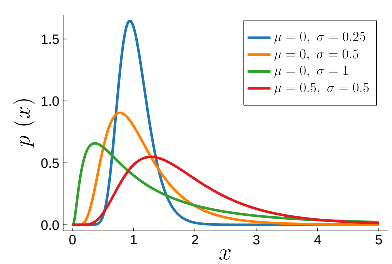

Figure 5.7: Some log-normal distributions.

log-normal distribution

I will use Zellner's prior $\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}(\mu_0, \sigma^2\Omega_0^{-1})$ with $\Omega_0 = \kappa_0(\mathbf{X}^\top\mathbf{X})$, and experiment with κ_0 to see the effect of this prior hyperparameter.

The prior mean of $\boldsymbol{\beta}$ is set to $\boldsymbol{\mu}_0 = (b_0, b_1, b_2, 0, 0, 0, 0)$,. This prior

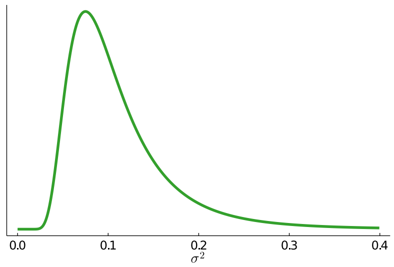

Figure 5.8: Prior for σ^2 in university salaries data.

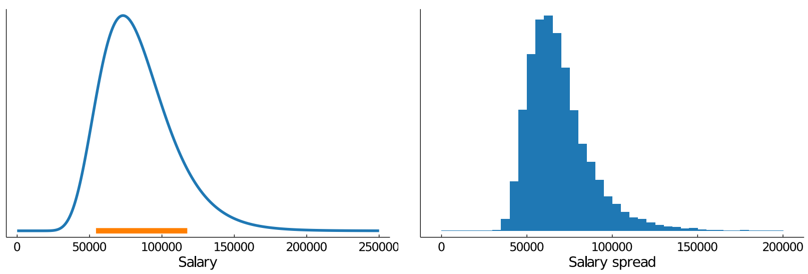

Figure 5.9: Prior elication for σ^2 in university salaries data. Left: Implied log-normal distribution of salaries from assuming a median salary of 80,000 and $\sigma_0^2 = 0.3^2$; the orange line marks out the wage spread as measured by the difference between the 90% and 10% salary percentiles. Right: implied prior distribution on the wage spread from the Inv$-\chi^2(\nu_0 = 10, \sigma_0^2 = 0.3^2)$ prior.

implies that the most probable model a priori is the simplified model

$$\texttt{logsalary} = b_0 + b_1 \cdot \texttt{phdage} + b_2 \cdot \texttt{phdagesqr} + \varepsilon,$$

and we can determine values for b_0, b_1, b_2 and κ_0 that are sensible given our knowledge of university wages. I set $b_0 = \log(70{,}000)$ so that the median salary for newly graduated professor (phdage$=$ 0) is \$70,000, i.e. \$10,000 below the median salary for middle rank professors found in my online search. The coefficient on phdage is set to $b_1 = 2$ and $b_2 = -1.5$ is used for phdagesqr; these values imply a median salary of middle age professors (phdage$=$ 0.5) around $\exp(\log(70{,}000) + 2 \cdot 0.5 - 1.5 \cdot 0.5^2) \approx \$130{,}777$ and a median salary for the oldest professors (phdage$=$ 1) around \$115,410.

	mean	std	lower95	upper95
intercept	11.20	0.03	11.13	11.26
phdage	1.36	0.20	0.96	1.75
phdagesqr	−1.11	0.19	−1.48	−0.74
rank2	0.04	0.03	−0.02	0.11
rank3	0.17	0.04	0.10	0.25
sex	0.03	0.02	−0.02	0.07
discipline	−0.07	0.01	−0.09	−0.04
σ	0.20	0.01	0.19	0.21

Table 5.2: Summary of the posterior distribution for the regression for the salaries data. The summaries for the regression coefficients were computed analytically from their marginal student-t posterior. The summaries for σ was computed by taking the square root transformation of 10,000 posterior draws of σ^2.

It remains to determine κ_0 which determines the precision in the prior for $\boldsymbol{\beta}$. One way to determine κ_0 is to simulate from the prior for different values of κ_0 and determine if the simulated prior agrees with our prior beliefs. Figure 5.10 explores the prior by simulation for four different κ_0: 0.01, 0.1, 1 and 10 by plotting the implied median salary curve over phdage for each of ten $\boldsymbol{\beta}$ simulated from the prior. The think curve is the median salary at prior mean

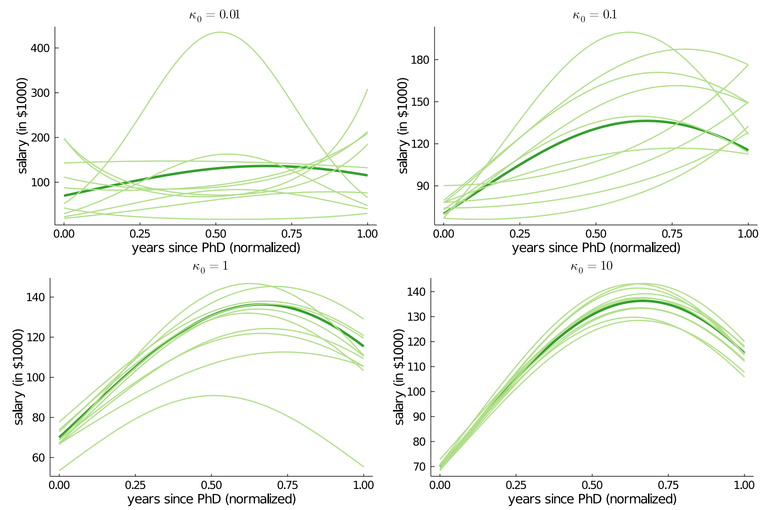

Figure 5.10: Implied relationship between `salary` and `phdage` from 10 simulations from the prior for $\boldsymbol{\beta}$ for four different κ_0 values. The thick line is the median salary at the prior mean $\boldsymbol{\mu}_0 = (\log(70{,}000), 2, -1.5, 0, 0, 0, 0)$.

$\boldsymbol{\mu}_0 = (b_0, b_1, b_3, 0, 0, 0, 0)$ and the thinner lighter curves are the median salary curves for the prior draws of $\boldsymbol{\beta}$. The smallest κ_0 gives too much uncertainty about the relationship between `salary` and `phdage` and the largest κ_0 implies a prior that contains more information than I actually have about US academic wages. $\kappa_0 = 1$ seems like an appropriate value for my prior beliefs and I will continue the posterior analysis with this prior.

Table 5.2 presents a summary of the posterior distribution for the prior with $\kappa_0 = 1$. We see that all β coefficients except for `rank2` and `sex` have 95% posterior intervals that do not include zero and can therefore be said to be important ("significant") in the Bayesian analysis. Hence, associate professors (`rank2=1`) can not be shown to have higher salaries than assistant professors, but full professors are likely to have higher salaries than assistant professors (for the same academic age and other covariates). Figure 5.11 shows the marginal posteriors for the regression coefficients for both $\kappa_0 = 1$ and $\kappa_0 = 0.1$; the maximum likelihood (ML) estimate is also marked out with a green dot; the choice of κ_0 has some effect on the posteriors, which is expected since the sample size is fairly small ($n = 397$). Finally, Figure 5.12 displays the posterior distribution of the salary for female professors in Discipline B at different `phdage` for the three ranks; there is a clear decrease in salary in the last quarter of the career.

BIKE SHARE DATA. The **bike share dataset** collected by Fanaee-T and Gama [2013] and made available in the UCI repository[1] records the number of daily rides with the bike share company capital bikeshare. The dataset contains the number of daily bike rides on 731 days during the two years 2011 and 2012 and a number of variables that may affect the demand for bikes, e.g. weather conditions, day of the

bike share dataset

[1] `https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset`

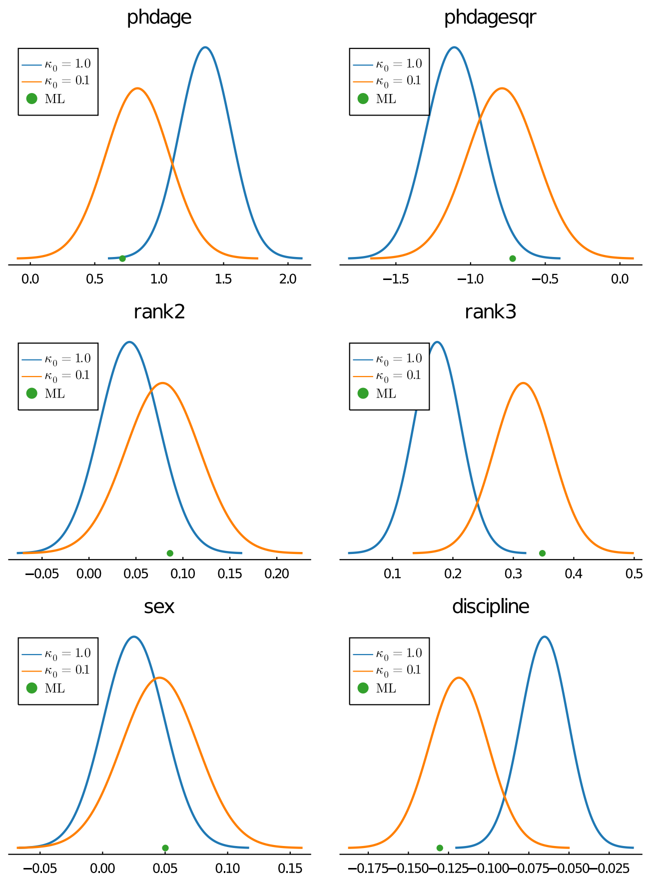

Figure 5.11: Marginal posterior densities for the regression coefficients in Gaussian linear regression fitted to the salaries data with two different priors. The maximum likelihood (ML) estimate is marked out with a green dot. See Table 5.1 for variable definitions.

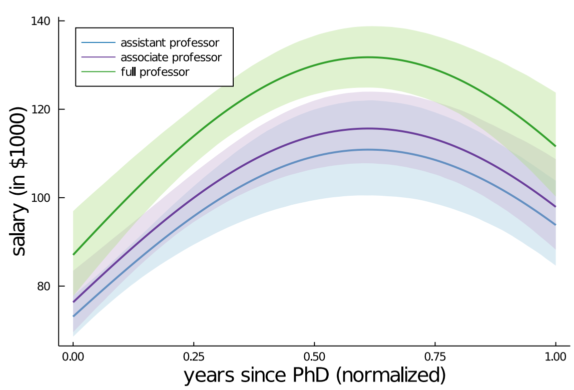

Figure 5.12: Posterior distribution (mean + 95% pointwise intervals) of the salary for female professors in Discipline B at different `phdage` for the three ranks.

week and holidays; Table 5.3 summarizes the dataset. Figure 5.13 plots the time series of daily rides.

We will ignore the time series nature of `nrides` in this chapter and model it by regression; in the next chapter on prediction the model will be extended with time series aspects. The variable `nrides` are count data, but we will nevertheless model it by a Gaussian linear regression since large counts are often approximately Gaussian; regression models for count data will be introduced in Chapter Classification.

variable	description	data type	values	comment
`nrides`	number of rides	counts	$\{0,1,\ldots\}$	min= 22, max= 8714
`feeltemp`	perceived temp	continuous	$[0,1]$	min= 0.07, max= 0.85
`hum`	humidity	continuous	$[0,1]$	min= 0.00, max= 0.98
`wind`	wind speed	continuous	$[0,1]$	min= 0.02, max= 0.51
`year`	year	binary	$\{0,1\}$	year 2011 $= 0$
`season`	season	categorical	$\{1,2,3,4\}$	winter $\rightarrow$ fall
`weather`	weather	ordinal	$\{1,2,3\}$	clear $\rightarrow$ rain/snow
`weekday`	day of week	categorical	$\{0,1,\ldots,6\}$	sunday $\rightarrow$ saturday
`holiday`	holiday	binary	$\{0,1\}$	holiday $= 1$

Table 5.3: Summary of the bike share data.

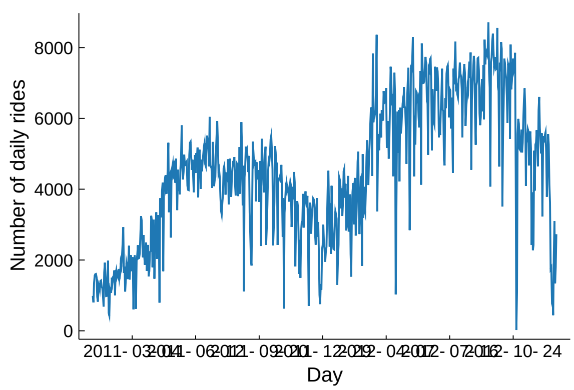

Figure 5.13: Time series plot of `nrides` in the Bike share data.

The dataset contains several categorical covariates which again needs to be one-hot encoded into several binary variables. For example, the variable `season` is coded into the $K = 3$ new binary variables: `season2`, `season3`, and `season4`, i.e. the first season (winter) is the reference category.

Figure 5.14 show scatterplots of `nrides` against the most important continuous covariate, the perceived temperature `feeltemp`. It is clear that `feeltemp` can only explain a smaller portion of the rather sizeable variability in `nrides`. The relationship between `nrides` and `feeltemp` seems slightly nonlinear: there are less biking on the hottest days, but it hard to tell when plotting on against only one covariate as the decrease in rides at high temperatures may be explained by other covariates, and we choose not to add higher order polynomial terms here. There are also some days with extremely low number of rides; these **outliers** correspond to hurricanes and will be more discussed when we revisit this example in the chapter Prediction and Decision making.

outliers

Figure 5.14 also shows the effect of some of the categorical variables by color coding the observations with respect to the levels: rainy weather accounts for some of the low `nrides` observations, and fall (`season`$= 4$) seems to have more biking than winter (`season`$= 1$) for the same temperature.

I use Zellner's unit information prior for simplicity by setting $\kappa_0 = 1/n$. The prior mean μ_0 for $\boldsymbol{\beta}$ is set to the zero vector with the exception of the intercept which is 1000 to reflect a rough guess of the number of rides on a day where all covariates are hypothetically zero (a very cold, dry and clear winter sunday with no wind). I set $\sigma_0^2 = 1000^2$ as a rough guess of σ^2, with $\nu_0 = 5$ so that my prior information about σ^2 is only worth five observations.

