Mattias Villani

Bayesian Learning

A GENTLE INTRODUCTION

Available at: https://github.com/mattiasvillani/BayesianLearningBook

PUBLISHED BY AVAILABLE AT: HTTPS://GITHUB.COM/MATTIASVILLANI/BAYESIANLEARNINGBOOK

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

First edition, April 2023

Contents

To all who did Bayes before it was hip.

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. It is written so that it can accommodate 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 heterogeneous students, with roughly half of the students from statistics and the other half from engineering and computer science. 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. Students from both camps thought that the course was on the right level for them. 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. There are of course several overlapping concepts and probability is the underlying technical language (although with highly different interpretations), but there is nevertheless a lot of effort spent 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 to 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, before moving over to territory unknown to all students. This book is written in the same style using Tufte style mar-

gin notes and figures to fill in potential missing gaps in probability and statistics, without breaking the flow of the main text.

Some programming experience 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 the more traditional Bayesian *inference* or 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, statistical modelling, and computing. The ideal class is therefore a mix of students from nearby disciplines that learn from each other's 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.

1.1 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 inferring a one-dimensional measure of a person's intelligence from IQ tests.

parameter space

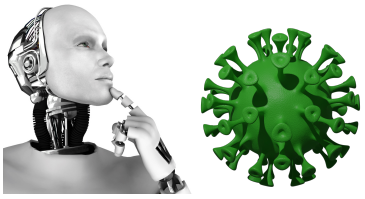


Figure 1.1: Artificial intelligence and infectious disease models are examples where Bayesian learning is often used for quantifying uncertainty.

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 its **probability mass function (pmf)** is

DISTRIBUTION BERNOULLI

probability mass function (pmf)

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

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. All distributions in this book are have interactive versions which can be explored by clicking on the distribution badge in the PDF version of the book.

Bernoulli trials

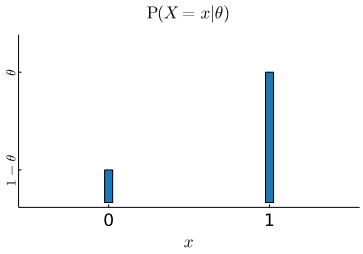


Figure 1.2: Bernoulli distribution with success probability $\theta = 0.8$.

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.

1.2 *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

likelihood function

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

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

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.

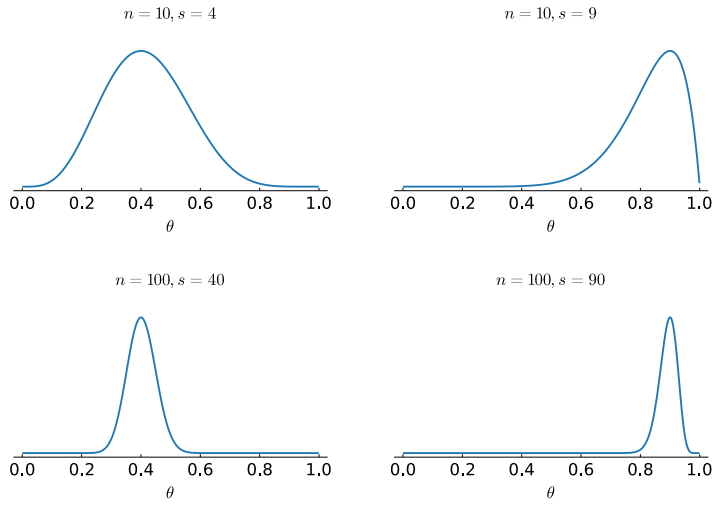


Figure 1.3: Bernoulli likelihood function for $n = 10$ and $s = 4$.

It is essential to have a 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 two grahs in the lower part of Figure 1.3 show results for $n = 100$ trials with the same success ratio s/n as in corresponding graphs 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 $p(x_1, \ldots, x_n|\theta)$ directly with respect to θ it is often easier to maximize the *log-likelihood function*

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

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

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

gives the unique solution $\hat{\theta}_{\text{MLE}} = s/n$, the fraction of successes in the data. It is straightforward to show that this is 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:

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

where we have written out explicitly that an estimator is function of the sample. Note that the number of successes is random in this calculation as we are considering the variability over all possible samples, hence the use of capital letter S. We have also used that if $X_1, \ldots, X_n|\theta \overset{iid}{\sim}$ Bernoulli then $S|\theta \sim$ Binomial(n, θ) with mean $E(S) = n\theta$; see Figure 1.5 for an example of a **Binomial distribution**.

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

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

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

It is important to understand that the above mean and variance of $\hat{\theta}_{MLE}$ are computed with respect to the **sampling distribution**, i.e. the distribution of the estimator as we repeatedly sample new datasets of size n from the assumed data generating process. They are **long-run properties** of the estimation method, telling us how the estimator would perform on average over many repeatedly sampled

Binomial distribution

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

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

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

Figure 1.4: The binomial distribution.

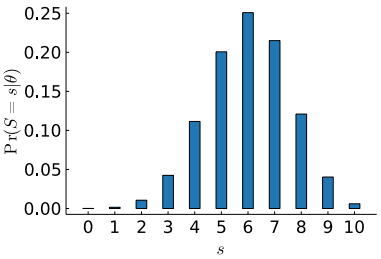


Figure 1.5: Binomial distribution with $n = 10$ and $\theta = 0.7$.

unbiased

Binomial distribution

sampling variance

sampling distribution

long-run properties

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.

When the data are recorded as **continuous random variables** the probability of any dataset is zero, and we instead define the likelihood function by letting $p(x_1, \ldots, x_n|\theta)$ be the joint probability density function (**pdf**) of the observed data. See Figure 1.6 for an illustration of a **probability density function**. When data are iid we can similarly define the likelihood as the product of the individual pdf's for each data point $p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$.

continuous random variables

pdf

probability density function

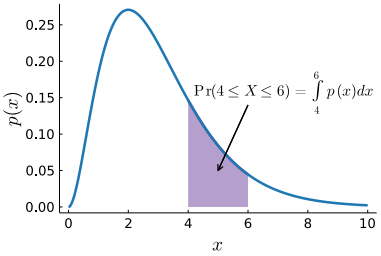


Figure 1.6: The probability density function (pdf) $p(x)$ for a continuous random variable X is a non-negative function that can be integrated to compute the probabilities of the form $\Pr(a \leq X \leq b)$.

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 make 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 a function of θ; the likelihood is therefore *not* a probability distribution for θ. Figure 1.7 reminds us of this error.

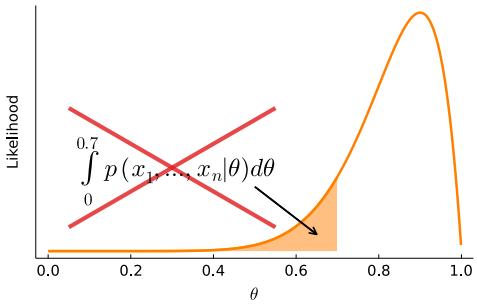
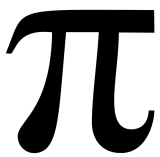


Figure 1.7: Areas under the likelihood function are **not** probabilities.

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, in some sense. And this is where our Bayesian story begins.

1.3 *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 Bayesian learning, 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 1.8: 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 persons have different 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, while someone that knows that this decimal is 5 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" in connection to quantum mechanics is interesting to ponder about, but has no 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 of 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 under certain conditions; the frequentist definition instead requires that the event must be infinitely repeatable. 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 $\Pr(\mathrm{A} \cup \mathrm{B}) = \Pr(\mathrm{A}) + \Pr(\mathrm{B}) - \Pr(\mathrm{A} \cap \mathrm{B})$ where $\mathrm{A} \cup \mathrm{B}$ and $\mathrm{A} \cap \mathrm{B}$ denotes the union and intersection

of the two events A and B, respectively. 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 for that bet? Surely not more than \$1 as then you would lose 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. One can show that your subjective probabilities must satisfy the usual axioms/rules for probabilities, otherwise you would be willing to enter a sequence of bets where you would lose an infinite amount with certainty; this is the **dutch book argument** for subjective probabilities. Objections have been raised against this argument, for example that the utility from the bet may not increase linearly with the monetary gain, and some people may even get utility just by the excitement in gambling; subsequent refinements of this argument have therefore completely disposed with the notion of money in favor of a more general notion of utility; see the Chapter Prediction and Decision making.

dutch book argument

1.4 *Bayesian Learning*

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

- Formulate your subjective *prior beliefs* $\Pr(A)$ 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 occurrence of some other event B as

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

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

Bayes' theorem

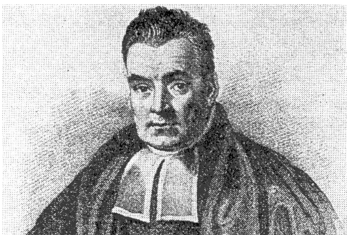


Figure 1.9: Reverend Thomas Bayes, ca 1701-1761, whose famous theorem was published after his death. Interestingly and somewhat ironically, we are not quite sure that the man in the photo actually is Thomas Bayes. *Probably* not.

Before moving on to Bayesian learning for model parameters, let us first use Bayes' theorem to solve a problem with meaningful real-life events. During the Covid pandemic it was common to take a quick home test to detect Covid, most commonly a cotton swab for nostrils and throat. Assume that you had just taken such a home test during the pandemic, with a positive result. Let us define the events of having Covid $A = \{\text{covid}\}$ and getting a positive test

$\mathrm{B} = \{\text{pos}\}$. The test that you are using contains a leaflet with the following information:

- The **sensitivity** of the test is 96.77%. This is the probability of a positive test given that one has Covid, hence $\Pr(\mathrm{B}|\mathrm{A}) = 0.9677$. sensitivity
- The **specificity** of the test is 99.20%. This is the probability of a negative test when one does not have Covid, hence $\Pr(\mathrm{B}^c|\mathrm{A}^c) = 0.9920$, where A^c is the complement to the event A. specificity

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 Covid given a positive home 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. Let us first assume we know nothing more than that the current prevalence of Covid in the population is around 5%, i.e. we use $\Pr(A) = 0.05$. Bayes' theorem reverses the conditioning and gives us the sought conditional probability:

$$\Pr(\mathrm{A}|\mathrm{B}) = \frac{\Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{B})} = \frac{\Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A}) + \Pr(\mathrm{B}|\mathrm{A}^c)\Pr(\mathrm{A}^c)} \approx 0.864,$$

where we have expressed $\Pr(\mathrm{B})$ in the numerator using a version of the **law of total probability** and the complement rule for e.g. $\Pr(\mathrm{B}|\mathrm{A}^c) = 1 - \Pr(\mathrm{B}^c|\mathrm{A}^c)$. Hence, even though the test has increased the probability of having the disease by a factor of $0.864/0.05 = 17.28$, the probability of actually having Covid is far from conclusive: there is a good chance $1 - 0.864 = 0.136$ of not having Covid after a positive test. A crucial assumption in this calculation is that your prior probability of having Covid before you took test is the prevalence of Covid in the population. This may be sensible if you were randomly selected to take the test *without any other symptoms*, but the reason why you took the test in the first place is probably that you had some symptoms of Covid (fever, coughing etc). Given such symptoms, you may assess your prior probability to be $\Pr(\mathrm{A}) = 0.7$ and the posterior probability after the positive test then rises to $\Pr(\mathrm{A}|\mathrm{B}) = 0.9965$. It is now almost certain that you have Covid. The lesson here is that prior probabilities matter. The Observable widget linked in the margin lets you experiment with different sensitivity, specificity and prior probability.

law of total probability

COVID BAYES' THEOREM

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 refers a belief about the event A *before* the data $x_1, \ldots, x_n$ was observed. In the same way, $\Pr(\mathrm{A}|x_1, \ldots, x_n)$ is referred to as the **posterior** since it is the probability of A *after* data was observed.

prior

posterior

The final step is to show how Bayes' theorem can be used to infer a parameter in a probability model $p(X_1, \ldots, X_n|\theta)$. One way to see the connection between a continuous parameter θ and the events A mentioned so far is by defining A to be the event that the model parameter θ belongs to an interval $\theta \in [a, b]$, for some constants $a < b$. We first take a simplified approach where the only possible parameter values are on a grid of values $\theta_1, \theta_2, \ldots, \theta_K$; for example $0.1, 0.2, \ldots, 0.9$ for the success probability $\theta \in [0, 1]$ in the iid Bernoulli model. 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 given by Bayes' theorem as

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

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

$$\Pr(\theta_k|x_1, \ldots, x_n) \propto \Pr(x_1, \ldots, x_n|\theta_k)\Pr(\theta_k), \tag{1.6}$$

where the symbol $\propto$ is read as 'is proportional to', i.e. the symbol $\propto$ indicates that 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 1.11 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).

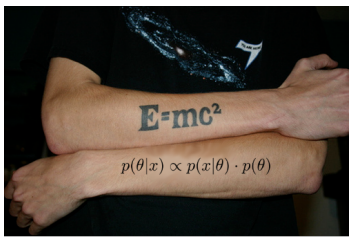


Figure 1.10: Great theorems make great tattoos.

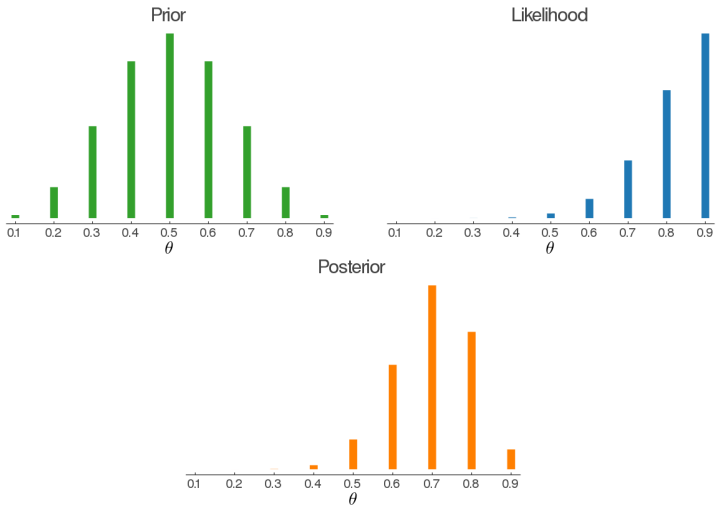


Figure 1.11: Prior, likelihood and posterior for Bernoulli model with $n = 10$ and $s = 9$.

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 1.12: Making a Bayesian omelette.

The ability to use prior information is a strength, especially when one has to make a decision based on 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, starting with regression and classification models in later chapters.

2.1 *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.

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

Beta distribution

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

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

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

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

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

Figure 2.1: The Beta distribution.

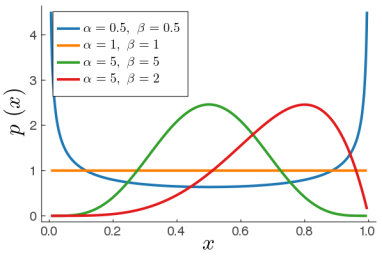


Figure 2.2: Some Beta distributions.

prior elicitation

DISTRIBUTION BETA

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.

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

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

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

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

Bayes estimator

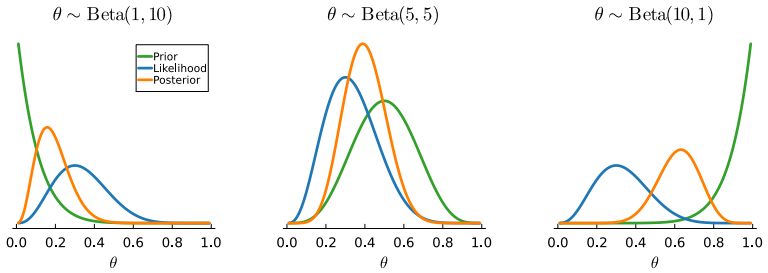


Figure 2.5: Bayesian analysis of $n = 10$ randomly chosen emails from the SpamBase data using three different priors. The likelihood is normalized.

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 $x_1,\ldots,x_n|\theta \overset{\text{iid}}{\sim} \text{Bern}(\theta)$ for the $n = 4601$ emails, where $x_i = 1$ if the email is spam and $x_i = 0$ for ham. The unknown quantity θ is the probability of spam.

SpamBase dataset

[1] Dua, D. and Graff, C. (2017). UCI machine learning repository

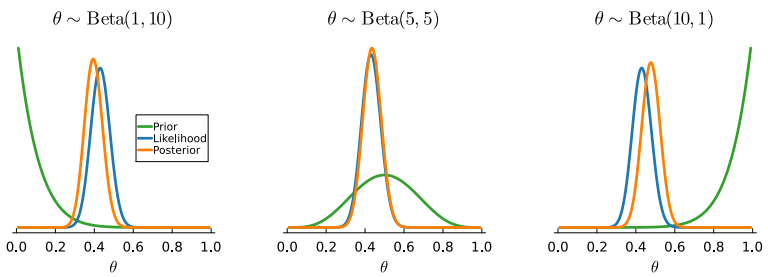


Figure 2.6: Bayesian analysis of $n = 100$ randomly chosen emails from the SpamBase data using three different priors. The likelihood is normalized.

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% of 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 achieve this in Chapter Classification when we 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.

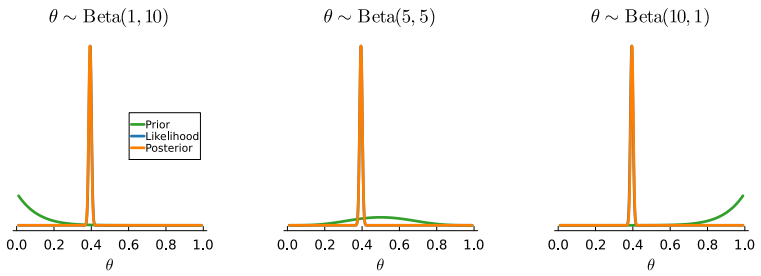


Figure 2.7: Bayesian analysis of all $n = 4601$ emails from the SpamBase data using three different priors. The likelihood is normalized.

2.2 *Bayesian learning and the likelihood principle*

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

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

- **Experiment 3**: sample data from $X_i|\theta \sim \text{Bern}(\theta)$ until exactly s, a predetermined number of successes, have been obtained.
  Stored data: the number of trials, n, until s successes have been obtained.

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

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

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

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

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

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

negative binomial distribution

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

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)

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

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

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

Likelihood principle

Informally, the likelihood principle says that all relevant information in an experiment about θ is contained in the likelihood function. The importance of the likelihood principle is that it can be mathematically derived from two simpler principles that everyone holds 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.

2.3 *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 practice 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.

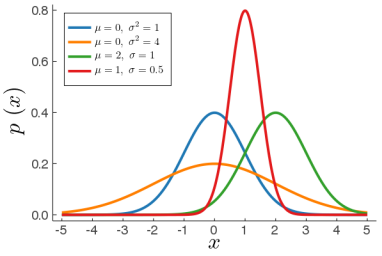


Figure 2.9: Some Normal distributions.

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 $\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i$ 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 interpretations 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: $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$, σ^2 known
Prior: $\theta \sim \text{N}(\mu_0, \tau_0^2)$
Posterior: $\theta | x_1, \ldots, x_n \sim \text{N}(\mu_n, \tau_n^2)$.
Posterior precision: $\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$
Posterior mean: $\mu_n = w\bar{x} + (1-w)\mu_0$, where $\bar{x} = n^{-1}\sum_{i=1}^n x_i$
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 days using a speed testing internet service; I will call this the **Internet speed dataset**. The measurements are assumed to be $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$, where θ is the average speed; we ignore for simplicity that the measurements cannot be negative. The measurements are reported to have a standard deviation of $\sigma = 5$ by the 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 $x_1 = 15.770$

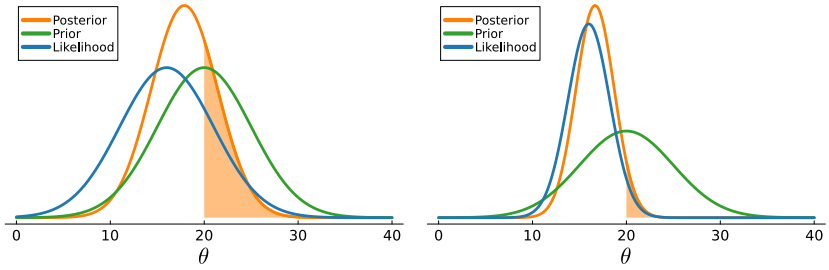


Figure 2.11: Internet speed data. Posterior updating based on $n = 1$ observation (left) and $n = 5$ observations (right). The orange shaded region marks out $\Pr(\theta > 20|x_1, \ldots, x_n)$.

Mbit/sec; the probability of interest $\Pr(\theta \geq 20|x_1, \ldots, x_n) \approx 0.275$ is marked out by the shaded orange region. Since the prior precision happened 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 (right) 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. An interactive Observable application for this example is linked from the button in the margin.

INTERNET SPEED PRIOR > POST

Online learning

Figure 2.11 illustrates a situation where the posterior is computed by combining the prior at day 0, $\mathrm{N}(\mu_0, \tau_0^2)$, 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** or **sequential learning**, where "yesterday's posterior becomes today's prior". Note that the online result in (2.9) is not specific for normal data with a normal prior, but is a general property of Bayesian updating. Figure 2.12 illustrates Bayesian online learning for the internet speed data.

online learning

sequential learning

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

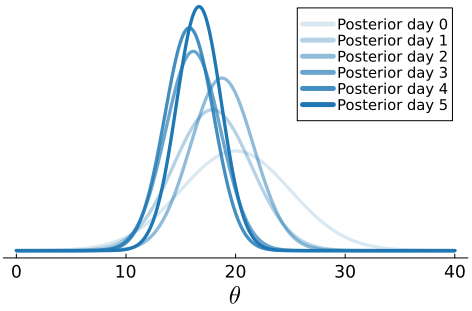


Figure 2.12: Illustration of Bayesian online learning for the internet speed data. The figure shows how the posterior changes when each new data point arrives. The posterior at Day 0 is just the original prior.

series, as is easily proved as follows

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

where the second line follows from the decomposition results in Figure 2.13. For iid data we have the additional simplication

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

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 $1 \leq m \leq n-1$

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

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

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

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

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

INTERNET SPEED SEQUENTIAL

batch learning

Decomposing distributions

For two random variables X, Y

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

For n random variables

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

and conditional on θ

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

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

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 the Chapters Gibbs sampling, Markov Chain Monte Carlo simulation and Variational inference.

2.4 *Poisson data*

Count data $X \in 0, 1, 2, \ldots$ is a quite frequently occurring 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 applications, but the model often fits many real datasets surprisingly well or can be extended to do so.

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

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

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:

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

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

Conjugate analysis - Poisson model

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

EXAMPLE: INTERNET AUCTION DATA. The **eBayCoin dataset** collected by Wegmann and Villani (2011) and made available in the

Poisson distribution

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

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

Figure 2.14: The Poisson distribution.

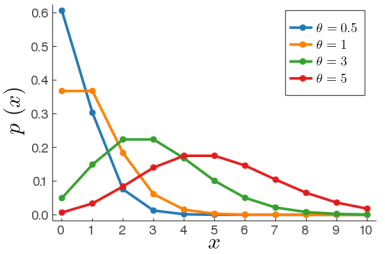


Figure 2.15: Some Poisson distributions.

DISTRIBUTION POISSON

Gamma distribution

Gamma distribution

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

Figure 2.17: Prior-to-Posterior updating for the Poisson data with a Gamma prior.

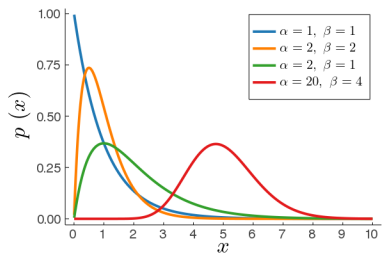


Figure 2.18: Some Gamma distributions.

eBayCoin dataset

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 bidders 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.

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

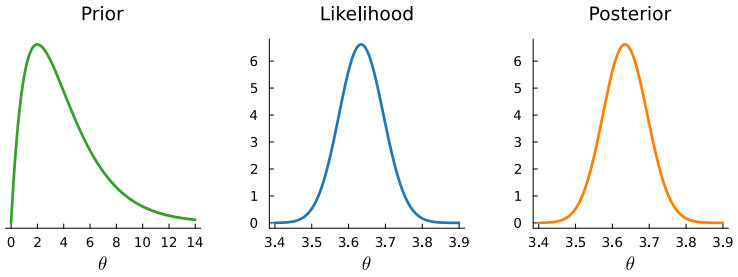


Figure 2.19: Bayesian analysis of the numbers of bidders in $n = 1000$ eBay coin auctions.

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.

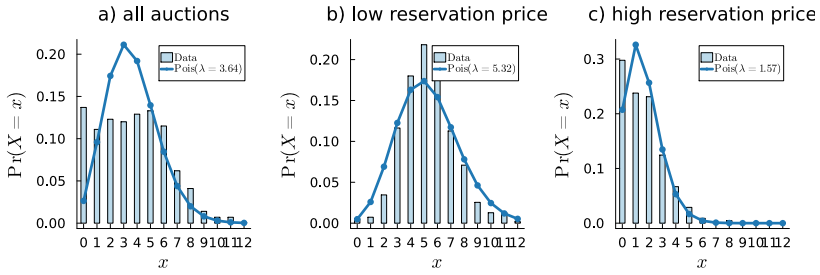


Figure 2.20: Assessing the fit of the Poisson model with the posterior mean estimate of θ.

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 to 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 Gamma$(4, 1/2)$ to reflect a belief that such auctions are likely to attract more bids. The prior for the auction where the reservation prices are high is set to 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 a covariate as well as other auction specific covariates.

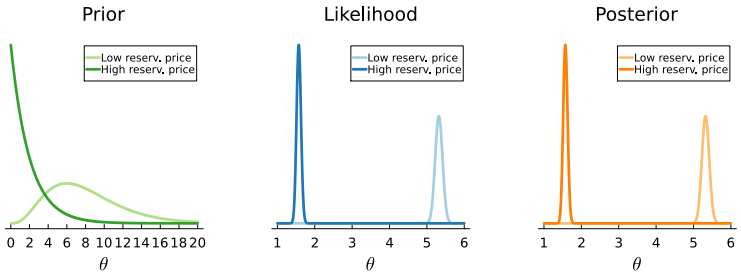


Figure 2.21: eBay auctions. Bayesian analysis of the numbers of bidders in $n = 550$ auctions with a low reservation price and $n = 450$ auctions with a high reservation price.

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*

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

2.5 *Summarizing a posterior distribution*

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

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

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

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

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

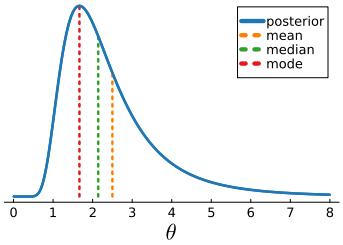


Figure 2.22: Three common point estimates for summarizing a posterior.

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 a 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 suggests,

equal tail credibility interval

is made up of the θ values with the highest posterior density. We use the word *region* instead of interval here since HPD regions need not be intervals. Here is the definition.

Highest Posterior Density (HPD) region

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:*

- $\Pr(\theta \in R | x_1, \ldots, x_n) = \gamma$ *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 the highest posterior density (middle graph) and how HPD regions can be disconnected (right hand graph).

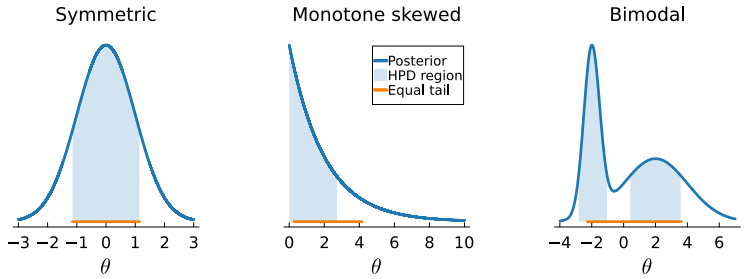


Figure 2.23: Illustration of HPD regions (shaded areas) and equal tail intervals (orange line).

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.

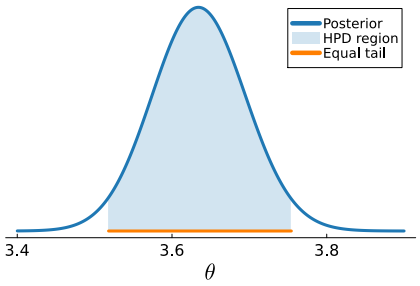


Figure 2.24: 95% credibility intervals for the Gamma posterior in the eBay auction data.

2.6 *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*

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

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 $\mathbf{t}(x_1, \ldots, x_n)$. □

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:

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

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

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 $\text{ExpFam}(\theta)$ as a generic notation for a distribution in the exponential family, leaving the specific $h(x)$, $A(\theta)$, $\eta(\theta)$ and $t(x)$ functions implicit.

The likelihood function for iid data from an $\text{ExpFam}(\theta)$ distribution is

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

Lemma 1 can be directly 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 distributions in 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)

$$p(\theta) = H(\tau_0, \nu_0) \exp\left(\eta(\theta)\tau_0 - \nu_0 A(\theta)\right), \quad (2.16)$$

where $H(\tau_0, \nu_0)$ is the normalizing constant. Note that this prior has two hyperparameter τ_0 and ν_0 that need 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 $\eta(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$ and $A(\theta) = \ln\left(\frac{1}{1-\theta}\right)$, for Bernoulli data. The prior in (2.16) is therefore

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

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 an imaginary prior sample of τ_0 success in ν_0 trials. The $\text{Beta}(\alpha, \beta)$

DISTRIBUTION STUDENT-T

Student-t distribution

$X \sim \text{t}(\mu, \sigma, \nu)$ 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 $\nu > 1$

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

Figure 2.25: The student-t distributions.

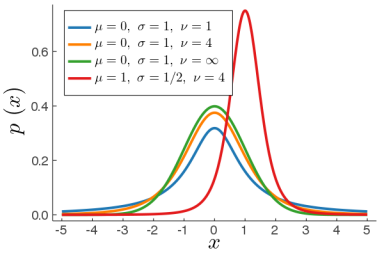


Figure 2.26: Some Student-t distributions.

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: $\theta | x_1, \ldots, x_n \sim \text{ExpFamConj}\left(\tau_0 + \sum_{i=1}^n t(x_i), \nu_0 + n\right)$

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 multiplying the likelihood in (2.15) with prior (2.16)

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

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 person's normal prior, i.e. compute θ_0 and τ_0 for this person.
3. (a) Let $x_1, \ldots, x_{10}$ 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.

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.

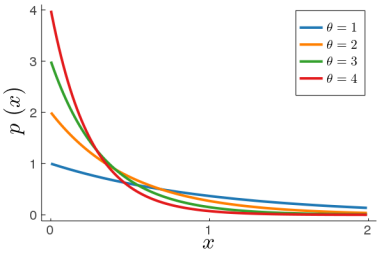


Figure 2.29: Some Exponential distributions.

(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 a computer available you may plot this unnormalized posterior over a grid of θ values. *Hint: the posterior distribution is not normal anymore when the measurements are truncated at* 3.