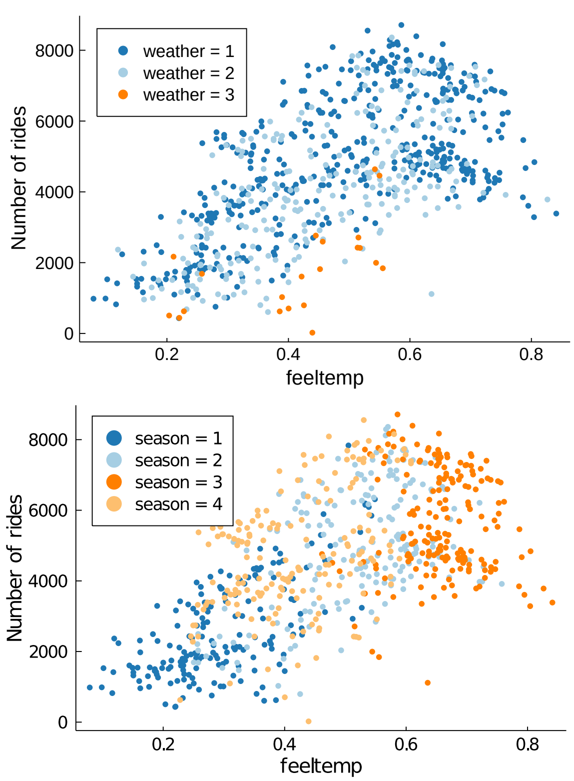

Figure 5.14: Bike share data. Sccatterplots of `nrides` against `feeltemp`, colorcoded by `weather` (top), `season` (bottom). See Table 5.3 for variable definitions.

TODO! RESIDUAL ANALYSIS. ADD LAGS. Pointer to prediction chapter.

	mean	std	2.5%	97.5%
`intercept`	1142.26	242.44	666.94	1617.57
`feeltemp`	5477.32	340.49	4809.79	6144.84
`hum`	−1245.12	301.81	−1836.83	−653.41
`wind`	−2494.02	435.24	−3347.32	−1640.72
`year`	2021.15	62.66	1898.30	2144.01
`season2`	1173.01	114.54	948.45	1397.58
`season3`	966.57	147.43	677.53	1255.61
`season4`	1541.81	98.33	1349.03	1734.58
`weather2`	−447.70	82.83	−610.09	−285.32
`weather3`	−1945.19	211.88	−2360.58	−1529.79
`weekday1`	203.28	118.65	−29.34	435.91
`weekday2`	298.03	115.94	70.73	525.34
`weekday3`	377.65	116.18	149.88	605.43
`weekday4`	392.76	116.15	165.04	620.47
`weekday5`	454.84	116.13	227.16	682.53
`weekday6`	446.26	115.54	219.75	672.77
`holiday`	−630.00	193.07	−1008.52	−251.48
σ	835.00	21.85	793.74	871.65

Table 5.4: Summary of the posterior distribution for the regression for the bike share data. The summaries for the regression coefficients were computed analytically from their marginal student-t posterior. The summaries for σ was computed by taking the square root transformation of 10,000 posterior draws of σ^2.

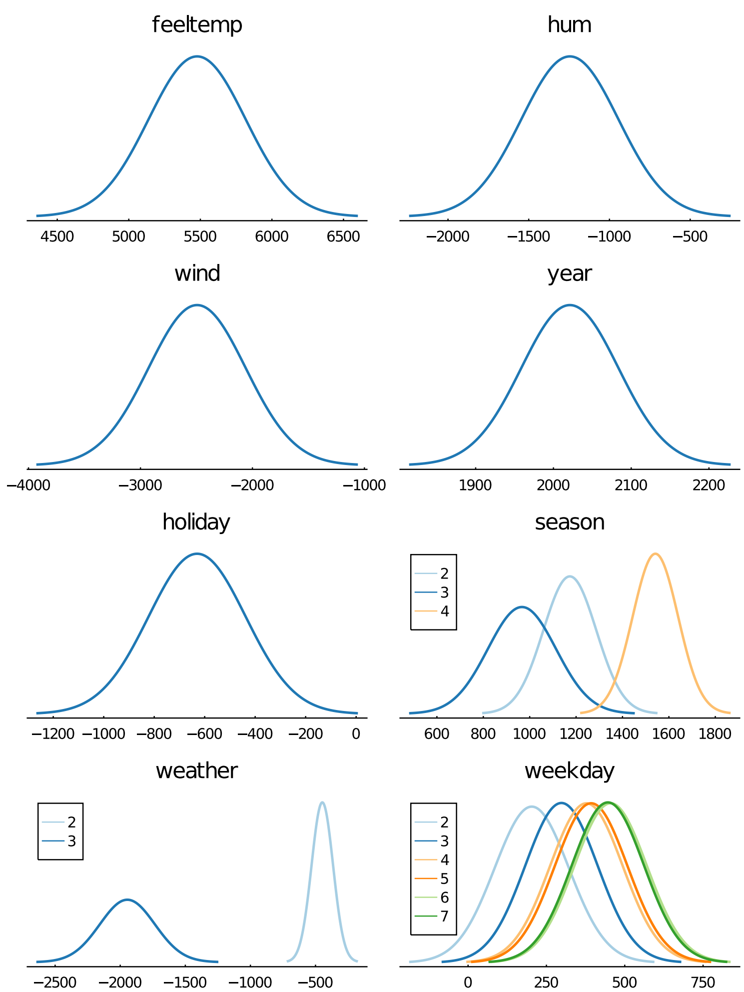

Figure 5.15: Marginal posterior densities for the regression coefficients in Gaussian linear regression fitted to the bike share data. See Table 5.3 for variable definitions.

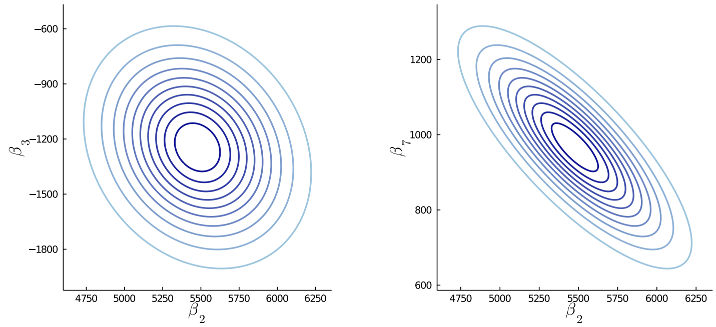

Figure 5.16: Bivariate student-*t* posterior densities for the regression coefficients on `feeltemp` and `hum` (left), and `feeltemp` and `season3` (right) in the bike share data. See Table 5.3 for variable definitions.

PROOFS

This section derives the posterior distribution for linear regression with a conjugate prior in Figure 5.3.

The joint posterior is

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) \\
&\propto |2\pi\sigma^2 I_n|^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{X}\boldsymbol{\beta})^\top (y - \mathbf{X}\boldsymbol{\beta})\right) \\
&\times |2\pi\sigma^2 \Omega_0^{-1}|^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \Omega_0 (\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right) \\
&\times (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}\nu_0\sigma_0^2\right) \\
&\propto (\sigma^2)^{-((\nu_0+n+p)/2+1)} \exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2 + (n-p)s^2)\right) \\
&\times \exp\left(-\frac{1}{2\sigma^2}((\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \Omega_0 (\boldsymbol{\beta} - \boldsymbol{\mu}_0))\right),
\end{aligned}
$$

where $s^2 = (y - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (y - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p)$ as before. Completing the squares in the exponents using the result in Figure 3.25 gives

$$
\begin{aligned}
&(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \Omega_0 (\boldsymbol{\beta} - \boldsymbol{\mu}_0) = \\
&(\boldsymbol{\beta} - \boldsymbol{\mu}_n)^\top \Omega_n (\boldsymbol{\beta} - \boldsymbol{\mu}_n) + (\boldsymbol{\mu}_n - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\mu}_n - \hat{\boldsymbol{\beta}}) + (\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)^\top \Omega_0 (\boldsymbol{\mu}_n - \boldsymbol{\mu}_0),
\end{aligned}
$$

where $\boldsymbol{\mu}_n = \Omega_n^{-1}(\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} + \Omega_0 \boldsymbol{\mu}_0)$. Hence,

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-((\nu_n+p)/2+1)} &\exp\left(-\frac{\nu_n \sigma_n^2}{2\sigma^2}\right) \\
\times &\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_n)^\top \Omega_n (\boldsymbol{\beta} - \boldsymbol{\mu}_n)\right)
\end{aligned}
\tag{5.12}
$$

where $\nu_n = \nu_0 + n$ and $\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n-p)s^2 + (\boldsymbol{\mu}_n - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\mu}_n - \hat{\boldsymbol{\beta}}) +$

$(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)^\top \Omega_0 (\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)$. Now,

$$\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) &\propto (\sigma^2)^{-((\nu_n+p)/2+1)} \exp\left(-\frac{\nu_n \sigma_n^2}{2\sigma^2}\right) |2\pi\sigma^2\Omega_n^{-1}|^{1/2} \\
&\quad \times |2\pi\sigma^2\Omega_n^{-1}|^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top \Omega_n (\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right) \\
&\propto (\sigma^2)^{-(\nu_n/2+1)} \exp\left(-\frac{\nu_n \sigma_n^2}{2\sigma^2}\right) \\
&\quad \times |2\pi\sigma^2\Omega_n^{-1}|^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top \Omega_n (\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right).
\end{aligned}$$

From the second factor we see that $\boldsymbol{\beta}|\sigma^2, \mathbf{y} \sim N(\boldsymbol{\mu}_n, \sigma^2\Omega_n^{-1})$ and from the first factor that $\sigma^2|\mathbf{y} \sim \text{Inv}-\chi^2(\nu_n, \sigma_n^2)$.

The marginal posterior of $\boldsymbol{\beta}$ is obtained by integrating $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ in (5.12) with respect to σ^2 using properties of the Inv$-\chi^2$ distribution

$$\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y}) &\propto \int (\sigma^2)^{-((\nu_n+p)/2+1)} \times \exp\left(-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2 + (\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top \Omega_n (\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right)\right) d\sigma^2 \\
&\propto \left(\left(\nu_n\sigma_n^2 + (\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top \Omega_n (\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right)/2\right)^{-(\nu_n+p)/2} \\
&\propto \left(1 + \frac{1}{\nu_n}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top \sigma^{-2}\Omega_n (\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right)^{-(\nu_n+p)/2}
\end{aligned}$$

which is proportional to the multivariate student-t density

$$\boldsymbol{\beta}|\mathbf{y} \sim t_{\nu_n}(\mu_n, \sigma_n^2\Omega_n^{-1}).$$

EXERCISES

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

NOTEBOOKS

1. See the notebook regression.

6 Prediction and Decision making

TODO! write intro text.

Bayesian prediction

We often want a prediction for an unknown quantity. That unknown quantity can be future yet unobserved value x_t of a time series, or the response observation for a person y given that persons covariate values $\mathbf{x}$ in a regression problem; for example the expected effect of some medical treatment for a person with some given characteristics such as age, weight, smoking and exercise habits.

We will use the tilde ($\sim$) symbol to make explicit that a variable is the aim for prediction. Hence, $\tilde{y}$ is for example the regression response that we want to predict based on observed covariates $\tilde{\mathbf{x}}$ for a given subject; in a time series problem we let $\tilde{y}_{T+h}$ denote the time series h time periods in the future relative to the time T where the prediction is made.

Having already observed n training data points, $\mathbf{y} = (y_1, \ldots, y_n)$, we now want a prediction of a new observation $\tilde{y}$. Consider first an iid model for the data: $y_i|\theta \overset{iid}{\sim} p(y|\boldsymbol{\theta})$. The Bayesian **predictive distribution** is the distribution of the unknown $\tilde{y}$ given the known training data $\mathbf{y}$:

predictive distribution

$$p(\tilde{y}|\mathbf{y}) = \int_\theta p(\tilde{y}|\theta)p(\theta|\mathbf{y})d\theta, \tag{6.1}$$

where $p(\theta|\mathbf{y})$ is the posterior distribution for the model parameters $\boldsymbol{\theta}$. The predictive distribution is therefore a weighted average of the model distribution $p(\tilde{y}|\theta)$ with respect to $\boldsymbol{\theta}$, with the posterior density $p(\theta|\mathbf{y})$ as weights.

The predictive distribution in (6.1) can be summarized by a **point prediction**, for example the predictive mean $\mathbb{E}(\tilde{y}|\mathbf{y})$, and predictive variance $\mathbb{V}(\tilde{y}|\mathbf{y})$ or by a 95% **predictive interval**, just as we summarized the posterior distribution for a parameter θ. But it is important to remember that the Bayesian approach gives a complete probabil-

point prediction

predictive interval

ity distribution for the unknown $\tilde{y}$, not just a point prediction and variance. As we will see, the predicted value can even be a vector in which case the predictive distribution $p(\tilde{\mathbf{y}}|\mathbf{y})$ is a multivariate distribution.

When observations are not necessarily iid, for example in time series problems, we have the slightly more general form

$$p(\tilde{y}|\mathbf{y}) = \int_\theta p(\tilde{y}|\theta, \mathbf{y})p(\theta|\mathbf{y})d\theta, \tag{6.2}$$

where the distribution of the predicted value $\tilde{y}$ now depends on all the training data $\mathbf{y}$. In many cases it is enough to condition on just a few of the training data point. For example, in the AR(p) process

$$y_t = \mu + \sum_{k=1}^{p} \phi_k(y_{t-k} - \mu) + \varepsilon, \tag{6.3}$$

we only need to condition on the p values preceeding the time period we want to predict; the AR(p) is said to be a **Markov process** of order p, as explained later in this chapter.

The following subsections presents a series of prediction examples with varying degree of complexity.

Iteration laws

Law of iterated expectation:

$$\mathbb{E}_X(X) = \mathbb{E}_Y(\mathbb{E}_{X|Y}(X))$$

Law of total variance:

$$\mathbb{V}_X(X) = \mathbb{E}_Y(\mathbb{V}_{X|Y}(X)) + \mathbb{V}_Y(\mathbb{E}_{X|Y}(X))$$

Figure 6.1: Law of iterated expectations and law of total variance.

Iteration laws for Bayes

Marginal posterior mean:

$$\mathbb{E}_{\theta_1|\mathbf{y}}(\theta_1) = \mathbb{E}_{\theta_2|\mathbf{y}}(\mathbb{E}_{\theta_1|\theta_2,\mathbf{y}}(\theta_1))$$

Marginal posterior variance:

$$\mathbb{V}_{\theta_1|\mathbf{y}}(\theta_1) = \mathbb{E}_{\theta_2|\mathbf{y}}(\mathbb{V}_{\theta_1|\theta_2,\mathbf{y}}(\theta_1)) + \mathbb{V}_{\theta_2|\mathbf{y}}(\mathbb{E}_{\theta_1|\theta_2,\mathbf{y}}(\theta_1))$$

Figure 6.2: Iteration laws applied to compute marginal posterior moments given some data $\mathbf{y}$.