4. Derive the posterior distribution for the normal model with a normal prior in Figure 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 the number of eBay bidders with a Poisson model.

3 *Multi-parameter models*

3.1 *Joint posterior distributions*

Most models have more than one parameter, and many models are incredibly rich in parameters. Datasets are increasing rapidly in size which 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 distributions are now both multidimensional joint distributions. Figure 3.1 gives an illustration of a bivariate ($K = 2$) normal distribution.

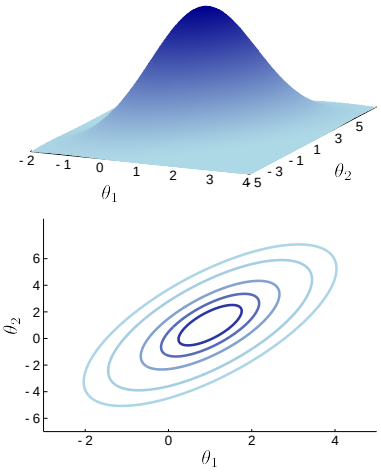


Figure 3.1: Surface and contour plot of the bivariate normal distribution. The contour levels contain 25, 50, 75, 95 and 99% of the probability mass, respectively.

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

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

3.2 *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

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

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

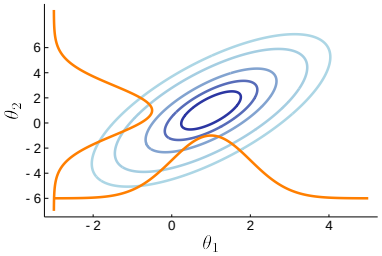


Figure 3.2: Contour plot of the bivariate normal distribution in Figure 3.1 along with the marginal distributions.

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 then 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.

3.3 *Gaussian data with unknown variance*

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

turns out 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 $\text{Inv}{-}\chi^2(\nu, \tau^2)$ in general; see Figures 3.3 and 3.4. This distribution is a specific parametrization of the **inverse Gamma distribution**. The name comes from the characterization

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

so that a $\text{Inv}{-}\chi^2(\nu, \tau^2)$ variable is an inverted χ^2_ν variable scaled by $\nu\tau^2$. Note from Figure 3.3 that the parameter τ^2 is close to the mean when ν is large. The mode is $\nu\tau^2/(\nu+2)$, so τ^2 is somewhere between the mode and the mean. We will therefore call τ^2 the location of $\text{Inv}{-}\chi^2(\nu, \tau^2)$, or sometimes just sloppily as "our best guess".

The conjugate prior in (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 normal distribution with variance σ^2. The hyperparameter ν_0 plays the same role for σ^2.

Figure 3.5 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 3.5, 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)

DISTRIBUTION SCALED INVERSE CHI2

inverse Gamma distribution

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: The $\text{Inv}{-}\chi^2$ distribution.

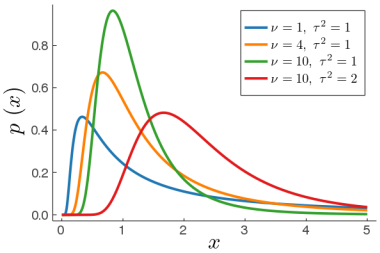


Figure 3.4: Some $\text{Inv}{-}\chi^2$ distributions.

Gaussian iid data with conjugate prior

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

Figure 3.5: Prior-to-Posterior updating for the iid Gaussian model with unknown mean and variance using the conjugate prior.

- the data variance σ^2 does not appear in w, as it did when the variance was known. The reason for this difference 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 3.5. In Exercise 1 you are asked 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, the marginal posterior of θ is

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

where μ_n, σ_n^2, κ_n and ν_n are all defined as in Figure 3.5. 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 3.7, which I find sensible.

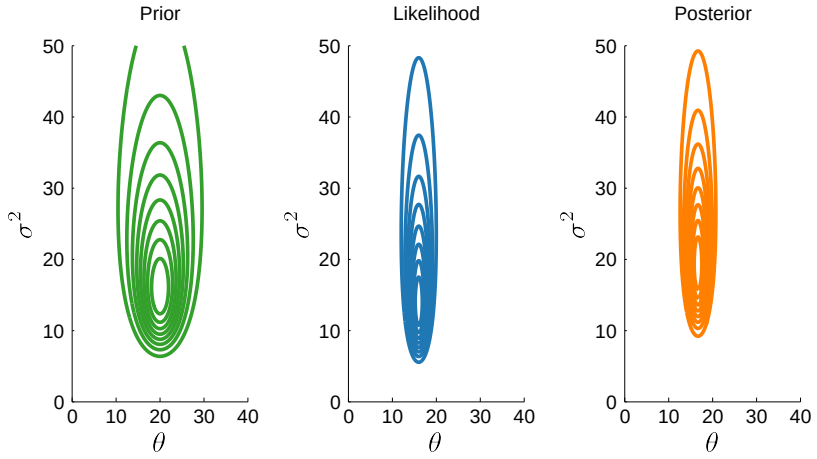


Figure 3.6: Prior-to-Posterior updating for the internet speed data in the iid Normal model. Contours of joint distributions of θ and σ^2.

Figure 3.6 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 3.7. 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.

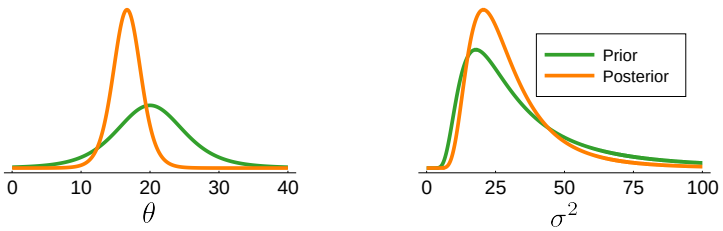


Figure 3.7: Marginal posteriors for the internet speed data in the iid Normal model.

3.4 *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 **??** 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 $\mathbf{x} = (x_1,\ldots,x_n)$
number of posterior draws m.
compute μ_n, σ_n^2, κ_n and ν_n using Figure 3.5.
for i in $1{:}m$ **do**
  $\sigma^2 \leftarrow$ rInvChi2(ν_n,σ_n^2)
  $\theta \leftarrow$ rNormal$(\mu_n,\sigma^2/\kappa_n)$
end
Output: m draws for θ and σ^2 from joint posterior.

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

Figure 3.8: 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 3.8 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 3.8.

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.

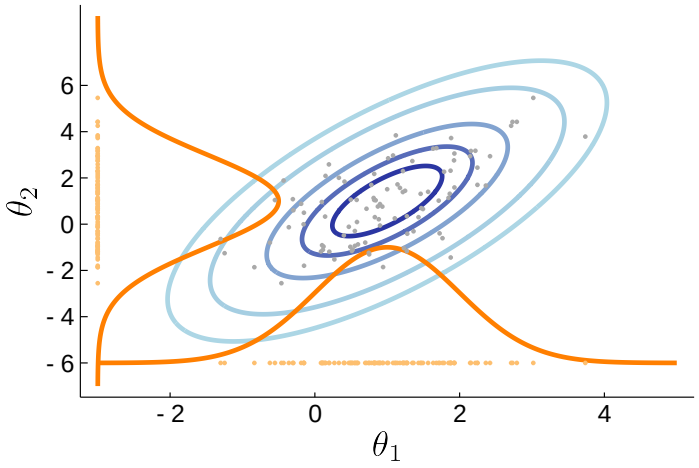


Figure 3.9: Illustrating marginalization by selection. The figure plots the contours of a joint distribution with the marginal distributions overlayed as orange curves. The gray points are 100 draws from joint distribution and the orange points are projections of the gray points on the two axes. The orange points correspond to the draws obtained by selecting out each parameter from the joint simulation and therefore 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 3.9.

Figure 3.12 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 3.12 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, the sample mean of the draws will converge to the true posterior expectation $\mathbb{E}(\theta|\mathbf{x})$ in large simulations. Formally, if we let $\theta^{(i)}$ denote the ith posterior draw of any of the parameters in a model, this result can be expressed as

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

where $\xrightarrow{p}$ denotes **convergence in probability**, see Figure 3.10. This result is a version of the **law of large numbers**, see Figure 3.11. The

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 3.10: 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 3.11: Weak law of large numbers.

simulation consistent

left side of Figure 3.13 illustrates this convergence by plotting the posterior mean estimates $\bar{\theta}_{1:m}$ for increasing m; note that the figure shows these cumulative estimates only up to $m = 1000$.

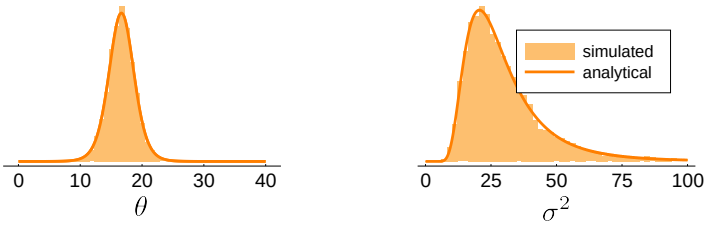


Figure 3.12: Histogram of simulated marginal posteriors for the internet speed data with analytical marginal posterior densities overlayed.

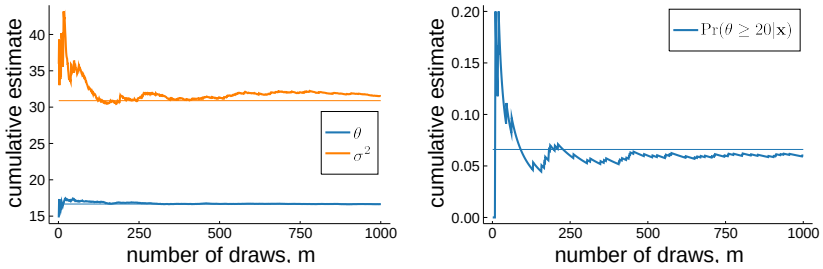


Figure 3.13: Convergence of the Monte Carlo estimate of the posterior expectation of θ and σ^2 (left) and $\Pr(\theta \geq 20|\mathbf{x})$ (right). The analytical posterior results are displayed as thin horizontal lines.

The **central limit theorem** (Figure 3.15) can be used to prove that $\bar{\theta}_{1:m}$ **converges in distribution** (Figure 3.14) to a normal distribution. Hence, the following approximation of the posterior estimate $\bar{\theta}_{1:m}$ is accurate when m is large:

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

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 $\boldsymbol{\theta}$.

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

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 functions (cdf) of X_n and X, respectively.

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

Figure 3.14: Convergence in distribution.

coefficient of variation

tedious multidimensional change-of-variables calculations. Here is a second attractive property of simulation: the posterior for $f(\boldsymbol{\theta})$ can be directly obtained from a posterior sample of $\boldsymbol{\theta}$ by simply computing the function $f(\boldsymbol{\theta})$ for each posterior draw. Provided the posterior variance of $f(\boldsymbol{\theta})$ exists, a central limit theorem of the form (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 3.16; 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

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

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

The CLT is often informally written as

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

Figure 3.15: The central limit theorem.

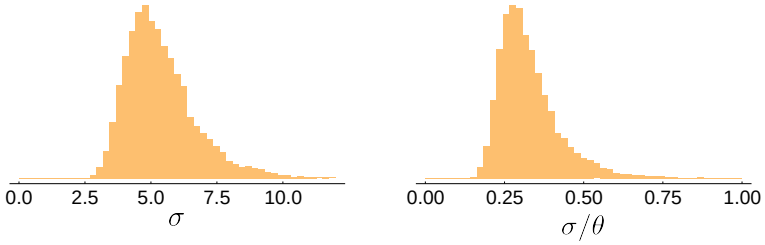


Figure 3.16: Histogram of simulated marginal posteriors for σ (left) and the coefficient of variation σ/θ (right) for the internet speed data.

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 3.13 illustrates the Monte Carlo convergence of this estimate.

3.5 *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 item sold in an auction may reported as: 'defective', 'normal quality', or 'new'; a continuous variable like age can be 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 categories 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 $x_j = 0$ 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

one-hot encoding

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

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 3.17 is an example of $\text{Cat}(\theta_1, \ldots, \theta_C)$ for $C = 4$.

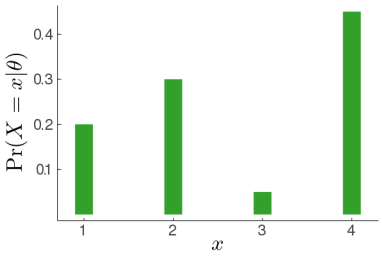


Figure 3.17: Categorical distribution with probabilities $\boldsymbol{\theta} = (0.20, 0.30, 0.05, 0.45)$.

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 3.18. 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

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

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

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 3.18: The multinomial distribution.

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 3.19.

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 3.19: The Dirichlet distribution.

unit simplex

Dirichlet distribution

The **Dirichlet distribution** is specified with the prior hyperparameters $\alpha_c > 0$, see Figure 3.20 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 3.20, 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 3.19; hence, the prior hyperparameters $\boldsymbol{\alpha} = (1.5, \ldots, 1.5)$ and $\boldsymbol{\alpha} = (5, \ldots, 5)$ in the upper part of Figure 3.20 have the same mean, but the latter has smaller variance. Finally, the bottom part of Figure 3.20 shows examples where the prior mean is different over the categories.

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

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 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

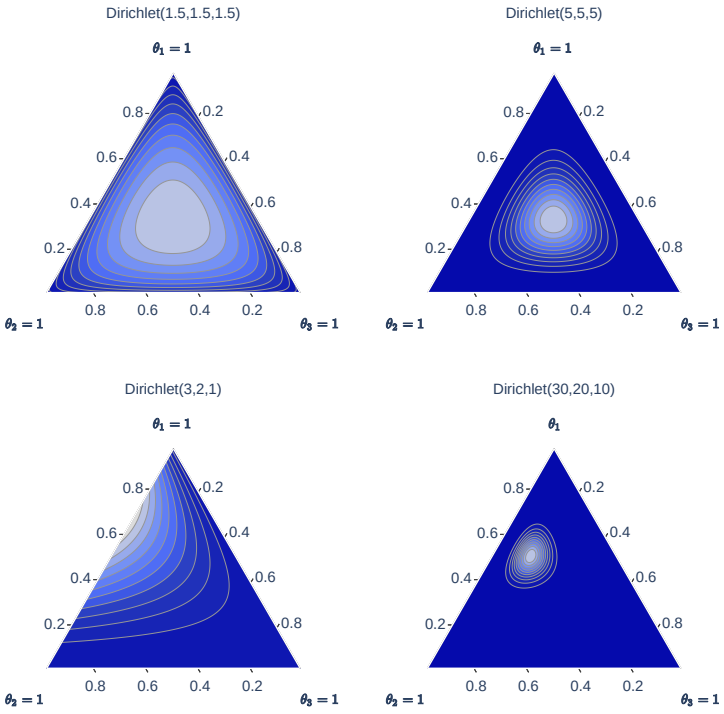


Figure 3.20: Probability density functions for some Dirichlet distributions for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. Lighter color means higher density. The corners of the simplex represent $\theta_1 = 1$, $\theta_2 = 1$ and $\theta_3 = 1$ respectively, as denoted in the figures.

category. This parallels and generalizes the binary case where a Beta(α, β) prior was updated to a posterior by adding the number of successes s to α and the number of failures f to β. Figure 3.21 summarizes the prior-to-posterior updating for multinomial data with a Dirichlet prior.

Multinomial data with Dirichlet prior

Model: $\mathbf{y}|\boldsymbol{\theta} \sim \text{Multinomial}(\boldsymbol{\theta})$, 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: $\boldsymbol{\theta}|\mathbf{y} \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{y})$

Figure 3.21: 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 years 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, $\sum_{c=1}^{4} \alpha_c = 50$, so the prior information is equivalent to a survey based on 50 respondents, as required.