Prediction in normal model with known variance

My streaming service becomes unreliable and buffers at speeds below 5Mbit/sec. I am therefore particularly interested in this 'catastrophic' event happening tonight while watching my favourite movie. Finding the probability of a *single* measurement lower than 5MBit/sec is an exercise in prediction.

The Gaussian model $\tilde{y} \sim \mathrm{N}(\theta, \sigma^2)$ for my internet speed can be trivially expressed as $\tilde{y} = \theta + \tilde{\varepsilon}$, where $\tilde{\varepsilon} \sim \mathrm{N}(0, \sigma^2)$. Since we already know that the posterior for θ is $\mathrm{N}(\mu_n, \tau_n^2)$ we see that $\tilde{y}$ is the sum of two Gaussian variables, and the predictive distribution for $\tilde{y}$ is therefore also Gaussian (Figure 5.1). To obtain the mean and variance of this predictive distribution it is helpful to first condition on θ and then 'undo' the conditioning by integrating with respect to the posterior for θ. This two-step approach of computing the mean and variance of random variables by first conditioning on an another random variable are called the iteration laws; specifically the **law of iterated expectation** and the **law of total variance**. Figure 6.1 gives these laws in the case of two generic random variables X and Y as typically presented in introductory probability textbooks. Figure 6.2 are the exact same laws but written in the context of computing the marginal posterior mean and variance for a parameter. Note the use of subscripts on expectations to explicitly denote which distribution

law of iterated expectation

law of total variance

the expectation is taken with respect to; for example

$$\mathbb{E}_{\theta|\mathbf{y}}(\theta) \equiv \int \theta p(\theta|\mathbf{y}) d\theta.$$

The predictive mean of $\tilde{y}$ can now be computed by first computing the mean given θ

$$\mathbb{E}_{\tilde{y}|\mathbf{y},\theta}(\tilde{y}) = \theta$$

and then undo the conditioning in the second step by taking the posterior expectation

$$\mathbb{E}(\tilde{y}|\mathbf{y}) = \mathbb{E}_{\theta|\mathbf{y}}(\theta) = \mu_n,$$

since μ_n is by definition the posterior mean of θ. The predictive variance is similarly given by the law of total variance as

$$\begin{aligned}\mathbb{V}(\tilde{y}|\mathbf{y}) &= \mathbb{E}_{\theta|\mathbf{y}}[\mathbb{V}_{\tilde{y}|\mathbf{y},\theta}(\tilde{y})] + \mathbb{V}_{\theta|\mathbf{y}}[\mathbb{E}_{\tilde{y}|\mathbf{y},\theta}(\tilde{y})] \\ &= \mathbb{E}_{\theta|\mathbf{y}}(\sigma^2) + \mathbb{V}_{\theta|\mathbf{y}}(\theta) \\ &= \sigma^2 + \tau_n^2.\end{aligned}$$

Hence, the posterior predictive distribution is

$$\tilde{y}|\mathbf{y} \sim N(\mu_n, \sigma^2 + \tau_n^2).$$

The predictive variance is the sum of the model variance σ^2 and the posterior variance of θ, τ_n^2, which represents the parameter uncertainty from not knowing θ when we make the prediction. The model variance σ^2 comes from each observation not being completely predictable even if the $N(\theta, \sigma^2)$ model was entirely known. The parameter uncertainty will disappear with more training data since $\tau_n^2 \to 0$ as $n \to 0$. These two sources of predictive uncertainty appear at least implicitly in all models, and their relative importance depends on size of the training sample, the fit and complexity of the model.

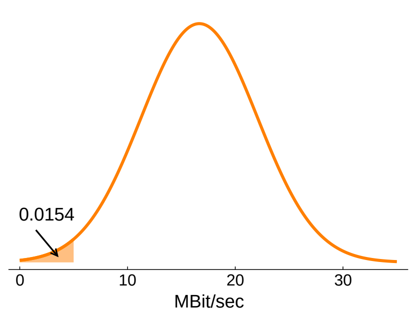

Figure 6.3: Predictive density for the internet download speed after observing $n = 5$. The probability of less than 5MBit/sec download speed is marked out by the orange region.

Figure 6.3 plots the predictive distribution for the internet download speed example with $n = 5$ training observations, and marks out the probability of interest, $\Pr(\tilde{y}|y_1, \ldots, y_5) \approx 0.0154$.

Prediction in linear regression

Consider the linear regresson model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 I_n). \tag{6.4}$$

We have already in Chapter Regression learned how to use a training dataset $(\mathbf{y}, \mathbf{X})$ with n observations to compute the posterior for the conjugate prior:

$$\sigma^2 | \mathbf{X}, \mathbf{y} \sim \mathrm{Inv}{-}\chi^2(\nu_n, \sigma_n^2)$$
$$\boldsymbol{\beta} | \sigma^2, \mathbf{X}, \mathbf{y} \sim N(\mu_n, \sigma^2 \Omega_n^{-1})$$

Interest now centers on predicting the response $\tilde{\mathbf{y}}$ for $\tilde{n}$ new observations using the $\tilde{n} \times p$ covariate matrix $\tilde{\mathbf{X}}$; the most common case is when $\tilde{n} = 1$ so that we predict a single response $\tilde{y}$ using a vector of covariates $\tilde{\mathbf{x}}$ for that observation. The joint posterior predictive distribution for all $\tilde{n}$ elements of $\tilde{\mathbf{y}}$ is

$$p(\tilde{\mathbf{y}} | \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}) = \iint p(\tilde{\mathbf{y}} | \tilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2 | \mathbf{X}, \mathbf{y}) d\boldsymbol{\beta} d\sigma^2. \tag{6.5}$$

I have here impliclity used some conditional independencies to reduce the notational clutter. We can for example write $p(\tilde{\mathbf{y}} | \tilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2)$ instead of the longer $p(\tilde{\mathbf{y}} | \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \sigma^2)$, since $\tilde{\mathbf{y}}$ is independent of the training data $\mathbf{X}, \mathbf{y}$ conditional on the parameters $\boldsymbol{\beta}, \sigma^2$; that is, given the true parameter values, there is no additional information in the training data that is useful for predicting $\tilde{\mathbf{y}}$.

The predictive distribution in (6.5) can be derived in two steps:

i) integrate out $\boldsymbol{\beta}$ to get $p(\tilde{\mathbf{y}} | \sigma^2, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}) = \int p(\tilde{\mathbf{y}} | \tilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} | \sigma^2, \mathbf{y}) d\boldsymbol{\beta}$

ii) integrate out σ^2 to obtain $p(\tilde{\mathbf{y}} | \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}) = \int p(\tilde{\mathbf{y}} | \sigma^2, \tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y}) p(\sigma^2 | \mathbf{y}) d\sigma^2$.

These two steps are derived in the Proof section at the end of the chapter. Figure 6.4 summarizes the end result: the joint predictive distribution of all test responses of $\tilde{\mathbf{y}}$ is a multivariate student-t.

The result in Figure 6.4 also shows that the predictive distribution includes uncertainty from two sources:

i) *observation noise* $\tilde{\boldsymbol{\varepsilon}}$, represented by the term $\sigma_n^2 \boldsymbol{I}_{\tilde{n}}$, and

ii) *parameter uncertainty*, represented by the term $\sigma_n^2(\tilde{\mathbf{X}} \Omega_n^{-1} \tilde{\mathbf{X}}^\top)$.

To see that the latter term is the uncertainty that comes from not knowing the parameters, note that the prediction of $\tilde{\mathbf{y}}$ conditional on the parameters is given by $\tilde{\mathbf{X}}\boldsymbol{\beta}$. This explains why the predictive variance is a quadratic form in $\tilde{\mathbf{X}}$; see the proof at the end of the chapter for a more precise explanation. The parameter uncertainty will vanish with large training samples since it can be shown that

Predictive density conjugate Gaussian linear regression

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

Posterior: $\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}_n, \sigma^2\Omega_n^{-1})$
$\sigma^2|\mathbf{y}, \mathbf{X} \sim \mathrm{Inv}{-}\chi^2(\nu_n, \sigma_n^2)$

Predictive density for $\tilde{n}$ observations with covariate matrix $\tilde{\mathbf{X}}$:
$$\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathbf{y} \sim t_{\nu_n}\left(\tilde{\mathbf{X}}\boldsymbol{\mu}_n, \sigma_n^2(\boldsymbol{I}_{\tilde{n}} + \tilde{\mathbf{X}}\Omega_n^{-1}\tilde{\mathbf{X}}^\top)\right)$$
using the posterior hyperparameters in Figure 5.3.

Figure 6.4: Predictive density in Gaussian linear regression with a conjugate prior.

$\Omega_n^{-1} \xrightarrow{p} \mathbf{0}$ and $\sigma_n^2 \xrightarrow{p} \sigma^2$ as $n \to \infty$, under the common assumption that $n^{-1}\mathbf{X}^\top\mathbf{X}$ converges to a constant non-singular matrix. In the chapter Model comparison we will see how the posterior predictive distribution can also incorporate model uncertainty, and in chapter Variable selection how to handle the uncertainty in the choice of covariates in regression and classification.

Some additional real-world prediction problems

I will now illustrate Bayesian prediction in two more complex problems. The problems are presented here as motivational examples to show how simulation can be used for Bayesian prediction in real applications; all details will not be given and the examples are not meant to be fully understood at this point.

Predicting inflation with an autoregressive process. Imagine that you have the task of predicting the future development of a time series, for example forecasting the Swedish inflation in the coming 12 quarters. Not only would you like to have a mean prediction, but also some notation of predictive uncertainty.

A popular model for macroeconomic time series forecasting is the autoregressive process with p lags, AR(p), introduced in the chapter Priors:

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \quad (6.6)$$

where y_t is the time series observed at time t, y_{t-k} is the kth lagged value of the time series and ε_t are future shocks to the time series.

Having observed training data $\mathbf{y}_{1:T} \equiv (y_1, \ldots, y_T)$ up to time T, we now want the joint predictive density of the time series in the h coming time periods $\tilde{\mathbf{y}}_{T+1:T+h} \equiv (\tilde{y}_{T+1}, \ldots, \tilde{y}_{T+h})$. This predictive density can as usual be written as an integral with respect to the posterior distribution,

$$p(\tilde{\mathbf{y}}_{T+1:T+h}|\mathbf{y}_{1:T}) = \int p(\tilde{\mathbf{y}}_{T+1:T+h}|\mathbf{y}_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) d\boldsymbol{\theta},$$

where $\boldsymbol{\theta} = (\mu, \phi_1, \ldots, \phi_p, \sigma^2)$ is the vector of parameters in the AR(p) process and $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ is the posterior distribution of the parameters based on the training data.

We can simulate from the predictive distribution $p(\tilde{\mathbf{y}}_{T+1:T+h}|\mathbf{y}_{1:T})$ by repeating the following two steps for $i = 1, \ldots, m$:

- simulate a posterior parameter draw $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$
- simulate a h-steps-ahead realization path $\tilde{\mathbf{y}}^{(i)}_{T+1:T+h}$ from the model $p(\tilde{\mathbf{y}}_{T+1:T+h}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(i)})$ conditional on $\boldsymbol{\theta}^{(i)}$.

The first step above will be described in Chapter Posterior simulation. The second step is implemented using the usual sequential decomposition of a joint distribution

$$\begin{aligned} p(\tilde{\mathbf{y}}_{T+1:T+h}|\mathbf{y}_{1:T}, \boldsymbol{\theta}) =& p(\tilde{\mathbf{y}}_{T+1}|\mathbf{y}_{1:T}, \boldsymbol{\theta}) p(\tilde{\mathbf{y}}_{T+2}|\mathbf{y}_{1:T+1}, \boldsymbol{\theta}) \\ & \cdots p(\tilde{\mathbf{y}}_{T+h}|\mathbf{y}_{1:T+h-1}, \boldsymbol{\theta}). \end{aligned} \tag{6.7}$$

We can simulate from each term in (6.7) forward in time, i.e. from left to right, by iterating on (6.6) with a new simulated future shock, ε_{T+j} injected at each time step. Since the AR(p) process is a **Markov process** of order p (see Figure 6.5) it is sufficient to condition on the p most recent time observations in each term instead of the full training sample $\mathbf{y}_{1:T}$. Note also that with exception of $p(\tilde{\mathbf{y}}_{T+1}|\mathbf{y}_{1:T}, \boldsymbol{\theta})$, all terms in (6.7) conditions on future, yet unobserved values, which have been simulated in earlier time steps. The algorithm is detailed in Figure 6.6 where this is made explicit by highlighting such data points in orange. Using this algorithm with $m = 10{,}000$ draws produces the $h = 12$-steps-ahead predictive distribution for Swedish inflation in Figure 6.7.

Markov process

A discrete-time stochastic process $X_1, X_2, \ldots$ is said to be **first-order Markov** if

$$\Pr(X_{n+1}|\mathbf{X}_{1:n}) = \Pr(X_{n+1}|X_n),$$

i.e. if the distribution of future values are independent of the past, conditional on the most recent value.

A process is **pth order Markov** if the distribution of future values are independent of the past, conditional on the p most recent values.

A Markov process in discrete time is also called a **Markov Chain**.

Figure 6.5: Markov processes.

Markov process

Predictive distribution - AR process.

Input: time series $\mathbf{y}_{1:T} = (y_1, \ldots, y_T)$
number of predictive draws m.
forecast horizon h.

for i in $1{:}m$ **do**

$\mu, \phi_1, \ldots, \phi_p, \sigma \leftarrow$ rPosteriorAR($\mathbf{y}_{1:T}$, Prior)

$\varepsilon_{T+1} \leftarrow \text{rNorm}(0, \sigma)$

$\tilde{y}_{T+1} \leftarrow \mu + \phi_1(y_T - \mu) + \ldots + \phi_p(y_{T+1-p} - \mu) + \varepsilon_{T+1}$

$\varepsilon_{T+2} \leftarrow \text{rNorm}(0, \sigma)$

$\tilde{y}_{T+2} \leftarrow \mu + \phi_1(\tilde{y}_{T+1} - \mu) + \ldots + \phi_p(y_{T+2-p} - \mu) + \varepsilon_{T+2}$

$\vdots$

$\varepsilon_{T+h} \leftarrow \text{rNorm}(0, \sigma)$

$\tilde{y}_{T+h} \leftarrow \mu + \phi_1(\tilde{y}_{T+h-1} - \mu) + \ldots + \phi_p(\tilde{y}_{T+h-p} - \mu) + \varepsilon_{T+h}$

end

Output: m draws from the joint predictive density:
$p(\tilde{y}_{T+1}, \ldots, \tilde{y}_{T+h} | \mathbf{y}_{1:T})$.

Figure 6.6: Algorithm for simulating from the joint h-step-ahead predictive distribution of an AR process. The function rPosteriorAR() uses Gibbs sampling and will be presented in Chapter Posterior simulation. The terms in orange font are future values used in the prediction which have been simulated in earlier time steps.