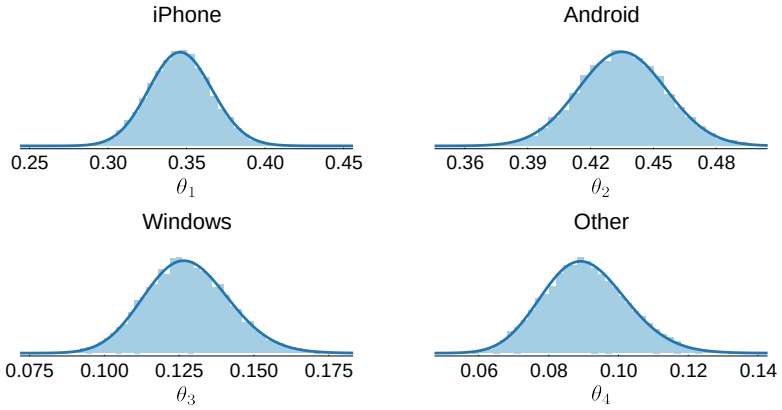


Figure 3.22: Marginal posteriors of the market shares for the mobile phone survey data. Simulated (histogram) draws and analytical density functions (solid curves).

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 $\theta_2 > \max(\theta_1, \theta_3, \theta_4)$.

The joint posterior distribution of all four shares is by Figure 3.21 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 3.22 as histograms from Monte Carlo simulation (see the algorithm in Figure 3.23); the analytical posteriors from Figure 3.19 are overlayed.

Figure 3.22 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

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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$.

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 3.23: 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.

$\theta_2 > \max(\theta_1, \theta_3, \theta_4)$ 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.

3.6 *Multivariate normal data with known covariance*

This section considers the iid **multivariate normal distribution** model for a p-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 3.24 for the density and properties, and Figure 3.25 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 $\mathbf{x}_i$

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

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 3.24: The multivariate normal distribution.

DISTRIBUTION MULTIVARIATE NORMAL

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):

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

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

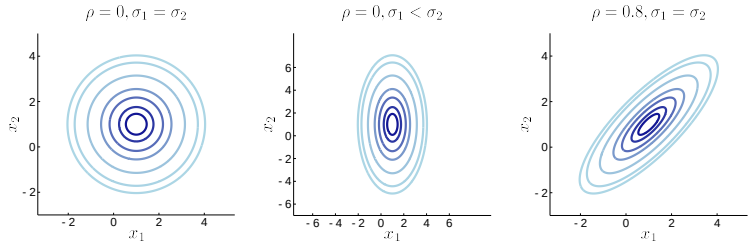


Figure 3.25: Contour plots of some bivariate normal distributions with correlation ρ.

Not too surprisingly, the multivariate normal prior

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

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 3.26 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:

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

where

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

Figure 3.26: Completing quadratic forms.

3.7 *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 suggests that the information in a dataset can be measured by how peaked the likelihood is around its mode. This section

formalizes this idea using the mathematical concept of Taylor approximations. If you are not familiar with the Taylor approximation of a function, pause your reading and go to the section Taylor approximation in the mathematical Appendix.

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

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

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

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

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)**.** *The **observed information** in a sample $\mathbf{x} = (x_1, \dots, x_n)$ 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 (Fisher 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 in the 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

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

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 in the 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 $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$.
2. Let $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$, where θ is assumed known. Show that the Inv$-\chi^2$ distribution is a conjugate prior for σ^2.
3. The monthly income (in thousands Swedish Krona) of ten randomly selected persons are: 14, 25, 45, 25, 30, 33, 19, 50, 34 and 67. The log-normal distribution (see Figures 3.27 and 3.28) is a commonly used model for income distributions. Let $y_1, \ldots, y_n|\mu, \sigma^2 \overset{\text{iid}}{\sim} \text{LN}(\mu, \sigma^2)$, where $\mu = \log(33)$ is assumed to be known but σ^2 is unknown with non-informative prior $p(\sigma^2) \propto 1/\sigma^2$.
   a) Show that posterior for σ^2 given μ is the Inv$-\chi^2(n, \tau^2)$ distribution, where
   $$\tau^2 = \frac{\sum_{i=1}^n (\log y_i - \mu)^2}{n}.$$

Log-Normal distribution

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

Figure 3.27: The log-normal distribution.

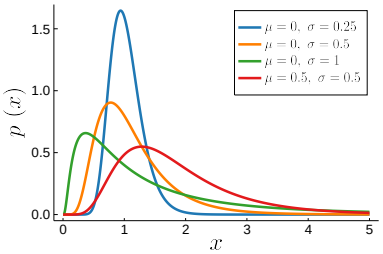


Figure 3.28: Some log-normal distributions.

b) Simulate $10,000$ draws from the posterior of σ^2 (assuming $\mu = \log(33)$) and compare graphically with the theoretical Inv$-\chi^2(n, \tau^2)$ posterior distribution.

c) A commonly used measure of income inequality is the Gini coefficient, $0 \leq G \leq 1$, where $G = 0$ is complete income equality, and $G = 1$ means complete income inequality. It can be shown that $G = 2\Phi\left(\sigma/\sqrt{2}\right) - 1$ when incomes follow a LN(μ, σ^2) distribution, where $\Phi(z)$ is the cumulative distribution function (CDF) for the standard normal distribution. Use the posterior draws in a) to compute the posterior distribution of the Gini coefficient G.

d) Use the posterior draws from c) to compute a 95% equal tail credible interval and a 95% Highest Posterior Density (HPD) interval for G. Compare the two intervals. To compute the HPD interval you will need an estimate of the posterior density for G; a common approach is to use a kernel density estimator.

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 subsequently 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.

4.1 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 occurring 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 displays 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 time series 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.

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

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

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 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 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 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 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 we 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

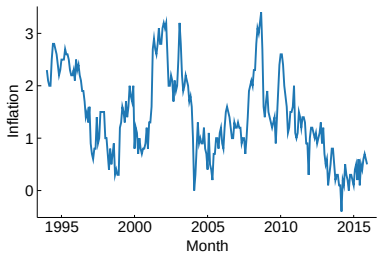


Figure 4.1: Swedish inflation 1995-2016 - annualized monthly observations.

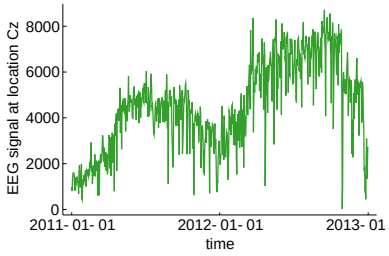


Figure 4.2: Daily number of rides with a bike sharing company.

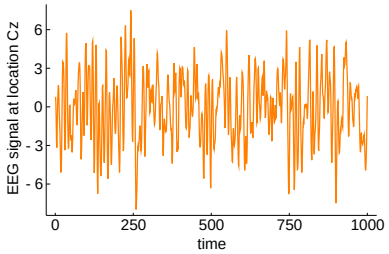


Figure 4.3: EEG recordings of electrical activity at one brain scalp location.

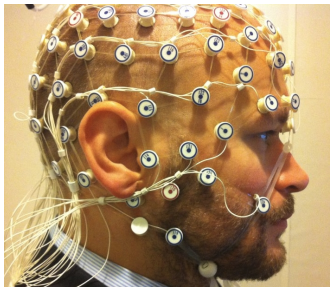


Figure 4.4: Positioning of EEG electrodes on a subject's brain scalp. Eeg electrodes are typically placed at 10-20 locations. A time series is recorded at each location.

autoregressive model

lagged value

steady-state form

stationary

often used in applications. We will also ignore the restrictions on $\phi_1, \ldots, \phi_p$ needed to guarantee stationarity when designing the prior. Such restrictions can imposed by simply truncating the parameter space of $\phi_1, \ldots, \phi_p$ to the stationary region. We will walk through a number of methods for prior elicitation and use different methods for different parameters.

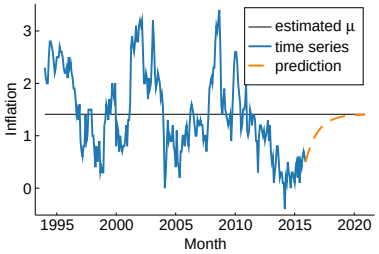


Figure 4.5: Swedish inflation 1995-2016 with 60 months ahead mean prediction in dashed orange.

4.2 *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 the 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 in section Multinomial 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.

4.3 *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 elicitation 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 a large area in itself, with help from cognitive science to account for the biases and shortcomings that are unfortunately part of being a human.

4.4 *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 a first understanding of the concept from a commonly used prior for the autoregressive parameters $\phi_1, \dots, \phi_p$ in the AR process. A regularization prior on $\phi_1, \dots, \phi_p$ 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

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

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 reason for scaling the prior variance of all ϕ_k by the error variance σ^2 is that the prior when become conjugate conditional on μ (compare with (3.3) in Section 3.3), which will turn out to be useful when we devise an algorithm for posterior simulation in Chapter Gibbs sampling.

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

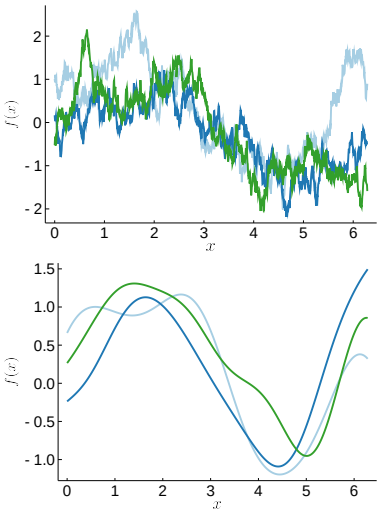


Figure 4.6: Three simulated draws from a prior over functions without smoothness beliefs (top) and with smooth beliefs (bottom).

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

smoothness beliefs

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 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.

4.5 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

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

where each $p(\phi_k|\tau^2, \sigma^2)$ is the previous $\mathrm{N}\left(\mu_k, \sigma^2 \frac{\tau^2}{k^2}\right)$, with independence now only conditionally on τ^2, and $p(\tau^2|\sigma^2)$ 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.

4.6 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 $\theta \sim \text{Beta}(\alpha,\beta)$ prior is $\theta|\mathbf{x} \sim \text{Beta}(\alpha+s,\beta+f)$. 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 variables 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.

4.7 *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

invariant prior

of the parameter.

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

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

so that the Fisher information is (using lowercase letters 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 $x_1, \ldots, x_n | \theta \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$. Let us also assume that θ is known and we use Jeffreys' rule to obtain the invariant prior for σ^2. The log-likelihood is

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

with first and second derivative

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

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

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

so Jeffreys' prior for the variance is

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

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

improper prior

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

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

reference prior

EXERCISES

1. Show that using Jeffreys' rule to obtain a prior for the log odds $\phi \equiv \log \theta/(1-\theta)$ in Bernoulli trials implies the same Beta$(1/2, 1/2)$ prior for θ (i.e. that Route A and B in the text give the same prior).
2. Derive Jeffreys' prior for the success probability θ in the negative binomial model for a dataset where n trials were needed to obtain a predetermined s number of successes. Compare with the Jeffreys prior derived for the Bernoulli model in the text. Discuss the implication for the likelihood principle.
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 a function of a set of input variables $x_1, \ldots, x_p$. The variable y can for example be the salary of a person that we are trying to explain using information on that person's `age` (recorded by the variable x_1) and `income` (recorded by x_2). The input variables are often called **covariates**, predictors or **features**, and the output variable is most commonly termed the **response variable** or target variable.

covariates

features

response variable

In the Chapter Classification we will see regression models for a binary response variable, for example a variable $y \in 0, 1$ that records if a person is employed ($y = 0$) or unemployed ($y = 1$). We will also encounter regression models for response variables of other data types, for example counts, where y may record the number of tickets sold to an event or the number of faulty products produced on any given day by a manufacturing machine. 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 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 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is the vector of covariates for the ith observation in the dataset, $\top$ denotes vector transpose 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 a 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. The reader who is not very familiar with vectors and matrices is encouraged to read the Linear Algebra section of Appendix Referencesch:appendixmath and check that the **matrix-vector product** $\mathbf{X}\boldsymbol{\beta}$ is a vector of length n with the ith element being $\mathbf{x}_i^\top \boldsymbol{\beta}$.

matrix-vector product

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 $\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$, the frequentist sampling distribution of $\hat{\boldsymbol{\beta}}$ is obtained by applying the result in Figure 5.1 with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, $\Sigma = \sigma^2 I_n$ 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 MLE can then be used to compute **fitted values** $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ or for making predictions for new covariate observations, see Chapter Prediction and Decision making.

fitted values

The result in (5.4) shows that the MLE is unbiased for $\boldsymbol{\beta}$, that is, $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$; using the MLE guarantees that your estimate of $\boldsymbol{\beta}$ is correct on average over all possible datasets of size n from the data generating process in (5.2).

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 however biased for σ^2, and the following unbiased estimator is typically used instead

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

Let us now consider what happens to the MLE $\hat{\boldsymbol{\beta}}$ when we rescale the covariates, i.e. when we multiply each covariate x_j with a constant c_j. Such transformations are very common, for example changing the measurement unit of x_j from meters to centimeters, which corresponds to multiplying all observations in x_j by $c_j = 100$. Now, the regression coefficient β_j measures how much the expected value of the response variable changes if you change x_j by one unit. We would therefore like that the rescaling $x_j \to c_j \cdot x_j$ of our covariate brings about a corresponding inverse scaling of the estimate: $\hat{\beta}_j \to (1/c_j) \cdot \hat{\beta}_j$. This desirable **equivariance** property holds for the MLE $\hat{\boldsymbol{\beta}}$, which we will now show.

equivariance

We will prove a more general equivariance result for the MLE where we linearly transform the whole vector of p covariates by a $p \times p$ invertible transformation matrix $\mathbf{A}$. The scaling of individual covariates is then the special case where $\mathbf{A} = \text{Diag}(1, \ldots, c_j, \ldots, 1)$ is a diagonal matrix with ones on the diagonal except in the jth position. So, let us consider what happens to the MLE under the general invertible transformation $\mathbf{x} \to \mathbf{A}\mathbf{x}$. The matrix of transformed covariates can then be written $\mathbf{X}_A = \mathbf{X}\mathbf{A}^\top$, since the ith row of $\mathbf{X}$ is the transpose of the column vector $\mathbf{x}_i$. The MLE for the coefficients $\boldsymbol{\beta}_A$ in the transformed covariate model is then easily obtained from the formula of the MLE and a little linear algebra:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_A &= (\mathbf{X}_A^\top\mathbf{X}_A)^{-1}\mathbf{X}_A^\top\mathbf{y} = (\mathbf{A}\mathbf{X}^\top\mathbf{X}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{X}^\top\mathbf{y} \\ &= (\mathbf{A}^\top)^{-1}(\mathbf{X}^\top\mathbf{X})\mathbf{A}^{-1}\mathbf{A}\mathbf{X}^\top\mathbf{y} = (\mathbf{A}^\top)^{-1}\hat{\boldsymbol{\beta}}.\end{aligned}$$

Hence, the MLE with transformed covariates is the inversely transformed MLE in with the original model: $\hat{\boldsymbol{\beta}}_A = (\mathbf{A}^\top)^{-1}\hat{\boldsymbol{\beta}}$. To further convince ourselves that this is sensible we can compute the fitted values in the model with transformed covariates

$$\hat{\mathbf{y}}_A = \mathbf{X}_A\hat{\boldsymbol{\beta}}_A = \mathbf{X}\mathbf{A}^\top(\mathbf{A}^\top)^{-1}\hat{\boldsymbol{\beta}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}},$$

which shows that the fit is not affected by transforming the covariates when maximum likelihood is used to estimate $\boldsymbol{\beta}$.