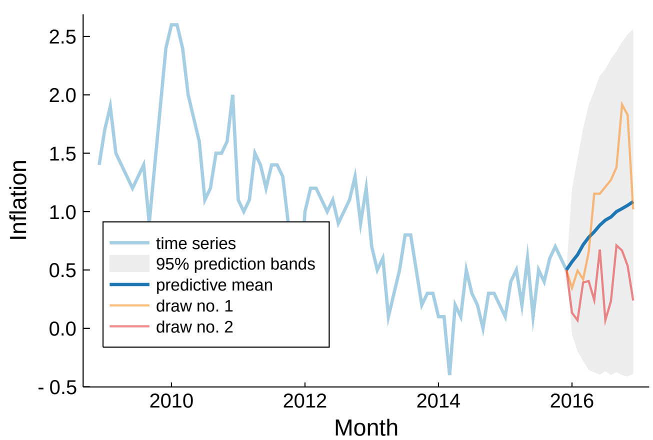

Figure 6.7: Predictive distribution $h = 12$ steps ahead for Swedish inflation represented by a mean prediction in dark blue and 95% predictive intervals as the gray region. Two of the $m = 10{,}000$ simulated paths from the algorithm in Figure 6.6 are marked out.

PREDICTING AUCTION PRICES. In Chapter Single-parameter models we inferred the mean number of bidders in a internet auction using the iid Poisson model. The Poisson model can be extended to a regression model with a mean depending on auction-specific covariates, such as the seller's reservation price, which we saw already in Chapter Single-parameter models was an important determinant of the number of bidders. A Poisson regression for the number of bidders r in an auction with covariates $\mathbf{w}$ is $r|\mathbf{w} \sim \text{Pois}(\lambda)$, where $\lambda = \exp(\mathbf{w}^\top \boldsymbol{\gamma})$. Note that the covariates $\mathbf{w}$ are auction-specific so each auction has its own Poisson mean λ reflecting the specifics (reservation price, seller's rating etc) of that specific auction. The reason for using the exponential function to link the covariates to the Poisson mean λ is that this guarantees that λ is positive.

We will return to a Bayesian analysis of the Poisson regression for the number of bidders in Chapter Classification; this model will here be a component in a bigger model for the final price of the auctioned item. The training dataset includes also all bids in each of the $n = 1{,}000$ auctions. There are two complications that needs to be considered when modeling the price. First, since bidders compete, they are strategic and their disclosed bids b need not be the same as their true valuation v of auctioned object. Game theory can be used to find the optimal bid function $b(v, r, \mu, \sigma)$, mapping the valuations to the observed bids; the optimal bid function depends on the number of bidders r, as well as the mean μ and standard deviation in the bidders evaluations, σ, for the given auction. The second complication is that the bidding system at eBay is such that the winner of the auction pays a price equal to the *second* largest bid. The following simulation scheme describes a simplified version of the model used in Wegmann and Villani [2011]:

1. Simulate the number of bidders $\tilde{r}|\tilde{\mathbf{w}} \sim \text{Pois}(\tilde{\lambda})$, where $\tilde{\lambda} = \exp(\tilde{\mathbf{w}}^\top \boldsymbol{\gamma})$ and $\tilde{\mathbf{w}}$ are the covariates for λ.
2. Simulate the valuations of all $\tilde{r}$ bidders from the linear regression $\text{N}(\tilde{\mu}, \sigma^2)$ where $\tilde{\mu} = \tilde{\mathbf{x}}^\top \boldsymbol{\beta}$ is the regression function depending on covariates $\tilde{\mathbf{x}}$ that determine the mean valuation.
3. Compute bids for all $\tilde{r}$ bidders using the optimal bid function $b(v, \tilde{r}, \tilde{\mu}, \sigma)$.
4. Return the second largest bid as the final price.

The algorithm in Figure 6.8 summarizes this process of simulating from the predictive distribution, including the sampling of the posterior for the model parameters.

Figure 6.9 shows the predictive distribution of two test auctions from the full model in Wegmann and Villani [2011]. The predictive

Predictive distribution - internet auction price.

Input: training auction bids $\mathbf{Y}$
training auction covariates $\mathbf{W}$ for λ.
training auction covariates $\mathbf{X}$ for μ.
test auction covariates $\tilde{\mathbf{w}}$ for λ.
test auction covariates $\tilde{\mathbf{x}}$ for μ.
number of predictive draws m.

for i in $1{:}m$ **do**
$\boldsymbol{\beta}, \gamma, \sigma \leftarrow$ rPostAuction($\mathbf{Y}$, $\mathbf{X}$, $\mathbf{W}$, Prior) # parameters
$\tilde{r} \leftarrow \text{rPois}(\tilde{\lambda})$, where $\tilde{\lambda} = \exp(\tilde{\mathbf{w}}^\top \gamma)$ # bidders
$\tilde{\mathbf{v}}_{1:\tilde{r}} \leftarrow \text{rNorm}(\tilde{\mu}, \sigma)$, where $\tilde{\mu} = \mathbf{x}^\top \boldsymbol{\beta}$ # all valuations v
$\mathbf{b}_{1:\tilde{r}} \leftarrow \text{BidFunction}(\tilde{\mathbf{v}}_{1:\tilde{n}}, \tilde{r}, \tilde{\mu}, \sigma)$ # all bids
$\tilde{p} \leftarrow \text{SecondLargest}(\mathbf{b}_{1:\tilde{n}})$ # final price is 2nd largest bid

end

Output: m draws from the predictive distribution of the final price $\tilde{p}$ in an auction.

Figure 6.8: Algorithm for simulating from the predictive distribution of selling price in internet coin auction with covariates $\tilde{\mathbf{x}}$ for the mean valuation and covariates $\tilde{\mathbf{w}}$ for modeling the mean number of bidders in the auction. The function rPostAuction simulates from the joint posterior of all model parameters using the Metropolis-Hasting algorithm described in Chapter Posterior simulation.

distribution contains three parts: i) the probabilility of no bids ($\tilde{r} = 0$), ii) the probabilility of exactly one bid ($\tilde{r} = 1$), in which the price is the seller's reservation price, and iii) the predictive density of the price given at least two bids. For the auction to the left in Figure 6.9 the actual outcome was more than two bids and a final price of \$41.00, which was fairly well predicted by the model: the probability of $\tilde{r} \leq 1$ was close to zero and \$41.5 has a relatively high density. The auction to the right in Figure 6.9 ended without any bids and this was predicted with a probability of 0.150; the probability of a price equal to the reservation price \$11.45 (the green dot) was 0.269.

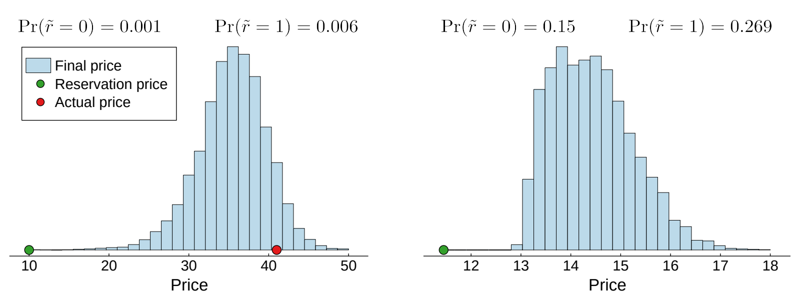

Figure 6.9: Predictive distributions of the final price in two auctions. The green dot marks out the sellers reservation price and the red dot marks out the realized price. The auction in the right hand graph has no red dot since the auction ended without any bids. The predicted probability of no bids and one bid (buyer pays the reservation price) are written in the graph titles. The histograms represent the predictive densities given at least two bids.

Bayesian decisions

Predictions play a major role in modern statistical analysis and machine learning, but the final aim is often **decision making under uncertainty**, with the predictive distribution as an essential component. This is obvious in AI applications, where self-driving cars or automatic stock trading apps need to constantly make decisions to reach pre-determined goals.

decision making under uncertainty

One can argue that decisions are nearly always the final aim, even when this is not as apparent as for automatic AI systems. Consider for example data from a clinical trial where the interest is to quantify the reduction in blood pressure from a given dose of beta-blocker medicine. A first idea would be to **infer** the regression coefficient β in linear regression of blood pressure (y) on the covariates dosage (x) and to check if the value $\beta = 0$ (no effect) is included in a 95% HPD credible interval.

A more interesting goal is **predicting** the blood pressure reduction for a given dosage, particularly if additional subject covariates, such as age, sex, exercise habits etc, are used in the regression model to obtain personalized predictions.

The ultimate goal however is to **make a decision** if a particular patient should be given the medicine. To answer this question we clearly need personalized predictions of the blood pressure reduction and its subsequent effet of reducing the probability of stroke, but also a valuation of the cost and the risk of potential side effects of taking the medicine. This section will introduce the Bayesian framework for making such decisions under uncertainty.

Actions and Utility

Let $a \in \mathcal{A}$ be an **action** in a set $\mathcal{A}$ of possible actions. Let $\theta \in \Theta$ represent an unknown quantity. The consequences of choosing action an a when θ turns out to be θ is quantified by a **utility function** $u(a, \theta)$. The utility function is subjective since the consequences of the actions typically varies from person to person. Table 6.1 present the utility of different action-unknown pairs in an example with a discrete set of actions and a discrete set of possible values for θ.

action

utility function

	θ_1	θ_2	$\cdots$	θ_K
a_1	$u(a_1, \theta_1)$	$u(a_1, \theta_2)$	$\cdots$	$u(a_1, \theta_K)$
a_2	$u(a_2, \theta_1)$	$u(a_2, \theta_2)$	$\cdots$	$u(a_2, \theta_K)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_J	$u(a_J, \theta_1)$	$u(a_J, \theta_2)$	$\cdots$	$u(a_J, \theta_K)$

Table 6.1: Utility table.

Table 6.2 present a toy decision problem where the choice is between bringing or not bringing an umbrella with you today. The consequences of this decision depends on the weather during the day. The best outcome is when you have chosen not to bring the umbrella and it turns out to be a sunny day. The worst outcome is when it rains and you left the umbrella home.

	Rain	Sun
No umbrella	−50	50
Umbrella	10	30

Table 6.2: Utility table.

Here are some more interesting decision problems.

SURGERY A surgeon needs to decide if a delicate surgery should be performed ($a = 1$) or not ($a = 0$). The surgery can be successful ($\theta = 1$) or lead to severe complications ($\theta = 0$). The probability of a successful operation can be computed based on the patient's characteristics. The utility function may be difficult to assess, but should involve the consequences of the patient as well as the cost of the operation. This is an example with discrete $\mathcal{A}$ and Θ.

CENTRAL BANK'S INTEREST RATE DECISIONS. An central bank with an explicit inflation target needs to continually decide the level of their lending rate (a) to simultanously reach an pre-determined target level for future inflation (θ_1) and to reduce future unemployment (θ_2). A simplified utility function could be

$$u(a, \boldsymbol{\theta}) = \omega \left(\theta_1(a) - \bar{\theta}_1\right)^2 - (1-\omega)\theta_2(a),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $\bar{\theta}_1$ is the inflation target, ω is the weight of the inflation target relative to the unemployment, and both unknowns θ_1 and θ_2 are functions of the interest rate, a. Here the set of actions $\mathcal{A}$ can be considered discrete (repo rate changes are in quarter percentage units) and Θ is two-dimensional and continuous.

PRICE REDUCTION ON ELECTRIC CARS. A government wants to give a price deduction on purchases of environmentally friendly electric cars (a) in an attempt to minimize future global warming. This is a complex decision problems with many unknowns. The government may settle for the intermediate goal of maximizing the expected utility from the C02 reduction from the price deduction (θ), net of the monetary cost of the deduction. Both $\mathcal{A}$ and Θ are continuous spaces here.

FIRMS' STOCKING DECISIONS. Deciding how much of a product to keep in stock is a balance act where too much stock is costly in storage, and too little stock runs the risk of not being able to deliver on time. Let a be the number of items in stock, θ the unknown number of items demanded by the customers in the coming period and p the set price for the product. A utility function for the firm may have the form

$$u(a, \theta) = \begin{cases} p \cdot \theta - c_1(a - \theta) & \text{if } a \geq \theta \\ p \cdot a - c_2(\theta - a)^2 & \text{if } a < \theta, \end{cases}$$

where c_1 and c_2 are positive constants. In the first case, too much stock was kept ($a \geq \theta$) and the utility is the profit, i.e. revenue $p \cdot \theta$ minus stocking costs for unsold items (c_1 each). In the second case, the firm kept too small stock, can only sell a units and suffers a reputation cost of not being a trustworthy firm that delivers on time. The reputation cost is considered to be quadratic in the number of undelivered items (many people complaining on social media etc).

Maximizing expected utility

There have been a large number of heuristic decision rules proposed in the literature. As an example, one such rule is the **maximin rule**: choose the action that gives the highest utility if the worst possible outcome of θ happens. In the umbrella example in Table 6.2 we see that the maximin decision is to always carry un umbrella since the worst utility for this choice is 10 (it rains) wheras if you choose not to carry un umbrella, the utility could be as low as -50 if it rains. The problem with the minimax rule, and many other heuristics, is that it completely ignores the probability of rain. Always bringing an umbrella may be a decent rule for rainy Bergen in Norway, but not for sunny California.

maximin rule

The Bayesian solution to a decision problem is instead based on the **posterior expected utility** of an action

posterior expected utility

$$\bar{u}(a) \equiv \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}} \left[u(a, \boldsymbol{\theta}) \right] = \int u(a, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}, \tag{6.8}$$

from which the **optimal Bayesian decision** is to choose the action $a \in \mathcal{A}$ that maximizes posterior expected utility:

optimal Bayesian decision

$$a^\star = \arg\max_{a \in \mathcal{A}} \bar{u}(a). \tag{6.9}$$

The Bayesian decision rule is naturally based on averaging over the unknown $\boldsymbol{\theta}$ with respect to your best quantification of uncertainty, the posterior distribution; brake the Bayesian eggs and you can enjoy a Bayesian omelette.

Figure 6.10 illustrates the optimal Bayesian decision in the umbrella toy decision problem in Table 6.2. Note how the probability for rain must be at least 0.25 for the Bayesian to make the same decision

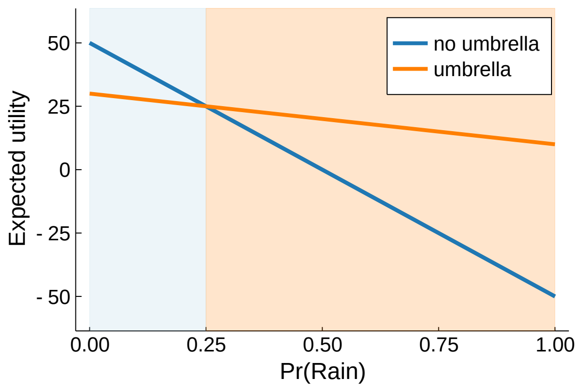

Figure 6.10: Expected utility of bringing an umbrella as function of the probability of rain. The shaded regions mark out the action that maximizes expected utility.

as the constantly umbrella carrying pessimist following the maximin rule.