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 Section Invariant priors.

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

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

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

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

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

$$
p(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-[1+(n-p)/2]} \exp\left\{-\frac{1}{2\sigma^2}(n-p)s^2\right\}, \tag{5.8}
$$

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

Gaussian linear regression with non-informative prior

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

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

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

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. Note that since we have used a noninformative prior the posterior mean of $\boldsymbol{\beta}$ is exactly the MLE and the posterior covariance matrix of $\boldsymbol{\beta}$ is the same as the sampling covariance matrix of the MLE. The *interpretation* of the Bayesian results Figure 5.2 are very different though; the Bayesian posterior is still a distribution for the unknown parameters conditional on the observed dataset. We can make great use of this distribution in prediction and decision making, as we will see in the next chapter.

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 $x_1, \ldots, x_n | \theta, \sigma^2 \sim \mathrm{N}(\theta, \sigma^2)$ was of the form $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$ where

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

The conjugate prior in linear regression is very similar

$$\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\mu}_0, \sigma^2\Omega_0^{-1}) \tag{5.9}$$
$$\sigma^2 \sim \text{Inv}{-}\chi^2(\nu_0, \sigma_0^2), \tag{5.10}$$

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

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:
$\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\mu}_0, \sigma^2\Omega_0^{-1})$
$\sigma^2 \sim \text{Inv}{-}\chi^2(\nu_0, \sigma_0^2)$

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

where

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

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

A detailed elicitation of the matrix Ω_0 can be demanding. We will have more to say about Ω_0 in the Chapter Regularization, where the simple choice $\Omega_0 = \kappa_0 \boldsymbol{I}_p$ will be discussed in more detail. This prior assumes that the regression coefficients are a priori independent since $\sigma^2\Omega_0^{-1}$ is diagonal. Prior independence does often not reflect true prior beliefs, but is convenient since we do not have to specify all prior correlations between parameters. Note also that this prior will be updated with data with the effect that the parameters are dependent in the posterior distribution. One further thing to note is that this prior assumes the *same* prior precision κ_0 for all β coefficients.

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.

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

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu} \;\text{ if } \nu > 1$$
$$\mathbb{V}(\mathbf{x}) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma} \;\text{ if } \nu > 2$$

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

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

Figure 5.4: The multivariate student-*t* distribution.

This is only sensible if the covariates are roughly on the same scale, i.e. have similar mean and standard deviations. Similarly, in contrast to the MLE, this prior gives a posterior which is not equivariant with respect to scaling of the covariates. The problem is that we are insisting on using the *same* prior for the regression coefficient after the transformation, thereby ignoring that the interpretation of $\boldsymbol{\beta}$ is directly dependent on the scaling of covariates; hence, the prior really should change after rescaling the covariates. This can be achieved by inversely transforming the prior, but a simpler solution is to standardarize the covariates before the analysis, for example to have mean zero and unit variance.

A commonly used prior is Zellner's prior where $\Omega_0 = \frac{\kappa_0}{n}(\mathbf{X}^\top\mathbf{X})$. Note here that we are using covariate data to formulate a prior, which seems to go against the requirement a prior should not depend on data. However, covariates are typically assumed to be known in regression analysis and can then actually be used when formulating a prior; Zellner's prior does not depend on response data, and the prior therefore contains the information about $\boldsymbol{\beta}$ *before* observering $\mathbf{y}$. One way to understand the particular form of Zellner's prior is that its prior covariance matrix is $\frac{n}{\kappa_0}\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, which is 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 and can be shown to give a posterior which is equivariant. 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 $\kappa_0 = 1$, Zellner's prior therefore becomes a noninformative **unit information prior** with information content equal to the expected information from just a single observation.

unit information prior

Figure 5.3 shows that the prior in (5.9) is indeed a conjugate prior. Figure 5.3 also gives the marginal posterior of $\boldsymbol{\beta}$ as a **multivariate student-t distribution**, see Figure 5.4. The proofs of these results are given at the end of this chapter.

UNIVERSITY SALARIES DATA. The **salaries dataset**, described in the book Fox and Weisberg (2019) and made available as the data frame `Salaries` in the `R` package `carData`, contains salaries for $n = 397$ university professors. The professors have three different ranks (Assistant, Associate and Full professor) and work 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.

salaries dataset

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 the highest academic age in the dataset. The relationship 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

Table 5.1: Summary of the university salaries data.

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

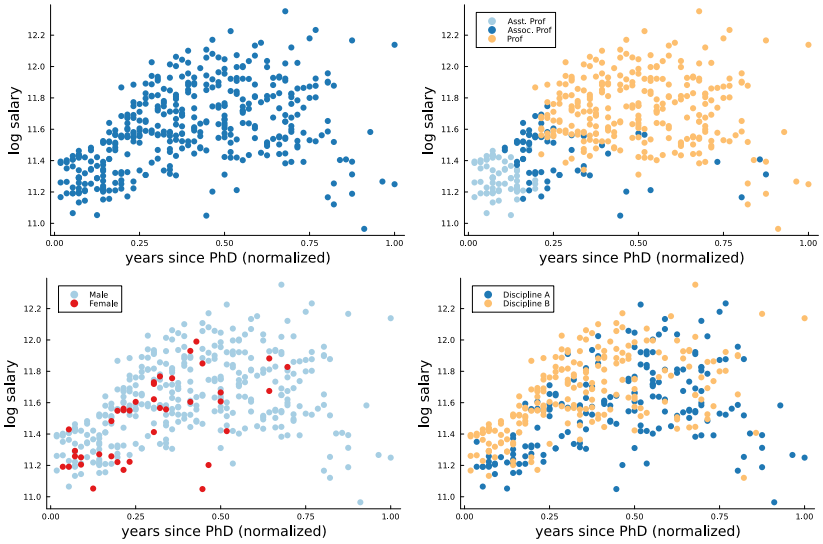


Figure 5.5: University salaries data. Scatterplot of `logsalary` against `phdage` (topleft), colorcoded by `rank` (top right), `sex` (bottom left) and `discipline` (bottom right). See Table 5.1 for variable definitions.

Datasets typically contain 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

dummy variables

one-hot encoding

associate professors, and C for full professors. This variable is coded 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 $y =$ `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 for 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 5.6. 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 3.27. 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 5.7, 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 5.7 agrees with my 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

I will use Zellner's prior $\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}(\mu_0, \sigma^2\Omega_0^{-1})$ with $\Omega_0 = \frac{\kappa_0}{n}(\mathbf{X}^\top\mathbf{X})$, 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

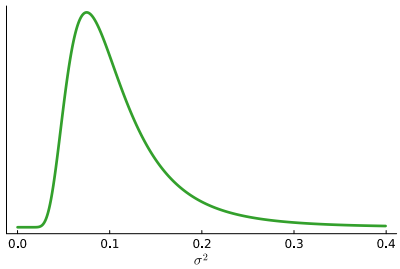


Figure 5.6: Prior for σ^2 in university salaries data.

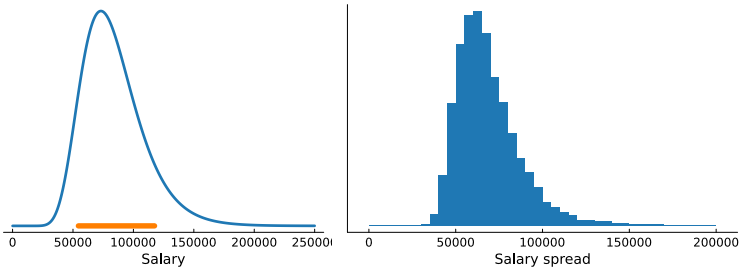


Figure 5.7: Prior elicitation for σ^2 in university salaries data. Left: Implied log-normal distribution of salaries from assuming a median salary of 80,000 and $\sigma_0^2 = 0.3^2$; the orange line marks out the wage spread as measured by the difference between the 90% and 10% salary percentiles. Right: 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 a 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 summary 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 5.8 explores the prior by simulation for four different κ_0: 5, 50, 397 and 1000 by plotting the implied median salary curve over `phdage` for each of ten $\boldsymbol{\beta}$ simulated from the prior. Note that the sample size $n = 397$, so $\kappa_0 = 397$ gives the same weight

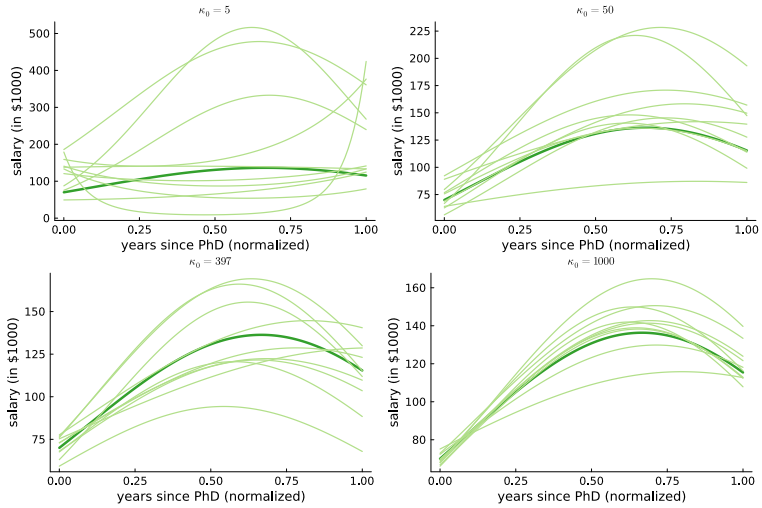


Figure 5.8: Implied relationship between `salary` and `phdage` from 10 simulations from the prior for $\boldsymbol{\beta}$ for four different κ_0 values. The thick line is the median salary at the prior mean $\boldsymbol{\mu}_0 = (\log(70{,}000), 2, -1.5, 0, 0, 0, 0)$.

to the prior and the likelihood. The think curve is the median salary at prior mean $\boldsymbol{\mu}_0 = (b_0, b_1, b_2, 0, 0, 0, 0)$ 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. $\kappa_0 = 397$ 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 $\kappa_0 = 397$. 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 5.9 shows the marginal posteriors for the regression coefficients for both $\kappa_0 = 397$ and $\kappa_0 = 50$; 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 5.10 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

bike share dataset

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

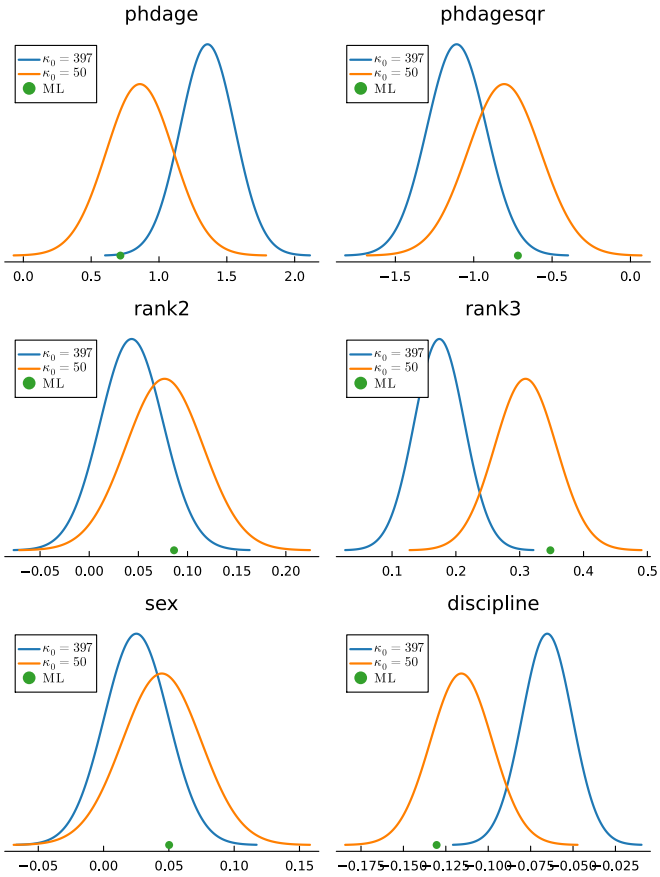


Figure 5.9: Marginal posterior densities for the regression coefficients in Gaussian linear regression fitted to the salaries data with two different priors. The maximum likelihood (ML) estimate is marked out with a green dot. See Table 5.1 for variable definitions.

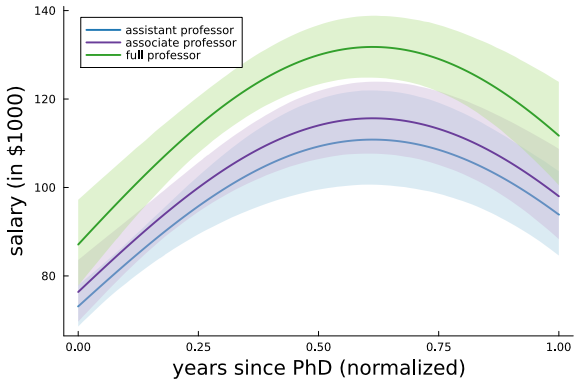


Figure 5.10: Posterior distribution (mean + 95% pointwise intervals) of the salary for female professors in Discipline B at different `phdage` for the three ranks.

may affect the demand for bikes, e.g. weather conditions, day of the week and holidays; Table 5.3 summarizes the dataset. Figure 5.11 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.

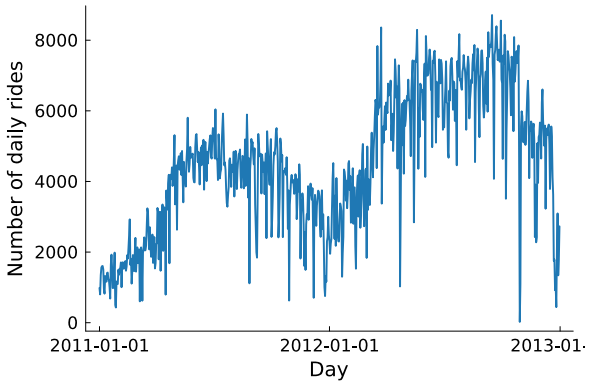


Figure 5.11: Time series plot of `nrides` in the Bike share data.

The dataset contains several categorical covariates which again need 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 5.12 shows 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 is less biking on the hottest days, but it is hard to tell when plotting 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 5.12 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 $\kappa_0 = n = 731$. 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.

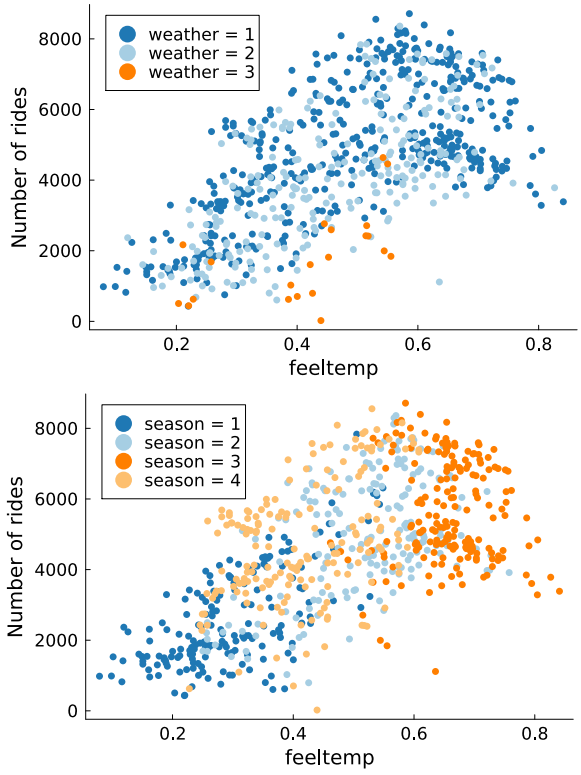


Figure 5.12: Bike share data. Scatterplots of `nrides` against `feeltemp`, color-coded by `weather` (top), `season` (bottom). See Table 5.3 for variable definitions.