An interesting feature of the Bayesian theory is that it implies the **separation principle**, i.e. that inference and decision problems can and should be kept separate:

separation principle

1. first learn a posterior distribution for the unknown state of the world $\boldsymbol{\theta}$ and then
2. set up a utility function $u(a, \boldsymbol{\theta})$ that values the consequence of actions $a \in \mathcal{A}$, to finally
3. choose the optimal action that maximizes posterior expected utility $\bar{u}(a)$.

Finding the optimal Bayesian decision involves computing the integral in (6.8), which is often analytically intractable. A simple approach is to compute the integral by Monte Carlo integration

$$\bar{u}(a) \equiv \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[u(a,\boldsymbol{\theta})\right] \approx m^{-1}\sum_{i=1}^{m} u(a,\boldsymbol{\theta}^{(i)}), \tag{6.10}$$

where $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)} \sim p(\boldsymbol{\theta}|\mathbf{x})$ are posterior draws. Expression (6.10) can be optimized numerically, see Chapter Classification, to find the approximate Bayes decision $a^{\star}$.

Point estimate as a decision problem

The chapter Single-parameter models presented ways of summarizing a posterior distribution by a measure of posterior location, e.g the posterior mean, median or mode. Choosing between these location measures is a decision problem where the action a is the **point estimate** of the unknown parameter θ. Reporting the estimate a when the unknown is really θ gives utility $u(a, \theta)$. For example, with a **quadratic utility** $u(a, \theta) = -(a-\theta)^2$, the optimal decision is to

point estimate

quadratic utility

summarize the posterior distribution $p(\theta|\mathbf{x})$ with the posterior mean, $\mathbb{E}(\theta|\mathbf{x})$. To see this, note that the negative posterior expected utility is

$$\mathbb{E}_{\theta|\mathbf{x}}(a-\theta)^2 = \mathbb{E}_{\theta|\mathbf{x}}(a-\mathbb{E}(\theta|\mathbf{x})-(\theta-\mathbb{E}(\theta|\mathbf{x})))^2 = (a-\mathbb{E}(\theta|\mathbf{x}))^2+\mathbb{V}(\theta|\mathbf{x}),$$

since the cross-term is zero by the fact that $\mathbb{E}_{\theta|\mathbf{x}}(\theta-\mathbb{E}(\theta|\mathbf{x})) = 0$. Maximizing the posterior expected utility is the same as minimizing $\mathbb{E}(a-\theta)^2$. Hence, since $\mathbb{V}(\theta|\mathbf{x})$ does not depend on a, the posterior mean $a = \mathbb{E}(\theta|\mathbf{x})$ is the optimal estimate for the quadratic utility function.

Similarly, one can show that the posterior median is optimal under the **linear utility** $u(a,\theta) = -|a-\theta|$. The posterior mode, the θ value with highest posterior density, seems like a sensible summary, but actually corresponds to the rather peculiar **zero-one utility**

linear utility

zero-one utility

$$u(a,\theta) = \begin{cases} 0 & \text{if } a = \theta \\ -1 & \text{if } a \neq \theta. \end{cases}$$

The zero-one utility hence gives a constant loss (negative utility) regardless of the size of the estimation error $a-\theta$, except when the estimate is spot on.

The linear, quadratic and zero-one utility are all symmetric in the error $a-\theta$. The following so called **lin-lin utility** function values over- and underestimation differently.

lin-lin utility

$$u(a,\theta) = \begin{cases} -c_1|a-\theta| & \text{if } a \leq \theta \\ -c_2|a-\theta| & \text{if } a > \theta. \end{cases}$$

where c_1 and c_2 are positive constants. A lin-lin loss is for example appropriate for budget spending prediction, where underestimation is worse than overestimation. The optimal estimate under lin-lin loss can be shown to be the $c_1/(c_1+c_2)\cdot 100\%$ **percentile** of the posterior distribution $p(\theta|\mathbf{x})$, i.e. the value that has exactly $c_1/(c_1+c_2)$ of the probability mass to the left. For example, with $c_1 = 9$ and $c_2 = 1$, i.e. the loss from underestimation is 9 times larger than for overestimation, the optimal estimate is the 90% percentile of $p(\theta|\mathbf{x})$.

percentile

The four presented utility function are plotted in Figure 6.11 as function of the estimation error $a-\theta$.

PROOFS

This section derives the predictive distribution for linear regression with a conjugate prior in Figure 6.4.

Since $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta}+\tilde{\varepsilon}$, and $\boldsymbol{\beta}$ and $\tilde{\varepsilon}$ are both normal, we immediately see that $p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}},\sigma^2,\mathbf{y})$ is multivariate normal with

$$\mathbb{E}(\tilde{\mathbf{y}}|\tilde{\mathbf{X}},\sigma^2) = \mathbb{E}(\tilde{\mathbf{X}}\boldsymbol{\beta})+\mathbb{E}(\tilde{\varepsilon}) = \tilde{\mathbf{X}}\boldsymbol{\mu}_n+0$$
$$\mathbb{V}(\tilde{\mathbf{y}}|\tilde{\mathbf{X}},\sigma^2) = \mathbb{V}(\tilde{\mathbf{X}}\boldsymbol{\beta})+\mathbb{V}(\tilde{\varepsilon}) = \tilde{\mathbf{X}}\sigma^2\Omega_n^{-1}\tilde{\mathbf{X}}^\top+\sigma^2\boldsymbol{I}_{\tilde{n}} = \sigma^2\tilde{\Sigma},$$

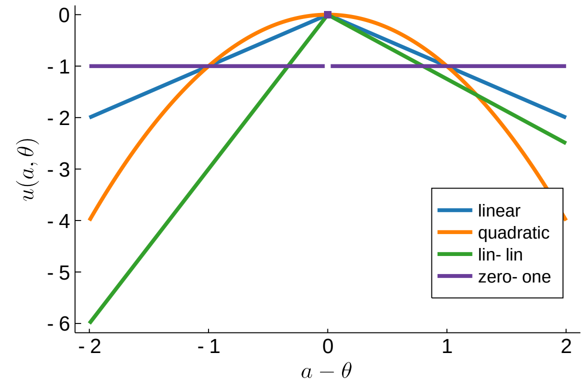

Figure 6.11: Utility functions for point estimation as a function of estimation error $a-\theta$. The lin-lin utility has $c_1 = 3$ and $c_2 = 1.25$.

where $\tilde{\Sigma} = \boldsymbol{I}_{\tilde{n}} + \tilde{\mathbf{X}}\Omega_n^{-1}\tilde{\mathbf{X}}^\top$; note that the expectation and variances are with respect to the posterior $p(\boldsymbol{\beta}|\sigma^2, \mathbf{y})$. Hence,

$$\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathbf{y}, \sigma^2 \sim N(\tilde{\mathbf{X}}\boldsymbol{\mu}_n, \sigma^2\tilde{\Sigma}).$$

Now, since $\sigma^2|\mathbf{y} \sim \text{Inv}-\chi^2(\nu_n, \sigma_n^2)$, we have

$$\begin{aligned}
p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathbf{y}) &= \int p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \sigma^2, \mathbf{y}) p(\sigma^2|\mathbf{y}) d\sigma^2 \\
&= \int |2\pi\sigma^2\tilde{\Sigma}|^{-1/2} \exp\Big(-\frac{1}{2\sigma^2}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\mu}_n)^\top\tilde{\Sigma}^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\mu}_n)\Big) \\
&\times \frac{(\nu_n\sigma_n^2/2)^{\nu_n/2}}{\Gamma(\nu_n/2)}(\sigma^2)^{-(\nu_n/2+1)} \exp\Big(-\frac{\nu_n\sigma_n^2}{2\sigma^2}\Big) d\sigma^2 \\
&= |2\pi\tilde{\Sigma}|^{-1/2}\frac{(\nu_n\sigma_n^2/2)^{\nu_n/2}}{\Gamma(\nu_n/2)} \\
&\times \int (\sigma^2)^{-((\nu_n+\tilde{n})/2+1)} \exp\left(-\frac{\nu_n\sigma_n^2 + a(\mathbf{y})}{2\sigma^2}\right) d\sigma^2 \\
&= (2\pi)^{-\tilde{n}/2}|\tilde{\Sigma}|^{-1/2}\frac{(\nu_n\sigma_n^2/2)^{(\nu_n+\tilde{n})/2}\Gamma((\nu_n+\tilde{n})/2)}{((\nu_n\sigma_n^2 + a(\mathbf{y}))/2)^{(\nu_n+\tilde{n})/2}\Gamma(\nu_n/2)}
\end{aligned}$$

where $a(\tilde{\mathbf{y}}) = (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\mu}_n)^\top\tilde{\Sigma}^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\mu}_n)$, and the last equality follows from the integrand begin proportional to a Inv$-\chi^2$ distribution. The density above can with a little bit of simple algebra be written as

$$\begin{aligned}
p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathbf{y}) &= \frac{\Gamma((\nu_n+\tilde{n})/2)}{\Gamma(\nu_n/2)(\pi\nu_n)^{\tilde{n}/2}|\sigma_n^2\tilde{\Sigma}|^{1/2}} \\
&\times \Big(1 + \frac{1}{\nu_n}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\mu}_n)^\top(\sigma_n^2\tilde{\Sigma})^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\mu}_n)\Big)^{-(\nu_n+\tilde{n})/2},
\end{aligned}$$

which can be recognized as the density of a multivariate student-t distribution.

EXERCISES

1. (a) Let $x_1, ..., x_n \overset{iid}{\sim} \text{Bern}(\theta)$, 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 $\text{Beta}(1,1)$ 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. Let $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{Expon}(\theta)$. Derive the predictive distribution for a new observation $\tilde{x}_{n+1}$.

3. (a) Let x_i be the number of sales of a product on month i. Let $x_1, ..., x_n \overset{iid}{\sim} \text{N}(\theta, \sigma^2)$ 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.

7 *Classification*

EXERCISES

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

NOTEBOOKS

1. See the notebook Classification.

8 Posterior simulation

Gibbs sampling

Markov Chain Monte Carlo

Hamiltonian Monte Carlo

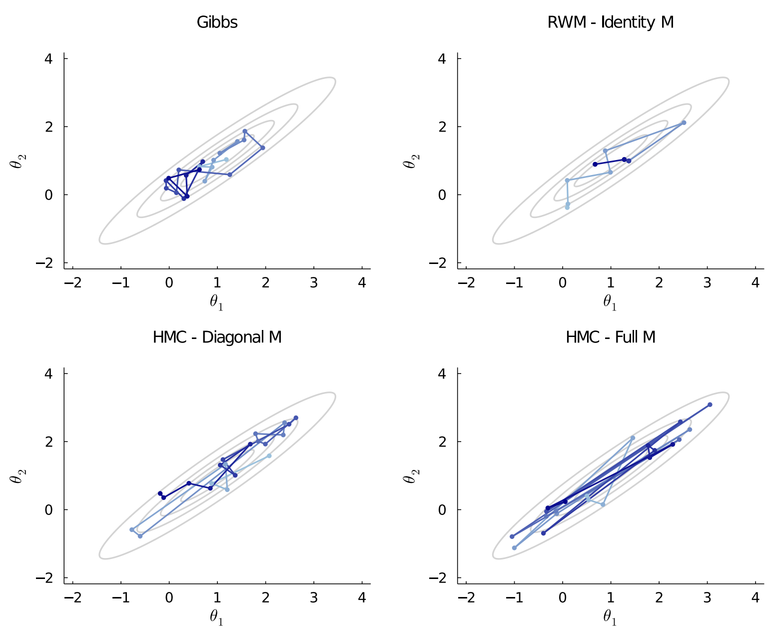

Figure 8.1: Comparing simulation paths of four algorithms for sampling from a bivariate normal target with $\boldsymbol{\mu} = (1, 1)^\top$, unit variances and correlation $\rho = 0.95$. The four compared algorithms are: i) Gibbs sampling, ii) random walk Metropolis with identity scaling, iii) HMC-NUTS with diagonal mass matrix and iv) HMC-NUTS with full mass matrix.

Probabilistic programming frameworks

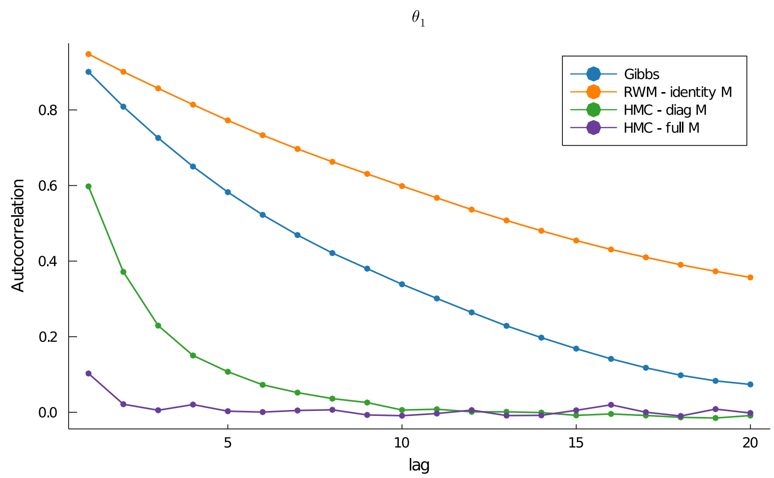

Figure 8.2: Comparing autocorrelation function of four posterior sampling algorithms for sampling from a bivariate normal target with $\boldsymbol{\mu} = (1,1)^\top$, unit variances and correlation $\rho = 0.95$. The four compared algorithms are: i) Gibbs sampling, ii) random walk Metropolis with identity scaling, iii) HMC-NUTS with diagonal mass matrix and iv) HMC-NUTS with full mass matrix.

Figure 8.3: Turing.jl code for the iid bernoulli model with a Beta prior.

```
using Turing, StatsPlots, Random

# Declare the Turing model:
@model function iidbern(y, α, β)
    θ ~ Beta(α,β)  # prior
    N = length(y)  # number of observations
    for n in 1:N
        y[n] ~ Bernoulli(θ) # model
    end
end

# Set up the observed data
data = [0,1,1,0,0,1,1,0,1,1]

# Settings for the Hamiltonian Monte Carlo (HMC) sampler.
niter = 10000
nburn = 1000
ϵ = 0.1
τ = 10

# Sample the posterior using HMC
postdraws = sample(iidbern(data, 1, 2), HMC(ϵ, τ), niter,
  discard_initial = nburn)
plot(postdraws)

# Print and plot results
display(postdraws)
plot(postdraws)
```

Figure 8.4: Turing.jl code for the iid normal model with a conjugate prior.

```
using Turing, StatsPlots, Random
ScaledInverseChiSq(ν,τ²) = InverseGamma(ν/2,ν*τ²/2) # Inv-χ² distribution

# Setting up the Turing model:
@model function iidnormal(x, μ₀, κ₀, ν₀, σ²₀)
    σ² ~ ScaledInverseChiSq(ν₀, σ²₀)
    θ ~ Normal(μ₀,σ²/κ₀)  # prior
    n = length(x)  # number of observations
    for i in 1:n
        x[i] ~ Normal(θ, √σ²) # model
    end
end

# Set up the observed data
x = [15.77,20.5,8.26,14.37,21.09]

# Set up the prior
μ₀ = 20; κ₀ = 1; ν₀ = 5; σ²₀ = 5^2

# Settings for the Hamiltonian Monte Carlo (HMC) sampler.
niter = 10000
nburn = 1000
α = 0.65 # target acceptance probability in No U-Turn sampler

# Sample the posterior using HMC
postdraws = sample(iidnormal(x, μ₀, κ₀, ν₀, σ²₀), NUTS(α), niter,
  discard_initial = nburn)

# Print and plot results
display(postdraws)
plot(postdraws)
```

Figure 8.5: Rstan code for the iid normal model with a conjugate prior.

```
library(rstan)

# Define the Stan model as a string
stanModelNormal = '
// The input data is a vector y of length N.
data {
  // data
  int<lower=0> N;
  vector[N] y;
  // prior
  real mu0;
  real<lower=0> kappa0;
  real<lower=0> nu0;
  real<lower=0> sigma20;
}

// The parameters in the model
parameters {
  real theta;
  real<lower=0> sigma2;
}

model {
  sigma2 ~ scaled_inv_chi_square(nu0, sqrt(sigma20));
  theta ~ normal(mu0,sqrt(sigma2/kappa0));
  y ~ normal(theta, sqrt(sigma2));
}
'

# Set up the observed data
data <- list(N = 5, y = c(15.77, 20.5, 8.26, 14.37, 21.09))

# Set up the prior
prior <- list(mu0 = 20, kappa0 = 1, nu0 = 5, sigma20 = 5^2)

# Sample from posterior using HMC
fit <- stan(model_code = stanModelNormal, data = c(data,prior), iter = 10000 )

# print and plot results
print(fit, pars = c("theta","sigma2"), probs=c(.1,.5,.9))
pairs(fit)
traceplot(fit, pars = c("theta", "sigma2"), nrow = 2)
```

9 Variational inference

10 Regularization

Polynomial regression As a first step toward flexible nonlinear regression modeling, let us consider the Gaussian polynomial regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

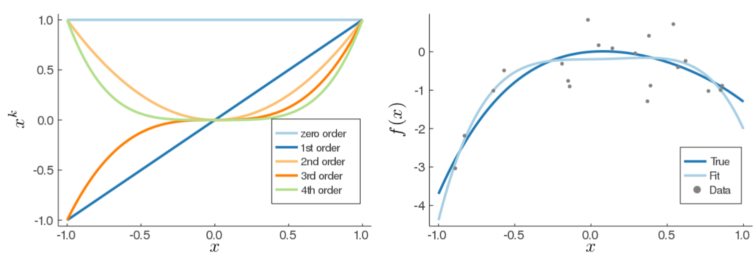

Figure 10.1: Polynomial regression.

L2-regularization and Ridge

L1-regularization and Lasso

Global-local regularization and Horseshoe

Laplace distribution

$X \sim \text{Laplace}(\mu, \beta)$ for $X \in \mathbb{R}$.

$$p(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right)$$

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

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

Figure 10.2: Laplace distribution.

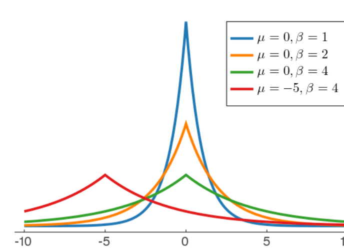

Figure 10.3: Some Laplace distributions.

11 Model comparison

Posterior model probabilities and the marginal likelihood

In most applications we have more than one potential model for the data. For example, count data can be modelled with a Poisson, geometric or negative binomial distribution. Income data can be modelled by a log-normal or a Gamma distribution. In regression analysis we usually have a multitude of models formed from different combinations of the covariates. This variable selection problem will be discussed in detail chapter Variable selection.

Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ denote the set of potential models for a dataset $\mathbf{x}$. Each model has its own set of parameters, $\boldsymbol{\theta}_k$ for model M_k. Consider first the rather unrealistic $\mathcal{M}$**-closed** case where one of these models are believed to be the **data generating process** (DGP). The Bayesian solution to the model comparison problem is then clear: compute the posterior distribution for the unknown true model $M \in \mathcal{M}$:

$\mathcal{M}$-closed

data generating process

$$\Pr(M = M_k|\mathbf{x}) \propto p(\mathbf{x}|M_k) \cdot \Pr(M_k), \tag{11.1}$$

where $\Pr(M = M_k)$ is the prior distribution over $\mathcal{M}$ and $p(\mathbf{x}|M_k)$ is the probability of the observed data $\mathbf{x}$ in model M_k. Table 11.1 is an example where a uniform prior distribution over four models $\mathcal{M} = \{M_1, \ldots, M_4\}$ is updated to posterior distribution; after observing the data, model M_2 is the most probable model.

	M_1	M_2	M_3	M_4
$\Pr(M_k)$	0.25	0.25	0.25	0.25
$\Pr(M_k|\mathbf{y})$	0.05	0.81	0.10	0.04

Table 11.1: Example of prior-to-posterior updating of model probabilities.

The likelihood contribution to (11.1), $p(\mathbf{x}|M_k)$, does not condition on the parameters $\boldsymbol{\theta}_k$ in model M_k; the parameters have been marginalized out and

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\boldsymbol{\theta}_k, M_k) p(\theta_k|M_k) d\theta_k, \tag{11.2}$$

is therefore usually called the **marginal likelihood**. The alternative name **evidence** is often used in machine learning. It is important to

marginal likelihood

evidence

note that the parameter are integrated out by the *prior* and that the marginal likelihood is the prior expected likelihood function:

$$p(\mathbf{x}|M_k) = \mathbb{E}_{\theta_k}\left(p(\mathbf{x}|\boldsymbol{\theta}_k, M_k)\right). \tag{11.3}$$

The marginal likelihood is therefore the **prior predictive distribution** for the training data $p(\mathbf{x}|M_k)$ when the parameters are drawn from the prior distribution. The marginal likelihood $p(\mathbf{x}|M_k)$ is therefore typically much more sensitive to the prior $p(\theta_k|M_k)$ than the posterior $p(\theta_k|\mathbf{x}, M_k)$ for the model parameters. We will explore this prior sensitivity in this chapter, and also present some alternative model comparison measures that are less sensitive to the prior.

prior predictive distribution

The **Bayes factor** comparing model M_1 to model M_2 is defined as

Bayes factor

$$B_{12}(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}. \tag{11.4}$$

The (modified) Jeffreys' scale of evidence [Kass and Raftery, 1995] is often used to interpret the strength of evidence of a Bayes factor:

- Barely worth mentioning: 1-3
- Positive: 3-20
- Strong: 20-150
- Very strong: > 150.

This scale is rather arbitrary, but can potentially be useful as a rough guide.

Bernoulli model

Let $x_1, ..., x_n \overset{\text{iid}}{\sim} \text{Bern}(\theta)$ and assume the prior $\theta \sim \text{Beta}(\alpha, \beta)$. The marginal likelihood is then

$$\begin{aligned}
p(x_1, ..., x_n) &= \int p(x_1, ..., x_n|\theta)p(\theta)d\theta \\
&= \int \theta^s(1-\theta)^f \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \frac{1}{B(\alpha,\beta)}\int \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}d\theta \\
&= \frac{B(\alpha+s, \beta+f)}{B(\alpha,\beta)},
\end{aligned}$$

where the last equality follows since the integral is with respect to the kernel of the $\text{Beta}(a+s, b+f)$ density. Note that we need to retain the normalizing constant $1/B(\alpha, \beta)$ in the prior when computing a marginal likelihood; we are not allowed to use the proportional form of Bayes' theorem here.

Normal model

Consider first the iid Normal model $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$ with known σ^2. We will compare two versions of this model: a null model M_0 where $\theta = \mu_0$ exactly, and a model M_1 with unrestricted θ following $\theta \sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0)$ a priori. This can be seen as the Bayesian equivalent of testing a sharp null hypothesis $H_0 : \theta = \mu_0$ vs $H_1 : \theta \neq \mu_0$. Note that the prior in the unrestricted model M_1 is centered on the null hypothesis, which is sensible given the hypothesis testing setup.

The marginal likelihood for model M_1 is obtained by integrating the likelihood with respect to the prior for the unknown θ:

$$p(\mathbf{x}|M_1) = \int \prod_{i=1}^{n} N(x_i|\theta, \sigma^2)\mathrm{N}(\theta|\mu_0, \sigma^2/\kappa_0)d\theta. \tag{11.5}$$

This integral can be calculated by completing the squares in the exponentials of the two Gaussian densities and integrating out θ using properties of the normal density. We will take a different route here that highlights the role of the sample mean $\bar{x}$ in the Bayes factor comparing M_0 to M_1.

Using the same algebra as when deriving the posterior for θ in the normal model in chapter Single-parameter models we can express the likelihood as

$$\begin{aligned} p(\mathbf{x}|\theta, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{(n-1)s^2}{2\sigma^2}\Big) \exp\Big(-\frac{1}{2\sigma^2} n(\bar{x}-\theta)^2\Big) \\ &= c(\sigma^2, s^2) N(\bar{x}|\theta, \sigma^2/n), \end{aligned}$$

where $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $c(\sigma^2, s^2) = (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{ns^2}{2\sigma^2}\Big)(2\pi\sigma^2/n)^{1/2}$ and $N(\bar{x}|\theta, \sigma^2/n)$ denotes the density function of the sample mean: $\bar{x}|\theta, \sigma^2 \sim N(\theta, \sigma^2/n)$. The constant $c(\sigma^2, s^2)$ will be shown to appear in both $p(\mathbf{x}|M_0)$ and $p(\mathbf{x}|M_1)$, and will therefore cancel out in the Bayes factor.

The marginal likelihood under M_0 is trivial since this model does not contain any unknown parameters, so we just insert $\theta = \mu_0$ in the likelihood:

$$p(\mathbf{x}|M_0, \sigma^2) = c(\sigma^2, s^2) N(\bar{x}|\mu_0, \sigma^2/n).$$

The marginal likelihood for model M_1 is

$$\begin{aligned} p(\mathbf{x}|M_1, \sigma^2) &= \int p(\mathbf{x}|\theta)p(\theta)d\theta \\ &= c(\sigma^2, s^2)\int N(\bar{x}|\theta, \sigma^2/n)\mathrm{N}(\theta|\mu_0, \sigma^2/\kappa_0)d\theta. \end{aligned}$$

We have seen an similar integral when deriving the predictive distribution for the iid Gaussian model, $p(\tilde{x}|\mathbf{x}) = \int \mathrm{N}(\tilde{x}|\theta, \sigma^2)\mathrm{N}(\theta|\mu_n, \tau_n^2)d\theta$

as $\mathrm{N}(\tilde{x}|\mu_n, \sigma^2 + \tau_n^2)$. Analogous arguments shows that

$$p(\mathbf{x}|M_1, \sigma^2) = c(\sigma^2, s^2) N\big(\bar{x}|\mu_0, \sigma^2(1/n + 1/\kappa_0)\big), \qquad (11.6)$$

and the Bayes factor for a given σ^2 is

$$\mathrm{BF}_{01}(\mathbf{x}, \sigma^2) = \frac{p(\mathbf{x}|M_0, \sigma^2)}{p(\mathbf{x}|M_1, \sigma^2)} = \frac{\mathrm{N}(\bar{x}|\mu_0, \sigma^2/n)}{\mathrm{N}\big(\bar{x}|\mu_0, \sigma^2(1/n + 1/\kappa_0)\big)}. \qquad (11.7)$$

The expression in (11.7) shows that the Bayes factor compares prior predictive densities for the two models with respect to the data compressed into the sufficient statistic $\bar{x}$. We can also clearly see the limiting behavior of BF_{01} with respect to the prior sample size κ_0:

- $B_{01} \to 1$ as $\kappa_0 \to \infty$. The prior under M_1 tends to a point mass at $\theta = \mu_0$ when $\kappa_0 \to \infty$, and M_0 and M_1 are therefore identical models in the limit.

- $B_{01} \to \infty$ as $\kappa_0 \to 0$, regardless of how close $\bar{x}$ is to μ_0. This is the case since the $\mathbb{V}(\bar{x}|M_1) = \sigma^2(1/n + 1/\kappa_0) \to \infty$ as $\kappa_0 \to 0$; model M_1 therefore assigns lower and lower predictive density to the observed $\bar{x}$ when $\kappa_0 \to 0$. A marginal likelihood evaluates the combination of a likelihood and a prior; if you make your prior "stupid" enough, the simpler null model M_0 will eventually win, even when $\bar{x}$ is not very likely to come from M_0.

The Bayes factor when the variance is assumed unknown is obtained by integrating $p(\mathbf{x}|M_0, \sigma^2)$ and $p(\mathbf{x}|M_1, \sigma^2)$ with respect to the $\sigma^2 \sim \mathrm{Inv}{-}\chi^2(\nu_0, \sigma_0^2)$ prior. The end result is a ratio of two student-t distributions for $\bar{x}$ and is not given here.

Internet speed data. Figure 11.1 plots the Bayes Factor comparing M_0: $N(20, 5^2)$ to M_1: $N(\theta, 5^2)$ for the internet speed data as a function of the prior sample size κ_0. The shaded region marks out the κ_0 where $\mathrm{BF}_{01} > 1$, i.e. where the evidence supports M_0. The regions for "barely worth mentioning" in the Jeffreys scale of evidence for is marked out by horizontal orange dashed lines. Unless the prior is very spread out, there is no evidence in favor of either model.

Figure 11.2 illustrates how the prior predictive density assigns increasingly lower density to the observed $\bar{x} = 15.99$ when κ_0 decreases.

Figure 11.3 illustrates the Bayes factor for the internet speed data with $\bar{x}$ artificially changed from 15.99 to $\bar{x} = 12$; the figure plots both the Bayes factor the Jeffreys scale of evidence in logs for visibility. With $\bar{x}$ so far from the null value $\mu_0 = 20$, there is now positive or even close to strong evidence in favor of M_1 for all $\kappa_0 \in (0.1, 1)$. This is also clear from Figure 11.2 if we move the purple data point to $\bar{x} = 12$.