An observable notebook for this example is available by clicking on the banner in the margin.

BIKE NOTEBOOK

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 summary for σ was computed by taking the square root transformation of 10,000 posterior draws of σ^2.

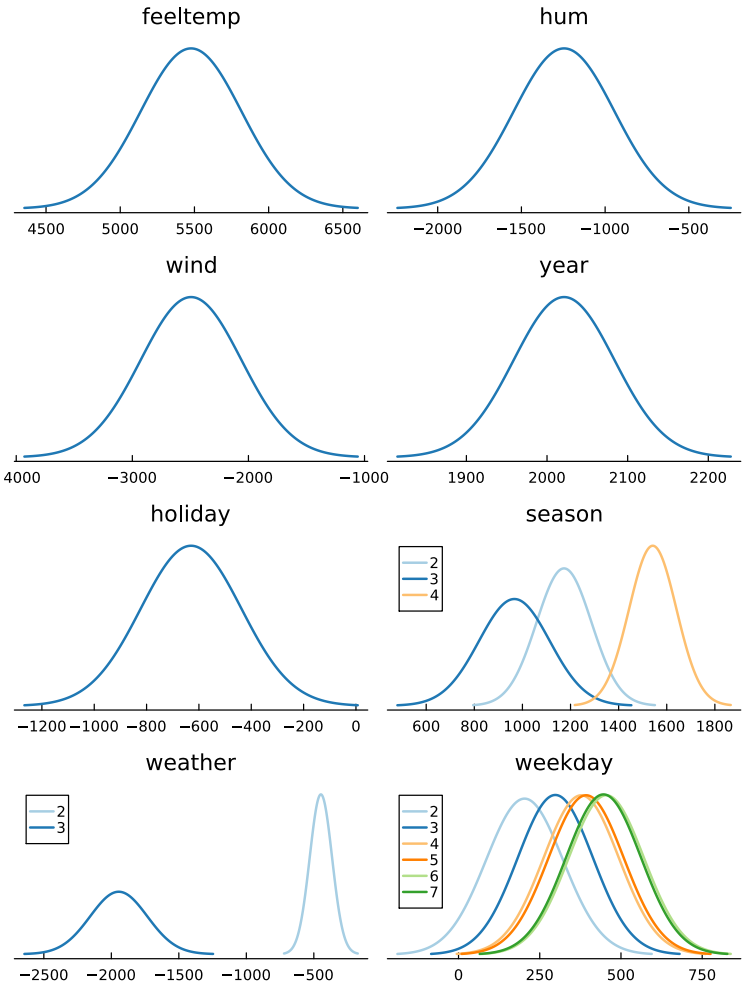


Figure 5.13: Marginal posterior densities for the regression coefficients in Gaussian linear regression fitted to the bike share data. See Table 5.3 for variable definitions.

BIKE PRIOR -> POST

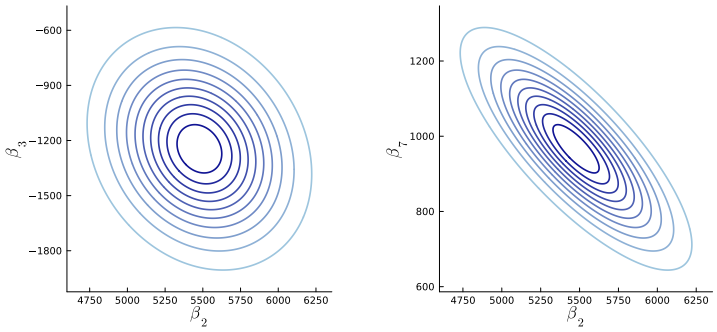


Figure 5.14: Bivariate student-t posterior densities for the regression coefficients on `feeltemp` and `hum` (left), and `feeltemp` and `season3` (right) in the bike share data. See Table 5.3 for variable definitions.

PROOFS

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

The joint posterior is

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

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 3.26 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,

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto (\sigma^2)^{-((\nu_n+p)/2+1)} \exp\left(-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right) \\
&\times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_n)^\top \Omega_n (\boldsymbol{\beta} - \boldsymbol{\mu}_n)\right) \qquad (5.11)
\end{aligned}
$$

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,

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

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 (5.11) 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.

6.1 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 person's 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 points. 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 preceding 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:

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

Law of total variance:

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

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

Iteration laws for Bayes

Marginal posterior mean:

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

Marginal posterior variance:

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

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 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 the size of the training sample, the fit and complexity of the model.

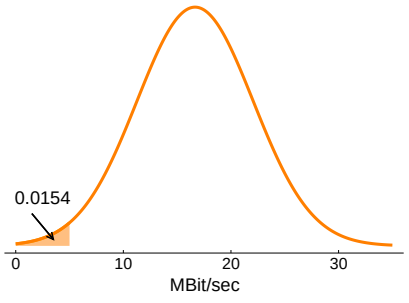


Figure 6.3: Predictive density for the internet download speed after observing $n = 5$. The probability of less than 5MBit/sec download speed is marked out by the orange region.

INTERNET SPEED PREDICTION

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, $\Pr(\tilde{y} < 5|y_1, \ldots, y_5) \approx 0.0154$.

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:

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

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 implicitly 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.

TODO! Make predictions for bike share data. Add lag to improve predictions.

Time series prediction 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:

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

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 Gibbs sampling. The second step is implemented using the usual sequential decomposition of a joint distribution

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

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 p**th 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 Gibbs sampling. The terms in orange font are future values used in the prediction which have been simulated in earlier time steps.

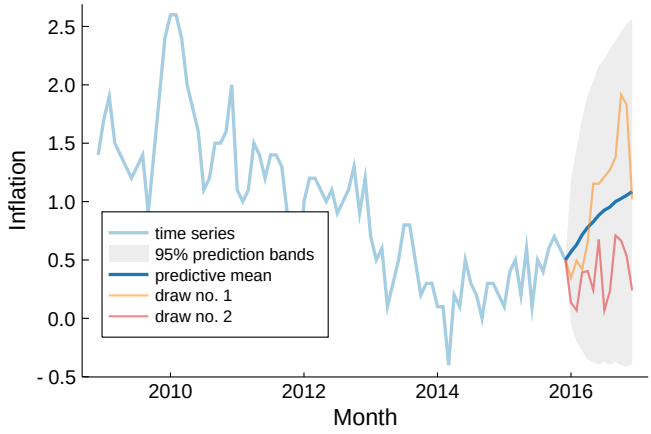


Figure 6.7: Predictive distribution $h = 12$ steps ahead for Swedish inflation represented by a mean prediction in dark blue and 95% predictive intervals as the gray region. Two of the $m = 10{,}000$ simulated paths from the algorithm in Figure 6.6 are marked out.

6.2 *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 effect in 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 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 vary from person to person. Table 6.1 presents 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_I	$u(a_I, \theta_1)$	$u(a_I, \theta_2)$	$\cdots$	$u(a_I, \theta_K)$

Table 6.1: Utility table.

Table 6.2 presents a toy decision problem where the choice is between bringing or not bringing an umbrella with you today. The

consequences of this decision depend 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 your umbrella at 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 for 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. A central bank with an explicit inflation target needs to continually decide the level of their steering rate (a) to simultaneously reach a 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 central bank's steering rate, a. Here the set of actions $\mathcal{A}$ can be considered discrete (steering 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 problem 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 balancing 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 an umbrella since the worst utility for this choice is 10 (it rains) whereas 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; break the Bayesian eggs and you too can enjoy a Bayesian omelette.

Figure 6.8 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 as

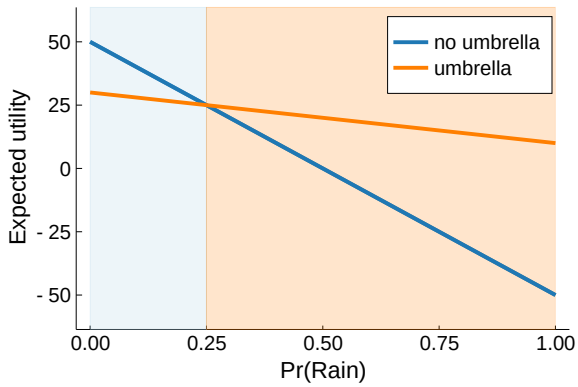


Figure 6.8: Expected utility of bringing an umbrella as function of the probability of rain. The shaded regions mark out the action that maximizes expected utility.

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

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

since the cross-term is zero by the fact that $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}(\theta - \mathbb{E}(\theta|\mathbf{x})) = 0$. 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 the 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.

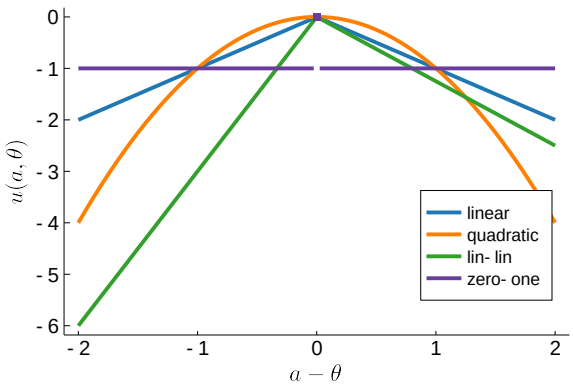


Figure 6.9: Utility functions for point estimation as a function of estimation error $a-\theta$. The lin-lin utility has $c_1 = 3$ and $c_2 = 1.25$.

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 $c_1/(c_1+$

percentile

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 $p(\theta|\mathbf{x})$.

The four presented utility functions are plotted in Figure 6.9 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 $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$, $\tilde{\mathbf{X}}$ is assumed known and $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\varepsilon}}$ are both normal, we immediately see that $p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \sigma^2, \mathbf{y})$ is multivariate normal with

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

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 $x_1, \ldots, x_n \overset{iid}{\sim} \text{Bern}(\theta)$, 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 of the weather today, the 11th day. Your utility for the action-state combinations are given in the table below. Assume a Beta$(1,1)$ prior for θ. Compute the Bayesian decision.

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

2. 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 $x_1, \ldots, x_n \overset{iid}{\sim} \text{N}(\theta, \sigma^2)$ 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 Normal posterior approximation

7.1 Intractable posterior and approximation

So far in this book we have analyzed models where we could always find a conjugate prior. The posterior then belongs to the same distributional family and updating the prior with new data is straightforward, often by adding some summary of the data to the prior hyperparameters. Unfortunately, in many models we simply cannot find a conjugate prior, or even a prior that gives the posterior in a mathematically tractable form. Here is an example.

Beta distribution as a model for proportions. Many problems involve data in the form of proportions, i.e. values in the unit interval $[0, 1]$. Some examples of data in the form of proportions are financial debt ratios for firms and the proportion of a certain substance in a container. The Beta(α, β) distribution can be used to model such data. Note that we are here using the Beta distribution as a model for the data, not as a prior for Bernoulli probability, and the interest is in the posterior distribution for the Beta parameters α and β. Assuming the model $x_1, \ldots, x_n | \alpha, \beta \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta)$, the likelihood function is

$$\begin{aligned} p(x_1, \ldots, x_n | \alpha, \beta) &= \prod_{i=1}^{n} \frac{1}{B(\alpha, \beta)} x_i^{\alpha-1} (1 - x_i)^{\beta-1} \\ &= \frac{1}{B(\alpha, \beta)^n} \Big(\prod_{i=1}^{n} x_i \Big)^{\alpha-1} \Big(\prod_{i=1}^{n} (1 - x_i) \Big)^{\beta-1}, \end{aligned}$$

where $B(\alpha, \beta)$ is as usual the Beta function. Since the model parameters α and β are partly located inside the fairly complicated Beta function, it is hard to see how one can find a prior that would make the posterior belong to a known distributional family.

The previous example is common: we can compute the unnormalized posterior $p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$ for any values of the parameters $\boldsymbol{\theta}$, but we cannot recognize the posterior as belonging to a known distribution, and we cannot compute the normalizing constant

$\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\theta$. We then say that our problem has an **intractable posterior**. There are three main ways to proceed whenever the posterior distribution is intractable.

intractable posterior

First, a brute force solution is to evaluate the unnormalized posterior $p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ over a rectangular grid of $\boldsymbol{\theta}$ values

$$(\theta_1, \theta_2, \ldots, \theta_p), \text{ for } \theta_1 \in \{\theta_1^{(1)}, \ldots, \theta_1^{(N)}\}, \ldots, \theta_p \in \{\theta_p^{(1)}, \ldots, \theta_p^{(N)}\},$$

and use numerical integration to compute the normalizing constant. The problem with this approach is that the number of grid points is N^p, which grows exponentially with the number of regression coefficients in $\boldsymbol{\theta}$; if 100 grid points are needed to get accurate numerical results in one dimension, then $100^2 = 10000$ grid points are needed in two dimensions, 1 million points in three dimensions and so on. Exercise 1 asks you use this technique in one and two dimensions.

Second, we can explore the posterior distribution by simulation, as we have already seen in Multi-parameter models and Regression. In those chapters the posterior was tractable, and we used simulation to compute the posterior distribution of transformations of the parameters. We will see in later chapters how powerful simulation algorithms can be used to simulate from intractable posteriors.

A third approach to dealing with intractable posteriors is to approximate the posterior with a simpler tractable distribution. There are several general approaches to approximating a posterior distribution, for example the variational inference method presented in Variational inference. This chapter presents a simple but often very accurate normal approximation of the posterior, and explains why the normal distribution is a particularly accurate approximation when the dataset is large. The method is shown to have great appeal for practical work since it can be automated on a computer with little effort.

7.2 *Taylor approximation of the posterior*

The section Taylor approximation in the mathematical appendix states that the Kth order Taylor approximation of a function $f(x)$ around the expansion point $x = a$ is

$$f(x) \approx \sum_{k=0}^{K} \frac{f^{(k)}(a)}{k!}(x-a)^k, \tag{7.1}$$

where $f^{(k)}(a)$ is the kth derivative of $f(x)$ evaluated at $x = a$. The 0th order derivative is just the function itself $f^{(0)}(x) = f(x)$ and $0! = 1$.

Taylor approximations are local approximations that are tailored to the function $f(x)$ around the point $x = a$, and are therefore accurate

in a region around $x = a$. Since the posterior distribution will be concentrated in a small region around its mode whenever we have large datasets, we can expect a Taylor approximation of the posterior to be accurate in large datasets. This also suggests that $a = \tilde{\theta}$, where $\tilde{\theta}$ is the posterior mode, is a natural expansion point for the approximation.