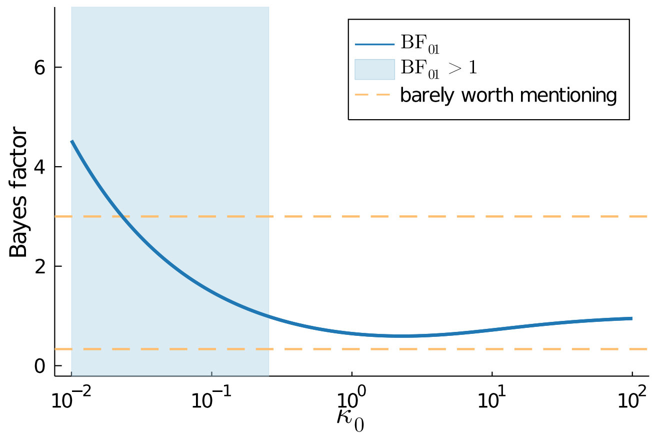

Figure 11.1: Bayes factor for the internet speed data with known variance $\sigma^2 = 5^2$. The graph plots the Bayes factor BF_{01} as function of the prior sample size κ_0 in log-scale. The shaded region shows the values for κ_0 where $BF_{01} > 1$, i.e. where there is support in favor of the null model. The limits for "barely worth mentioning" in the Jeffreys scale of evidence is marked out horizontal orange dashed lines.

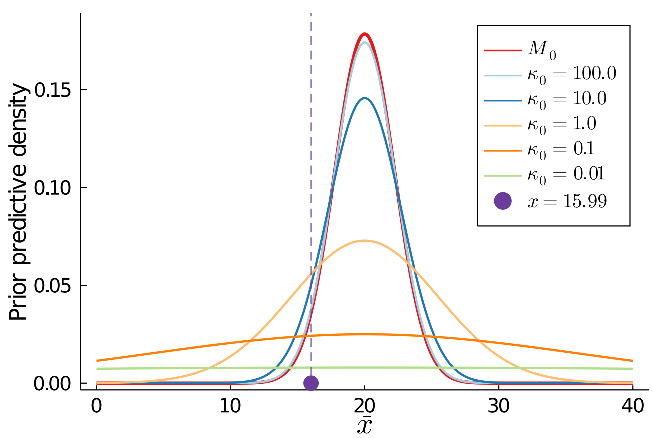

Figure 11.2: Internet speed data with known variance $\sigma^2 = 5^2$. Prior predictive densities for $\bar{x}$ in the models M_0 and M_1 for different values of the prior hyperparameter κ_0. The realized data of $\bar{x} = 15.99$ is shown as the purple dot with dashed line.

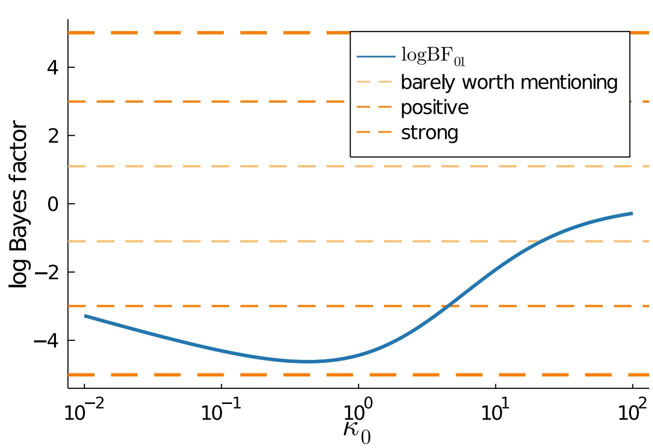

Figure 11.3: Log Bayes factor for the internet speed data with $\bar{x}$ artifically set to $\bar{x} = 12$ instead of the actually observed $\bar{x} = 15.99$. The graph plots the log Bayes factor BF_{01} as function of the prior sample size κ_0 in log-scale. The limits for Jeffreys scale of evidence (in logs) is marked out horizontal dashed lines.

Properties of posterior model probabilities

GEOMETRIC VS POISSON

Consider count data and the comparison of the two models:

- M_1: $x_1, \ldots, x_n | \theta_1 \overset{\text{iid}}{\sim} \text{Geo}(\theta_1)$ with prior $\theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$
- M_1: $x_1, \ldots, x_n | \theta_2 \overset{\text{iid}}{\sim} \text{Poisson}(\theta_2)$ with prior $\theta_2 \sim \text{Gamma}(\alpha_2, \beta_2)$.

The marginal likelihoods are (see Exercise X)

$$\begin{aligned} p(x_1, \ldots, x_n | \mathrm{M}_1) &= \int p(x_1, \ldots, x_n | \theta_1, \mathrm{M}_1) p(\theta_1 | \mathrm{M}_1) d\theta_1 \\ &= \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(n + \alpha_1)\Gamma(n\bar{y} + \beta_1)}{\Gamma(n + n\bar{y} + \alpha_1 + \beta_1)}. \end{aligned}$$

and

$$\begin{aligned} p(x_1, \ldots, x_n | \mathrm{M}_2) &= \int p(x_1, \ldots, x_n | \theta_2, \mathrm{M}_2) p(\theta_2 | \mathrm{M}_2) d\theta_2 \\ &= \frac{\Gamma(n\bar{y} + \alpha_2)\beta_2^{\alpha_2}}{\Gamma(\alpha_2)(n + \beta_2)^{n\bar{y} + \alpha_2}} \frac{1}{\prod_{i=1}^n y_i!}. \end{aligned}$$

For consistency, we set $\alpha_1/\beta_1 = \beta_2/\alpha_2$ so that both models have the same prior predictive mean, $\mathbb{E}(\tilde{x}|\mathrm{M}_1) = E(\tilde{x}|\mathrm{M}_2)$ [Bernardo and Smith, 2009]. We will specifically use $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 10$ in the illustrations, and equal prior model probabilities $\Pr(\mathrm{M}_1) = \Pr(\mathrm{M}_2) = 1/2$.

To investigate how the posterior model probabilities $\Pr(M_1|\mathbf{x})$ and $\Pr(M_2|\mathbf{x})$ behave the sample size grows large, I simulate a data set with $n = 500$ from the $\text{Pois}(\theta_2 = 1)$ model, so the M_2 is the true data generating process. We then compute $\Pr(M_2|\mathbf{x})$ sequentially using a larger and larger sample size until all $n = 500$ observations have been used up. Figure 11.6 shows the results from this experiment repeated four times to also see the sampling variation. The graph to the left in Figure 11.6 zooms in on the first $n = 100$ observations; there is quite some sampling variability in the model probabilities, but there is a clear tendency for the posterior probability on the Poisson model to tend to 1. The right hand graph shows the results for the full sample of $n = 500$ observations; the probability $\Pr(M_2|\mathbf{x})$ clearly tends to 1 for all four replications.

The asymptotic behavior in Figure 11.6 is what one would expect, and one can indeed prove that Bayesian posterior model probabilities are consistent in the $\mathcal{M}$-closed setting where the data generating process is among the compared models:

$$\Pr(\mathrm{M}_k^\star|\mathbf{x}) \xrightarrow{p} 1 \text{ as } n \to \infty, \tag{11.8}$$

Geometric distribution

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

$$p(x) = (1-\theta)^x\theta$$

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

$$\mathbb{V}(X) = \frac{1-\theta}{\theta^2}$$

Figure 11.4: The Geometric distribution.

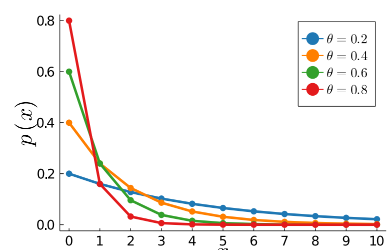

Figure 11.5: Some Geometric distributions.

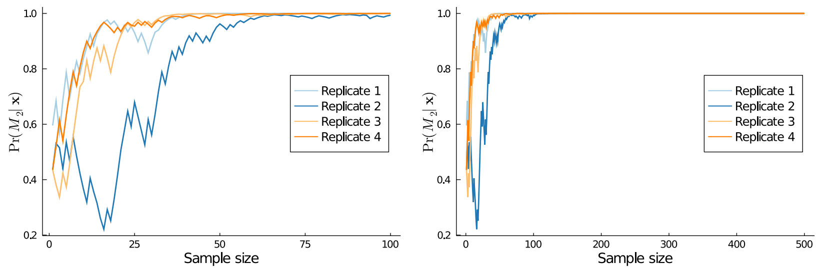

Figure 11.6: Asymptotic behavior of posterior model probabilities in $\mathcal{M}$-closed when comparing the models:
M_1: Geo(θ_1), $\theta_1 \sim$ Beta$(10, 10)$
M_2: Pois(θ_2), $\theta_2 \sim$ Gamma$(10, 10)$.
The graphs shows the evolution of the posterior probability for the Poisson model as the sample size increases. Each line corresponds to a replication of the experiment. The data are generated from the iid Pois(1) model. The left graph shows the subset of the first 100 data points and the right graph shows all 500 data points.

where $M_k^\star$ is the data generating process.

What happens asymptotically when the data generating process is not among the compared models? This $\mathcal{M}$**-open** setting is more realistic since models are typically just approximations to reality. To explore this let us change previous experiment and generate data from a **negative binomial distribution** in a slightly different form from the one encountered in chapter Single-parameter models:

$\mathcal{M}$-open

negative binomial distribution

$$p(x) = \binom{x+r-1}{x}(1-\theta)^r\theta^x, \text{ for } x = 0, 1, 2, ... \qquad (11.9)$$

Negative binomial distribution

$X \sim$ NegBin(r, θ)
Support: $X \in \{0, 1, \ldots\}$

$$p(x) = \binom{x+r-1}{x}(1-\theta)^r\theta^x$$

$$\mathbb{E}(X) = \frac{r\theta}{1-\theta}$$

$$\mathbb{V}(X) = \frac{r\theta}{(1-\theta)^2}$$

Figure 11.7: The Negative binomial distribution.

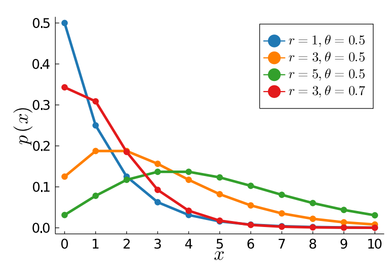

Figure 11.8: Some Negative binomial distributions.

The negative binomial in Single-parameter models was the total number of trials until a certain number of successes. The negative binomial in (11.9) instead counts the number of successes x before the rth failure occurs.

Figure 11.9 (left) shows the asymptotic behaviour of the posterior model probabilities for the Poisson and Geometric models when both models are wrong and data actually comes from the NegBin$(2, 0.5)$ distribution; the posterior probabilities seem to converge to a solution where the Geometric model gets a probability of one as n grows.

The right hand graph of the figure explains why this is happening by plotting the NegBin$(2, 0.5)$ data generating distribution as a bar chart with the optimal fit of each compared model overlayed. The optimal fit is defined as the fit that minimizes the Kullback-Leibler divergence of the model from the data generating process. Specifically, let $g_\theta(x)$ be the data density of a model and let $f(x)$ denote the data generating process. The optimal fit for the model $g_\theta(x)$ is then obtained by minimizing the Kullback-Leibler divergence

$$d(f, g) = \int \log\left(\frac{f(x)}{g_\theta(x)}\right) f(x) dx$$

with respect to the model parameters θ.

The legend of Figure 11.9 (right) shows that the Geometric model is closer to the data generating process (smaller KL divergence) than the Poisson model which explains why the geometric model wins asymptotically. The asymptotic tendency seen in Figure 11.9 can proved to hold quite generally in that

$$\Pr(\mathrm{M}_k^\star|\mathbf{x}) \xrightarrow{p} 1 \text{ as } n \to \infty, \tag{11.10}$$

where $\mathrm{M}_k^\star$ is the model in $\mathcal{M}$ with smallest Kullback-Leibler divergence to the data generating process.

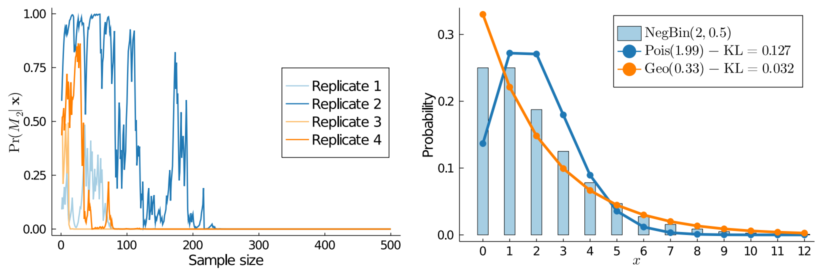

Figure 11.9: Asymptotic behavior of posterior model probabilities in $\mathcal{M}$-open when comparing the models:
M_1: $\mathrm{Geo}(\theta_1)$, $\theta_1 \sim \mathrm{Beta}(10, 10)$
M_2: $\mathrm{Pois}(\theta_2)$, $\theta_2 \sim \mathrm{Gamma}(10, 10)$.
The left graph shows the evolution of the posterior probability for the Poisson model as the sample size increases. Each line corresponds to a replication of the experiment. The data are generated from the iid $\mathrm{NegBin}(2, 0.5)$ model. The right graph shows the fit of the models with KL-optimal parameters.

Marginal likelihood in linear regression

The marginal likelihood for the linear regresson model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 I_n), \tag{11.11}$$

is given by

$$p(\mathbf{y}|\mathbf{X}) = \iint p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) p(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2. \tag{11.12}$$

The marginal likelihood is a special case of the posterior predictive distribution in Figure 6.4 when the posterior is based on $n = 0$ data points, i.e. when the parameters are integrated with respect to the prior, and the object of prediction is the training data $\mathbf{y}$; for this reason, the marginal likelihood is sometimes called the **prior predictive distribution**. Note that the marginal likelihood is not measuring in-sample training error since the prediction for the training data $\mathbf{y}$ is only using prior information for the model parameters $\boldsymbol{\beta}$ and σ^2. Hence setting $n = 0$ and $\tilde{\mathbf{y}} = \mathbf{y}$ we immediately have the marginal likelihood for the linear regression model

prior predictive distribution

$$\mathbf{y}|\mathbf{X} \sim t_{\nu_0}\left(\mathbf{X}\boldsymbol{\mu}_0, \sigma_0^2(\boldsymbol{I}_n + \mathbf{X}\Omega_0^{-1}\mathbf{X}^\top)\right). \tag{11.13}$$

TODO! add comparison of models with different predictors for the salaries data. Then point forward to variable selection chapter.

The Laplace approximation of the marginal likelihood

There are many methods for approximating the marginal likelihood when it cannot be derived analytically. An obvious approach comes from the marginal likelihood being the prior expected likelihood

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \mathbb{E}_{p(\boldsymbol{\theta})}p(\mathbf{x}|\boldsymbol{\theta}),$$

and can therefore be computed by simple Monte Carlo simulation

$$\widehat{p(\mathbf{x})} = \frac{1}{m}\sum_{i=1}^{m} p(\mathbf{x}|\boldsymbol{\theta}^{(i)}), \qquad (11.14)$$

where $\boldsymbol{\theta}^{(i)} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta})$ are m draws from the prior.

Unfortunately, the simple Monte Carlo estimator in (11.14) usually has disastrously large variance and is rarely used in practice. The problem with the estimator in (11.14) is that the likelihood is often much more concentrated than the prior and the estimate will then be dominated by the few prior draws that happen to end up where the likelihood is concentrated. Importance sampling can be used to reduce the variance, see for example the modified harmonic estimator in [Geweke, 1999]. There are also many methods based on MCMC, in particular Chib's methods for Gibbs sampling [Chib, 1995] and its extension to Metropolis-Hastings [Chib and Jeliazkov, 2001]. We will here present a simple but often quite accurate method for approximating the marginal likelihood, the Laplace approximation.

The Laplace approximation of the log marginal likelihood for a model with p parameters is

$$\ln \hat{p}(\boldsymbol{x}) = \ln p(\boldsymbol{x}|\hat{\theta}) + \ln p(\hat{\theta}) + (1/2)\ln\left|J^{-1}_{\boldsymbol{x},\hat{\theta}}\right| + (p/2)\ln(2\pi),$$

where $\hat{\theta}$ is the posterior mode and $|J_{\boldsymbol{x},\hat{\theta}}|$ is the determinant of the observed information matrix as in Chapter Multi-parameter models, but here defined for the posterior instead of the likelihood:

$$J_{\hat{\boldsymbol{\theta}},\mathbf{x}} = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \qquad (11.15)$$

where $\hat{\boldsymbol{\theta}}$ is the posterior mode.

BERNOULLI MODEL We have already computed the marginal likelihood for the Bernoulli model in closed form earlier in this chapter, so there is really no need to approximate it. However, it gives us a chance to practice deriving the marginal likelihood and we can also assess how accurate the approximation is since we know the true answer here. We have:

$$\ln p(\boldsymbol{y}|\theta)p(\theta) = (\alpha + s - 1)\ln\theta + (\beta + f - 1)\ln(1-\theta)$$
$$\frac{\partial \ln p(\boldsymbol{y}|\theta)p(\theta)}{\partial\theta} = \frac{\alpha+s-1}{\theta} - \frac{\beta+f-1}{1-\theta}$$
$$\frac{\partial^2 \ln p(\boldsymbol{y}|\theta)p(\theta)}{\partial\theta^2} = -\frac{\alpha+s-1}{\theta^2} - \frac{\beta+f-1}{(1-\theta)^2}$$

Solving $\partial \ln p(\boldsymbol{y}|\theta)p(\theta)/\partial\theta = 0$ for θ gives the posterior mode

$$\hat{\theta} = \frac{\alpha+s-1}{\alpha+\beta+n-2}.$$

and therefore

$$J^{-1}_{\boldsymbol{x},\hat{\theta}} = -\left[\frac{\partial^2 \ln p(\theta|\boldsymbol{x})}{\partial\theta^2}|_{\theta=\hat{\theta}}\right]^{-1} = \frac{(\alpha+s-1)(\beta+f-1)}{(\alpha+\beta+n-2)^3}.$$

To examine the accuracy of this approximation, let us consider a dataset with $s = 6$ successes in $n = 10$ trials and the uniform prior with $\alpha = \beta = 1$. Here, $\hat{\theta} = s/n = 0.6$ and $J^{-1}_{\boldsymbol{x},\hat{\theta}} = sf/n^3 = 0.024$. So, $\ln \hat{p}(\boldsymbol{x}) = -7.676$ which is quite close to the true log marginal likelihood $\ln p(\boldsymbol{x}) = -7.745$. Consider for example using this marginal likelihood for comparing a model against a null model where $\theta = 1/2$. The true Bayes factor is then $0.5^{10}/\exp(-7.745) \approx 2.559$ and the Bayes factor from the Laplace approximation is $0.5^{10}/\exp(-7.676) \approx 2.105$; the approximate Bayes factor and the exact Bayes factor both lead to the conclusion that the evidence in favor of the null model is 'barely worth mentioning' according to the Jeffreys' scale of evidence.

Log predictive score

The marginal likelihood is by construction usually sensitive to the exact specification of the prior. A precise prior elicitation is sometimes hard, or at least time-consuming, particularly in models with many parameters where the prior dependence can be especially hard to get right. Several alternative measures for Bayesian model comparison that are less sensitive to prior have therefore been developed. The log predictive score measure in this section sacrifices some data to make the marginal likelihood more robust to variations in the prior.

The marginal likelihood is the joint prior predictive distribution for all observations and can therefore be decomposed as sequences of conditional densities:

$$p(x_1,\ldots,x_n) = p(x_1)p(x_2|x_1)\cdots p(x_1|x_1,x_2,\ldots,x_{n-1}) \tag{11.16}$$

The ith factor in this decomposition is the intermediate predictive density

$$p(x_i|x_1,...,x_{i-1}) = \int p(x_i|x_1,\ldots,x_{i-1},\boldsymbol{\theta})p(\boldsymbol{\theta}|x_1,...,x_{i-1})d\boldsymbol{\theta},$$

where $p(\boldsymbol{\theta}|x_1,...,x_{i-1})$ is the intermediate posterior for $\boldsymbol{\theta}$ conditional on the data subset $x_1,...,x_{i-1}$. For iid data we have the usual simplification $p(x_i|x_1,\ldots,x_{i-1},\boldsymbol{\theta}) = p(x_i|\boldsymbol{\theta})$.

In a time series context where the observations has a natural ordering in time, the factor $p(x_i|x_1,...,x_{i-1})$ in the decomposition in (11.16) is the one-step-ahead predictive distribution for the observation at time i given data up to time $i-1$. When the data are not specifically ordered, for example iid data, the decomposition in (11.16) can be done in many different ways by ordering the observations differently; we return this interpretation later in this section.

The decomposition in (11.16) is interesting for at least three reasons. First, it can be used to diagnose why a model has a low marginal likelihood by inspecting each of the terms in the decomposition to see which observations are poorly predicted. Second, it gives a clear connection between the marginal likelihood and sequential out-of-sample predictive performance of a model, particularly for time series data. Third, the decomposition in (11.16) can be used to highlight the effect of the prior on the marginal likelihood and suggest a way to reduce the influence of the prior in Bayesian model comparisons using the marginal likelihood.

To elaborate on this last point consider the iid Normal model with known variance: $x_1,\ldots,x_n|\theta \overset{\text{iid}}{\sim} \mathrm{N}(\theta,\sigma^2)$ with prior $\theta \sim N(\mu_0,\sigma^2/\kappa_0)$. We will be particularly interested in the sensitivity of the marginal likelihood with respect to κ_0. The intermediate predictive distribution for observation x_i in decomposition (11.16) is

$$x_i|x_1,\ldots,x_{i-1} \sim \mathrm{N}\left(\mu_{i-1},\sigma^2\left(1+\frac{1}{i-1+\kappa_0}\right)\right), \tag{11.17}$$

where $\mu_{i-1} = w_{i-1}\bar{x}_{i-1} + (1-w_{i-1})\mu_0$, $\bar{x}_{i-1}$ is the sample mean of the first $i-1$ observations, and $w_{i-1} = (i-1)/(i-1+\kappa_0)$. This result is simply the predictive distribution for the Gaussian model with known variance in Prediction and Decision making with $n = i-1$ data points in the posterior.

Consider now $n = 100$ observations simulated from the $\mathrm{N}(20,5)$ distribution, to mimic the setting in the Internet speed data; the original dataset with only $n = 5$ observations is too small for the point I want to make here. The upper graph in Figure 11.10 plots the log of the marginal likelihood decomposition

$$\log p(x_1,\ldots,x_n) = \log p(x_1) + \log p(x_2|x_1) + \ldots + \log p(x_1|x_1,\ldots,x_{n-1}), \tag{11.18}$$

for three different values of κ_0. The log marginal likelihood for the model with $\kappa_0 = 1$ is for example the sum of the values in the orange line. A careful examination of the graph shows that the prior sensitivity of log marginal likelihoods is entirely driven by the first term

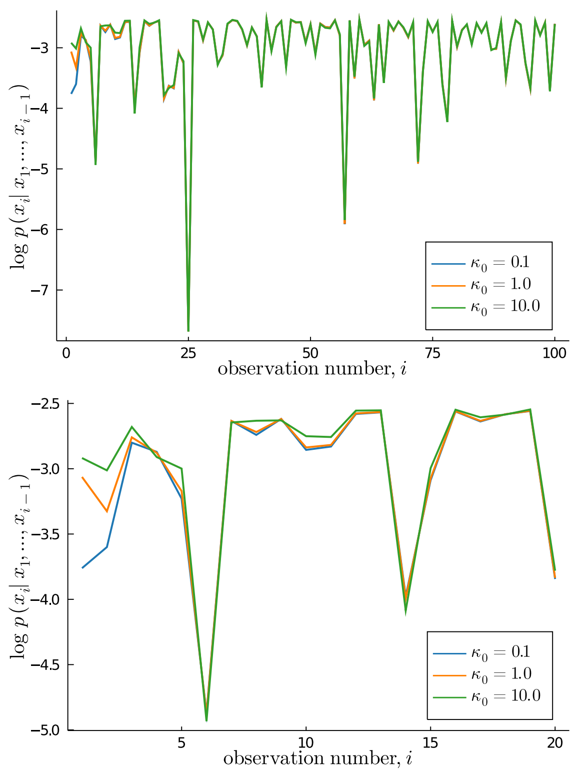

Figure 11.10: Decomposition of the log marginal likelihood for the simulated internet speed data with $n = 100$ observations for three different value for the prior sample size κ_0. The bottom graph zooms in on the gray shaded region with the first 20 observations.

in the decomposition (11.18). The lower graph in Figure 11.10 makes this more visible by zooming in on the 20 first observations. This comes as no surprise since it is clear from (11.17) that the first terms will be affected by κ_0, but the later terms in the sequence where i is large remain essentially unaffected by κ_0.

An obvious way of reducing the prior sensitivity while still remaining close to the marginal likelihood is therefore to discard the first terms in (11.18). This is the **log predictive score** (**LPS**):

log predictive score

$$\text{LPS} = \sum_{i=i^\star+1}^{n} \log p(x_i | x_1, \ldots, x_{i-1}). \tag{11.19}$$

The LPS is effectively using the first $i^\star$ observations to train the prior $p(\boldsymbol{\theta})$ into a intermediate posterior $p(\boldsymbol{\theta} | x_1, \ldots, x_{i^\star})$ which is then used as the new prior for the remaining test data $x_{i^\star+1}, \ldots, x_n$. There are also variants of LPS which scales by $1/(n - i^\star)$ so that the LPS is the average log predictive observation per test observation. The form in (11.19) has the advantage that Jeffreys scale of evidence can still be used since the number of terms in the LPS is the number of test observations; the training data have been sacrificed to reduce the sensitivity to the prior and can therefore not be used in the evidence for the model.

Figure 11.11 plots the LPS for the three κ_0 as a function of the training fraction $f = i^\star/n$. The LPS in the figure is scaled by $n/(n - i^\star)$ to keep the same scale on the LPS for all training fractions for presentation purposes. The LPS in Figure 11.11 with $f = 0$ is the original log marginal likelihood where the prior (κ_0) has a substantial effect. Already with a training fraction of 15% is the LPS is insensitive to $\kappa_0 \in [0.1, 10]$.

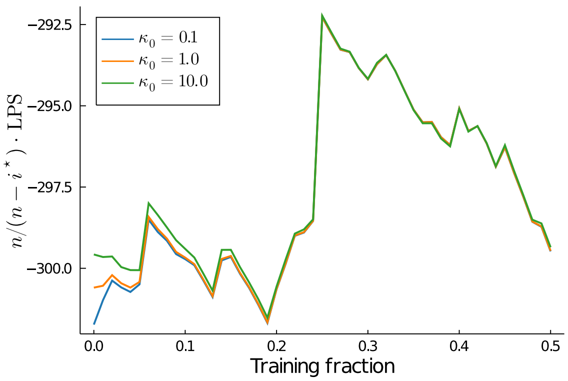

Figure 11.11: Log predictive score (scaled) as a function of the training fraction for the simulated internet speed data with $n = 100$ observations.

The LPS in (11.19) discards the *first* $i^\star$ observations. This makes sense for time series where the observations are ordered in time. For cross-sectional data, e.g. iid data, there is no natural ordering and it is then common practise to use a cross-validated version of the LPS. The idea with **K-fold cross-validated LPS** is to split, or partition, the data into K folds, use one of the K folds for training and then evaluating the preditive performance on the $K - 1$ folds left out. This is repeated K times, each time with a new fold as the training fold. Table 11.2 illustrates the data partitioning. Note that this is different from the usual cross-validation used in the machine learning where instead $K - 1$ folds would be used for training with one fold left out and used for testing. The reason is that cross-validation in machine learning aims at estimating the generalization performance of the model on future data. The cross-validated LPS still aims for something close to the marginal likelihood, but uses cross-validation to lessen the arbitrary choice of which observations to use in the training and test when computing the LPS. Bayesian cross-validation that aims to estimate the generalization performance of the model are discussed in the next section.

K-fold cross-validated LPS

Bayesian estimators of generalization performance

Leave-one-out and cross-validation. WAIC

	n data observations				
	$1, 2, \dots\dots\dots\dots, n-1, n$				
Split 1:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Table 11.2: The data partitioning for 5-fold cross-validation of the LPS. For each of the K splits, the observations in each fold in blue is used to train the prior into a posterior. The observations in the remaining folds in the same row are used to compute the LPS for the split.

12 Variable selection

13 Gaussian processes

Gaussian processes

14 Mixture models

Finite mixtures

Mixtures of regressions

Latent Dirichlet allocation

Infinite mixtures

15 Dynamic models and sequential inference

Dynamic models

- Time-varying regression models
- State-space models (with control)

Bayesian filtering and smoothing

- The Kalman filter (Bayesian approach)
- Forward filtering backward smoothing

Sequential Monte Carlo

Basic particle filter

Sequential decision making

- Bayesian updating is key
- Markov Decision process
- Reinforcement learning
- Bellman's equation?

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.

Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

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

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013. DOI: 10.1007/s13748-013-0040-3. URL http://dx.doi.org/10.1007/s13748-013-0040-3.

John Fox and Sanford Weisberg. *An R companion to applied regression*. Sage publications, 2019.

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.

John Geweke. Using simulation methods for bayesian econometric models: inference, development, and communication. *Econometric reviews*, 18(1):1–73, 1999.

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.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

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