To give a first illustration of using the Taylor approximation for posterior approximation, consider a posterior distribution of the form

$$p(\theta|\mathbf{y}) \propto \exp\left(-\exp(\theta/\kappa_0)(\theta-\bar{y})^2\right), \tag{7.2}$$

where κ_0 is a prior hyperparameter and $\bar{y}$ is the sample mean, and $\theta \in (-\infty, \infty)$. Let us not worry about what kind of model, prior and data gave rise to the posterior in (7.2), just note that the posterior is *not* normal. We will do a Taylor approximation of the *logarithm* of the posterior distribution, $\log p(\theta|\mathbf{y}) \propto -\exp(\theta/\kappa_0)(\theta-\bar{y})^2$. The reason for expanding the *log* posterior is that it can often be approximated well with a low order Taylor approximation (see Theorem 2 below). Using the product rule, the first derivative of the log posterior is

$$\frac{\partial \log p(\theta|\mathbf{y})}{\partial\theta} = -\frac{1}{\kappa_0}\exp\left(\frac{\theta}{\kappa_0}\right)(\theta-\bar{y})^2 - \exp\left(\frac{\theta}{\kappa_0}\right)2(\theta-\bar{y}),$$

which is clearly zero at $\theta = \bar{y}$, so $\bar{y}$ is the posterior mode. The Taylor approximation will therefore be around $\theta = \bar{y}$. The second derivative is

$$\begin{aligned}\frac{\partial^2 \log p(\theta|\mathbf{y})}{\partial\theta^2} = &-\frac{1}{\kappa_0^2}\exp\left(\frac{\theta}{\kappa_0}\right)(\theta-\bar{y})^2 - \frac{2}{\kappa_0}\exp\left(\frac{\theta}{\kappa_0}\right)(\theta-\bar{y}) \\ &-\frac{2}{\kappa_0}\exp\left(\frac{\theta}{\kappa_0}\right)(\theta-\bar{y}) - 2\exp\left(\frac{\theta}{\kappa_0}\right),\end{aligned}$$

which is $-2\exp(\bar{y}/\kappa_0) < 0$ at $\theta = \bar{y}$. We can continue in the same fashion to compute the third and fourth derivative, to finally obtain a fourth order Taylor approximation of the log posterior by inserting the derivatives in (7.1):

$$\begin{aligned}\log p(\theta|\mathbf{y}) \approx &-\exp(\bar{y}/\kappa_0)(\theta-\bar{y})^2 - \frac{\exp(\bar{y}/\kappa_0)}{\kappa_0}(\theta-\bar{y})^3 \\ &-\frac{\exp(\bar{y}/\kappa_0)}{2\kappa_0^2}(\theta-\bar{y})^4.\end{aligned}$$

The graph to the left in Figure 7.1 shows the Taylor approximation of $\log p(\theta|\mathbf{y})$ for the case $\bar{y} = 2$ and $\kappa_0 = 20$. The approximation improves as we increase the polynomial order, and the fourth order approximation is very accurate for all $\theta \in [-10, 10]$. The second order approximation seems to be too crude, the approximation error

is large for all θ outside of the interval $(-1, 5)$. However, and this is the important part, the posterior is negligible outside of the interval $(-1, 5)$, so we really do not care if the approximation is poor there. This is shown in the graph to the right in Figure 7.1, which plots the posterior and the implied Taylor approximation on the original scale $p(\theta|\mathbf{y}) \propto \exp(\log p(\theta|\mathbf{y}))$. Even the second order approximation is more or less perfect on the original scale.

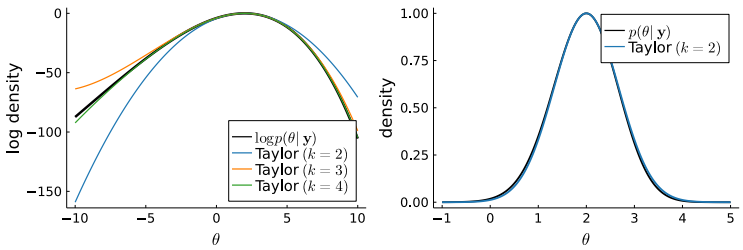


Figure 7.1: Taylor approximation of the posterior distribution $p(\theta|\mathbf{y}) = \exp\left(-\exp(x/20)(x-2)^2\right)$ around $\theta = 2$. The figure on the left show Taylor approximations of the log posterior for different polynomial orders. The figure on the right shows the implied second order approximation of the posterior on the original scale.

7.3 *Normal posterior approximation and large sample asymptotics*

Note that the second order Taylor approximation of the log posterior in the previous example implies the posterior approximation

$$p(\theta|\mathbf{y}) \approx \exp\Big(-2\exp(\bar{y}/\kappa_0)(\theta-\bar{y})^2\Big),$$

which can be recognized as the normal distribution

$$\theta|\mathbf{y} \sim \mathrm{N}\left(\bar{y}, \frac{1}{4\exp(\bar{y}/\kappa_0)}\right).$$

In fact, a posterior based on a second order Taylor approximation of the log posterior is always a normal distribution. This can be seen as follows. A second order approximation of the log posterior is

$$\begin{aligned}\log p(\theta|\mathbf{y}) \approx \log p(\tilde{\theta}|\mathbf{y}) &+ \frac{\partial \log p(\theta|\mathbf{y})}{\partial\theta}|_{\theta=\tilde{\theta}}(\theta-\tilde{\theta}) \\ &+ \frac{1}{2}\frac{\partial^2 \log p(\theta|\mathbf{y})}{\partial\theta^2}|_{\theta=\tilde{\theta}}(\theta-\tilde{\theta})^2 \qquad (7.3)\end{aligned}$$

where the first order term is zero since

$$\frac{\partial \log p(\theta|\mathbf{y})}{\partial\theta}|_{\theta=\tilde{\theta}} = 0$$

from the definition of the posterior mode. Hence, taking exponentials on both sides of (7.3) gives

$$\begin{aligned}p(\theta|\mathbf{y}) &\approx \exp\Big(\log p(\tilde{\theta}|\mathbf{y})\Big)\exp\left(\frac{1}{2}\frac{\partial^2 \log p(\theta|\mathbf{y})}{\partial\theta^2}|_{\theta=\tilde{\theta}}(\theta-\tilde{\theta})^2\right) \\ &\propto \exp\left(\frac{1}{2}\frac{\partial^2 \log p(\theta|\mathbf{y})}{\partial\theta^2}|_{\theta=\tilde{\theta}}(\theta-\tilde{\theta})^2\right)\end{aligned}$$

since $\exp(\log p(\tilde{\theta}|\mathbf{y}))$ does not depend on θ (the mode $\tilde{\theta}$ is just a number for a given dataset). Using the definition of the observed information

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

from the section Likelihood and Information we have the approximation

$$p(\theta|\mathbf{y}) \approx \exp\Big(-\frac{1}{2}J_{\theta,\mathbf{y}}(\tilde{\theta})(\theta-\tilde{\theta})^2\Big).$$

Hence, we have the following normal posterior approximation

$$\theta|\mathbf{y} \overset{\text{a}}{\sim} \text{N}\Big(\tilde{\theta}, J^{-1}_{\theta,\mathbf{y}}(\tilde{\theta})\Big),$$

where the symbol $\overset{\text{a}}{\sim}$ denotes "is approximately distributed as".

The following theorem shows that this normal posterior approximation will become more and more accurate with the size of the dataset.

Theorem 2 (large sample normality of posterior)**.** *The posterior distribution of θ conditional on data $\mathbf{y} = (y_1, \ldots, y_n)$ converges to a normal distribution in large samples:*

$$J^{1/2}_{\theta,\mathbf{y}}(\tilde{\theta})(\theta-\tilde{\theta}) \,|\, \mathbf{y} \xrightarrow{d} \text{N}(0,1), \text{ as } n \to \infty,$$

where $\tilde{\theta}$ is the posterior mode and

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

is the observed information at $\tilde{\theta}$.

The result in Theorem 2 is often called the **Bernstein-von Mises theorem** after the persons that proved a similar result. The result requires some regularity conditions, for example that the posterior distribution becomes concentrated in a small neighborhood around the posterior mode as the sample size grows large; see Bernardo and Smith (2009) for some details. These conditions are met in most models used in practise, but we will give an example where they are not, and the result in Theorem 2 fails to hold.

Bernstein-von Mises theorem

Theorem 2 does not tell us how large n must be for the approximation to be accurate, and this will be model specific. But the convergence happens quickly in many problems and the normal approximation is often accurate enough for practical applications.

NORMAL APPROXIMATION OF A GAMMA POSTERIOR. Consider the iid Poisson model $x_1, \ldots, x_n|\theta \overset{\text{iid}}{\sim} \text{Pois}(\theta)$ with the conjugate $\theta \sim \text{Gamma}(\alpha, \beta)$ prior. Here we know that the posterior is the

Gamma$(\alpha + \sum y_i, \beta + n)$ distribution, so there is no need to approximate it. Let us however as an exercise derive a normal approximation of this posterior and compare it with the exact posterior. The log posterior density is

$$\log p(\theta|\mathbf{x}) \propto (\alpha + \sum x_i - 1) \log \theta - \theta(\beta + n)$$

with first derivative

$$\frac{\partial \log p(\theta|\mathbf{x})}{\partial \theta} = \frac{\alpha + \sum x_i - 1}{\theta} - (\beta + n).$$

Setting the first derivative to zero and solving for θ give the posterior mode

$$\tilde{\theta} = \frac{\alpha + \sum x_i - 1}{\beta + n}. \tag{7.4}$$

The second derivative at the mode $\tilde{\theta}$ is

$$\frac{\partial^2 \ln p(\theta|\mathbf{x})}{\partial \theta^2}|_{\theta=\tilde{\theta}} = -\frac{\alpha + \sum x_i - 1}{\left(\frac{\alpha+\sum x_i-1}{\beta+n}\right)^2} = -\frac{(\beta+n)^2}{\alpha + \sum x_i - 1},$$

which is negative for all θ if $\alpha + \sum x_i > 1$, i.e. if at least one observation is non-zero or if $\alpha > 1$, so this is essentially always satisfied; hence $\tilde{\theta}$ in (7.4) is really the mode.

In summary, the normal posterior approximation is

$$\theta|\mathbf{x} \sim \mathrm{N}\left(\frac{\alpha + \sum x_i - 1}{\beta + n}, \frac{\alpha + \sum x_i - 1}{(\beta + n)^2}\right).$$

Figure 7.2 displays the normal approximation and the true posterior for the number of bidders in the eBay data with increasing sample sizes from the first $n = 5$ observations in the data to the first $n = 50$ observations. The normal approximation is crude for the smallest sample size, but already with $n = 10$ it is rather accurate and at $n = 50$ it is nearly perfect.

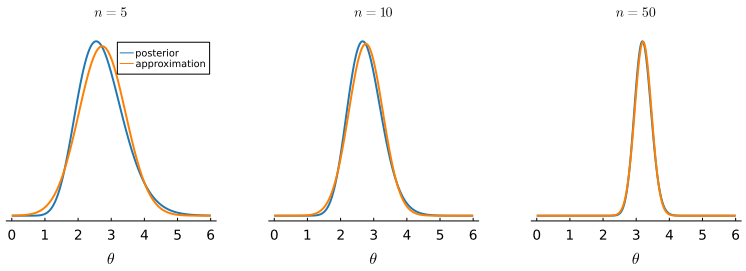


Figure 7.2: Normal approximation of the Gamma posterior in the Poisson model for the number of bidders in the eBay data for different sample sizes.

The posterior normality in large samples in Theorem 2 also holds when the model parameter is a vector, hence motivating the multivariate normal approximation of the posterior for a d-dimensional parameter vector $\boldsymbol{\theta}$ in Figure 7.3.

Reparametrization example

Simulation after normal approx - Gini

Normal posterior approximation

The posterior can in large samples be approximated by

$$\boldsymbol{\theta}|\mathbf{y} \overset{\text{a}}{\sim} \text{N}\left(\tilde{\boldsymbol{\theta}}, J^{-1}_{\tilde{\boldsymbol{\theta}},\mathbf{y}}(\tilde{\boldsymbol{\theta}})\right)$$

where $\tilde{\boldsymbol{\theta}}$ is the posterior mode and

$$J_{\tilde{\boldsymbol{\theta}},\mathbf{y}} = -\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$$

is the $d \times d$ observed information matrix at $\tilde{\boldsymbol{\theta}}$.

Figure 7.3: Multivariate normal approximation of a posterior distribution.

Approximating the posterior in a student-t model. This example gives an illustration where one of the conditions of the Bernstein-von Mises theorem is violated and the posterior is not asymptotically normal.

Let us first start with an example where asymptotic normality holds. Consider data coming from a standard student-t distribution with $\nu = 4$ degrees of freedom, i.e. $y_1, \ldots, y_n \overset{\text{iid}}{\sim} t_4(0,1)$. The same issues will appear for models with unknown location and scale in the student-t distribution. We will fit the model $y_1, \ldots, y_n|\nu \overset{\text{iid}}{\sim} t_\nu(0,1)$ to the data. The posterior $p(\nu|y_1, \ldots, y_n)$ is intractable for any prior. We will first use a non-informative uniform prior over $\nu \in (0, \infty)$ to expose problems in the likelihood for this model, and then add a more informative prior. Figure 7.4 shows that when the data comes from a $t_4(0,1)$ distribution, the normal posterior approximation improves as we increase the sample size, as suggested by the Bernstein-von Mises theorem.

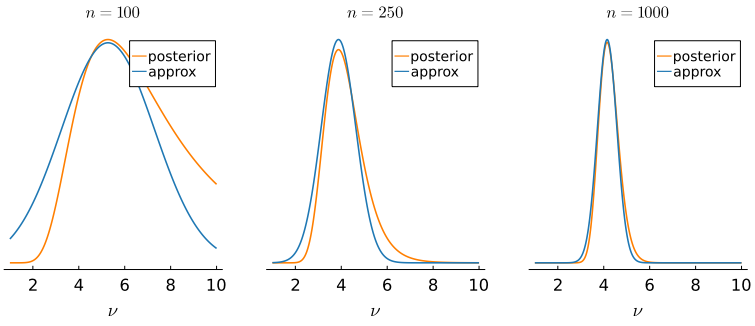


Figure 7.4: Normal posterior approximation for the degrees of freedom in the $t_\nu(0,1)$ model fitted to iid data from the $t_4(0,1)$ distribution. The posterior clearly tends toward normality as the sample size, n, increases.

Assume now that the data is generated from a $N(0,1)$ distribution, but we still fit a $t_\nu(0,1)$ model to the data. Note that the $N(0,1)$ data generating process is a student-t distribution where $\nu \to \infty$. Figure 7.5 shows that the normal posterior approximation is a disaster here.

The reason is that the likelihood is essentially flat for all $\nu > 50$ since a t_{50} or t_{100} are more or less identical models as both are very close to the Normal model, i.e. $\nu \to \infty$. The problem here is that the true parameter value ($\nu = \infty$) is at the boundary of the parameter space and all posterior mass will therefore 'pile up at infinity' for large sample sizes. This violates the so called steepness assumption needed for the Bernstein-von Mises theorem (see Bernardo and Smith (2009)), and the posterior does not tend to a normal distribution in large samples.

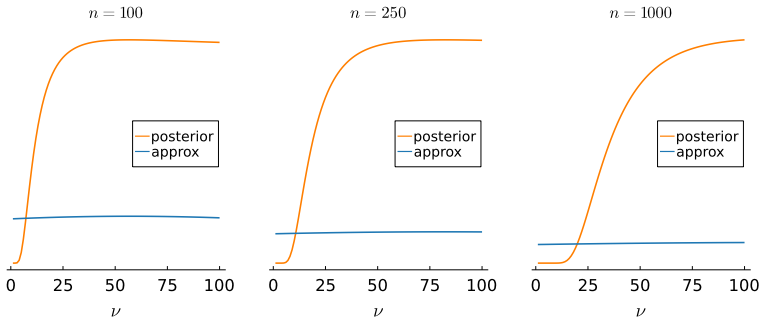


Figure 7.5: Normal posterior approximation for the degrees of freedom in the $t_\nu(0,1)$ model fitted to iid data from the $N(0,1)$ distribution. The prior for ν is uniform over $(0,\infty)$ to highlight properties of the likelihood. The posterior does not tend to normality even at large sample size.

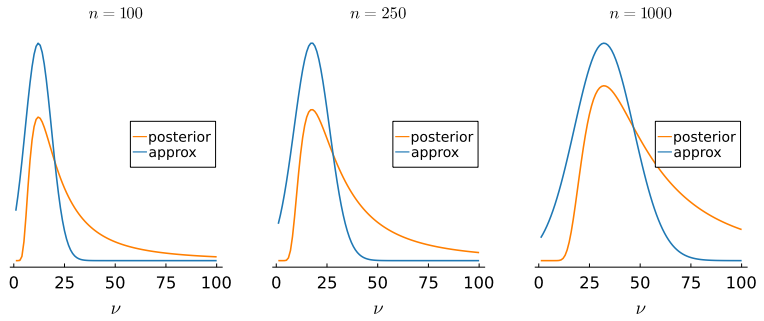


Figure 7.6: Normal posterior approximation for the degrees of freedom in the $t_\nu(0,1)$ model fitted to iid data from the $N(0,1)$ distribution. The prior $\nu \sim \text{Inv}{-}\chi^2(5,10)$ gives some curvature to the posterior, but the normal approximation is poor even at $n = 1000$.

Figure 7.6 shows the posterior when using a $\nu \sim \text{Inv}{-}\chi^2(5,10)$ prior. The prior adds curvature to flat regions of the likelihood, but even for $n = 1000$ is the normal approximation very poor. Moreover, the prior will be dominated by the likelihood as the sample size increases further, so the same problems we saw when using the flat prior will reappear as we add more and more data.

7.4 *Computing the normal posterior approximation numerically*

Computing the matrix of second derivatives needed in the normal approximation in Figure 7.3 can be tedious. Moreover, the posterior mode may not be available in closed form since the system of equations

$$\frac{\partial \log p(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = 0 \tag{7.5}$$

is often non-linear without analytical solution. This is for example the case in the logistic regression model.

We will now explain how a computer with numerical optimization routines can be used to automatically find the posterior mode $\tilde{\boldsymbol{\theta}}$ and the observed information matrix. It will be sufficient to code up the log-likelihood function and the log prior, and then let the computer do all the tedious equation solving and differentiation.

First, we can use Newton's method to solve the system of equations in 7.5 for the posterior mode. **Newton's method** starts with an initial value $\boldsymbol{\theta}^{(0)}$ and iterates for $t = 1, 2, \ldots$ until convergence:

Newton's method

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \mathbf{H}\big(\boldsymbol{\theta}^{(t-1)}\big)^{-1}\mathbf{g}\big(\boldsymbol{\theta}^{(t-1)}\big),$$

where $g\big(\boldsymbol{\theta}^{(t-1)}\big)$ is the gradient and $\mathbf{H}\big(\boldsymbol{\theta}^{(t-1)}\big)$ the Hessian matrix of $\log p(\boldsymbol{\theta}|\mathbf{y})$ at the previous parameter value $\boldsymbol{\theta}^{(t-1)}$.

Newton's method requires the gradient and Hessian of the log posterior, which can be obtained in most modern programming languages by **automatic differentiation**. Automatic differentiation is a technique that applies the chain rule for differentiation from Calculus in clever algorithmic ways to produce derivatives that are both numerically accurate and fast to compute, even for functions with many inputs. Alternatively, there are optimizers like the BFGS algorithm that returns both the posterior mode $\tilde{\boldsymbol{\theta}}$ and the Hessian (observed information), where the Hessian matrix is iteratively built up during the iterations of the algorithm. The bottom line is that there are many ways to obtain all we need for the normal approximation in Figure 7.3 using a computer. All we need to code is a function that computes $\log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$ for any value of $\boldsymbol{\theta}$ for a fixed dataset $\mathbf{y}$. Note that we only need to code the proportional form of Bayes' theorem (on the log scale), since the normalizing constant does not depend on $\boldsymbol{\theta}$ and will therefore not affect the optimization for the posterior mode or the observed information matrix.

automatic differentiation

EXERCISES

1. The wind direction was measured once a month at a given location. The measurements for the first ten months were

   $$\mathbf{y} = (-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02),$$

   recorded in radians $-\pi \leq y_i \leq \pi$ with South located at zero radians; see Figure 7.4. Assume that these data points are independent observations following the **von Mises** distribution for directional data (see Figure 7.8)

   von Mises

   $$y_1, \ldots, y_n | \mu, \kappa \overset{\text{iid}}{\sim} \text{VM}(\mu, \kappa).$$

(a) Assume that μ is known to be 2.39, and $\kappa \sim \text{Expon}(\theta = 1)$ a priori. Plot the posterior distribution of κ over a fine grid of κ values.

(b) Use numerical optimization to approximate the posterior distribution of κ and plot the approximation in the same graph.

(c) Assume now that both μ and κ are unknown. Plot the bivariate posterior $p(\mu, \kappa|\mathbf{y})$ over a two-dimensional grid of (μ, κ) pairs.

(d) Use numerical optimization to approximate the bivariate posterior distribution $p(\mu, \kappa|\mathbf{y})$.

2. Next problem!

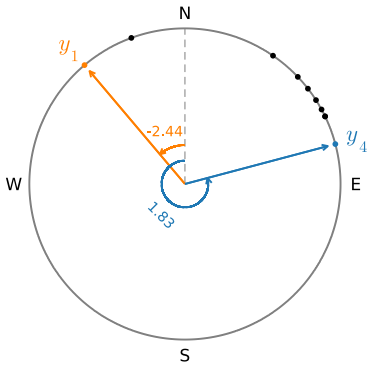


Figure 7.7: Wind direction data in radians $y \in [-\pi, \pi]$, with South at $y = 0$ radians.

Von Mises distribution

$X \sim \text{VM}(\mu, \kappa)$ for $X \in [-\pi, \pi]$

is a common distribution for directional data.

$$p(x) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}$$
$$\mathbb{E}(X) = \mu$$
$$\mathbb{V}(X) = 1 - I_1(\kappa)/I_0(\kappa),$$

where $I_v(\kappa)$ is the modified Bessel function of the first kind of order v, implemented as `besseli` in many programming languages.
The variance can be shown to be decreasing in κ, so κ is the precision of the distribution.

Figure 7.8: Von Mises distribution.

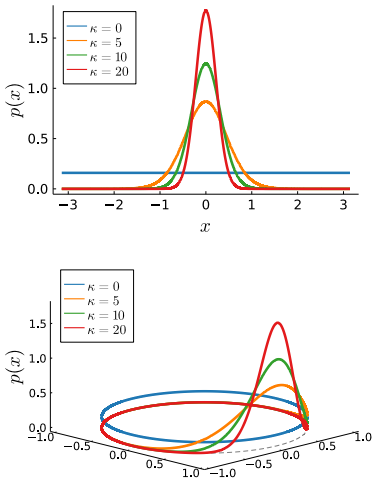


Figure 7.9: Some Von Mises distributions as functions of the angle (top) and over the circle (bottom).

8 Classification

The Regression chapter presented Bayesian learning and prediction for a *continuous* response variable y explained by a set of covariates $\mathbf{x}$; the covariates may be binary, categorical or continuous. Classification is instead when we are modeling a *binary* or *categorical response variable* as a function of covariates $\mathbf{x}$. This distinction in terminology between regression and classification is not always upheld, for example one of the more commonly used classification models is named logistic regression.

Modeling non-continuous response variables requires other models than the Gaussian linear regression model, and this chapter will present Bayesian learning, prediction and decision making for such models.

The likelihood function can be complex in classification problems, and the posterior distribution is therefore often mathematically intractable. We will therefore use the normal posterior approximation presented in Chapter Normal posterior approximation, and later chapters will present simulation-based methods to explore the posterior in models used for classification.

8.1 Classification problems

Binary classification

Many interesting problems involve modeling and predicting a **binary response variable** $y \in \{0, 1\}$. The coding of the two different values need not be $0/1$, but can equally well be `True/False`, or something application specific such as `Heads/Tails` in coin tossing; the coding $y \in \{-1, 1\}$ is common in machine learning. A binary variable is said to have two possible **classes**, for example the two classes `Heads` and `Tails` in coin tossing. It is also common to distinguish the two classes by the generic labels **`positive class`** and **`negative class`**, where `positive` does not necessarily mean positive in the usual sense, but may for example indicate the presence of a disease. Here are some examples of binary classification problems.

binary response variable

classes

positive class

negative class

EXAMPLE: SPAM PREDICTION. You want to build a spam filter that can determine if a newly arrived email is `spam` (perhaps coded as $y = 1$) or `ham` (coded as $y = 0$). The spam prediction can use covariates based on the processed text in the email. For example, dummy variables that indicate the presence of certain trigger words for spam, or covariates based on the number of $-signs or CAPITAL LETTERS in the given email; see Figure 8.1.

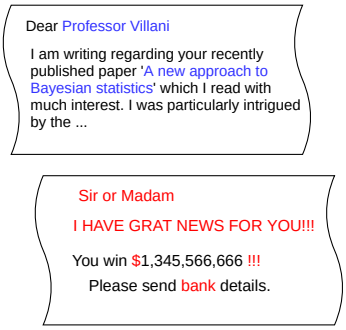


Figure 8.1: Indicators for spam in red text and ham in blue text.

EXAMPLE: INTERNET AUCTION SALE. What determines if an internet auction ends up in a sale? This can be analyzed by collecting data on many past auctions and recording the response variable `sold` ($y = 1$) and `not sold` ($y = 0$) for each auction. To aid in the classification one can also collect information (covariates) about each auction, for example the seller's reservation price, the feedback/review score of the seller, and measures of the auctioned object's condition determined from the seller's text description, or visual inspection of the posted images by a human.

The aim in binary classification problems is the probability of the positive class, $\Pr(y = 1|\mathbf{x})$, conditional on a set of covariates $\mathbf{x}$. We then of course immediately get the probability of the negative class $\Pr(y = 0|\mathbf{x}) = 1 - \Pr(y = 1|\mathbf{x})$. The importance of $\Pr(y = 1|\mathbf{x})$ in prediction problems should be clear: computing $\Pr(y = 1|\tilde{\mathbf{x}})$ for a new observation $\tilde{\mathbf{x}}$ gives the probability of the positive class, which can be directly used for Bayesian decision making. For example, consider a planned auction where we can use the reservation price, seller and object information as covariates to compute the probability that the object will be sold. This probability can be used by the seller to make a decision of whether or not to put the object up for auction, or for determining a more suitable reservation price that increases the sale probability.

Multi-class classification

Many problems involve more than two classes. **Multi-class classification** has a response variable y that belongs to exactly one of C possible classes or categories. We have seen such categorical data before in section Multinomial data where a Bayesian analysis of multinomial data with a Dirichlet prior was presented. Here we will model the class probabilities $\Pr(y = c|\mathbf{x})$, $c \in \{1, \ldots, C\}$, conditional on a set of covariates $\mathbf{x}$.

Multi-class classification

EXAMPLE: MARKETING BRAND PREDICTION. Customers can often choose from several competing brands when shopping. Marketers

can build multi-class classification models to predict the brand choice ($y \in \{1, \ldots, C\}$) of a customer from covariates ($\mathbf{x}$) constructed from personal data (age, income, residence area, sex) and characteristics of the product (price, price of competing brands, placement).

EXAMPLE: IMAGE CLASSIFICATION. An image consists of a large number of pixels, where each pixel is color-coded according to some color system; for example RGB where each pixel is described by a three-dimensional vector with numbers ranging from $0-255$. Each RGB vector gives the composition of red (first element), green (second element) and blue (third element) colors in the pixel. Consider a dataset of images where each observation is an image with a label (y) that describes the category of the image ("dog", "cat", "human", "car", "train" etc.) and three covariates for each pixel in the image. A self-driving car robot is using multi-class classification to determine the category of an object from camera images, and other sensors.

8.2 *Logistic regression*

We will use the notation $\Pr(y = y^\star|\mathbf{x})$, to denote the probability that the binary class variable y takes the value $y^\star \in \{0, 1\}$. **Logistic regression** assumes that the responses $y_1, \ldots, y_n$ are independent conditional on the covariates and the probability of the positive class is modelled by

$$\Pr(y = 1|\mathbf{x}, \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}^\top \boldsymbol{\beta})}, \tag{8.1}$$

where $\mathbf{x}$ is a p-dimensional vector with covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients. A common alternative form is obtained by multiplying both the numerator and denominator of (8.1) by $\exp(-\mathbf{x}^\top \boldsymbol{\beta})$ to obtain

$$\Pr(y = 1|\mathbf{x}, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{x}^\top \boldsymbol{\beta})}. \tag{8.2}$$

The function $f(x) = 1/(1 + \exp(-x))$ is called the **logistic function** (see Figure 8.2), hence the name logistic regression. Logistic regression is similar to the usual linear regression in that the linear combination $\mathbf{x}^\top \boldsymbol{\beta}$ is the connection between the covariates $\mathbf{x}$ and the response y. The role of the logistic function is to 'squash' $\mathbf{x}^\top \boldsymbol{\beta}$ so that the end result is a number between 0 and 1, which is required here since we are modeling a probability, $0 \leq \Pr(y = 1|\mathbf{x}) \leq 1$.

There are many other squashing functions that can be used instead of the logistic, for example the distribution function (CDF) $\Phi(z)$ of the standard normal distribution, which gives rise to the popular **probit regression** model

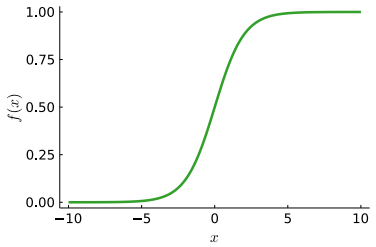


Figure 8.2: The logistic function $f(x) = 1/(1 + e^{-x})$.

Logistic regression

logistic function

probit regression

$$\Pr(y = 1|\mathbf{x}, \boldsymbol{\beta}) = \Phi(\mathbf{x}^\top \boldsymbol{\beta}). \tag{8.3}$$

The fact that $\Phi(z)$ is a CDF guarantees that $0 \leq \Pr(Y = y|\mathbf{x}) \leq 1$. We will return to this model later in the book.

The parameters in the logistic regression model are most easily interpreted in odds form. To see this, note first that the complementary probability is

$$\Pr(y = 0|\mathbf{x}, \boldsymbol{\beta}) = 1 - \Pr(y = 1|\mathbf{x}, \boldsymbol{\beta}) = \frac{1}{1 + \exp(\mathbf{x}^\top \boldsymbol{\beta})}, \tag{8.4}$$

and the odds of the positive class compared to the negative class is therefore

$$\text{Odds}(\mathbf{x}) \equiv \frac{\Pr(\text{y} = 1|\mathbf{x})}{\Pr(\text{y} = 0|\mathbf{x})} = \exp(\mathbf{x}^\top \boldsymbol{\beta}). \tag{8.5}$$

Consider now how changing the jth covariate by one unit affects the odds. In particular we will look at the **odds ratio** (OR) for the jth covariate

odds ratio

$$\text{OR}_j = \frac{\text{Odds}\big(\mathbf{x} = (x_1, \ldots, x_j + 1, \ldots, x_p)\big)}{\text{Odds}\big(\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_p)\big)} = \exp(\beta_j). \tag{8.6}$$

The important fact about an odds ratio from logistic regression is that it does not depend on the value of the covariates $\mathbf{x}$. A value of $\exp(\beta_j)$ of 1.01 has the interpretation that the odds for the positive class increases by 1% whenever x_j increases by one unit, regardless of the value for the other covariates, or the value of x_j before the unit change. For this reason, it is common to report inferences for $\exp(\beta_j)$ rather than β_j.

Finally, note that log odds is a linear function of the covariates

$$\text{LogOdds}(\mathbf{x}) \equiv \log \frac{\Pr(\text{y} = 1|\mathbf{x})}{\Pr(\text{y} = 0|\mathbf{x})} = \mathbf{x}^\top \boldsymbol{\beta}. \tag{8.7}$$

The logistic regression is therefore often said to be a linear model, even though the probability of the response is clearly a nonlinear function of the covariates. One important implication of the linearity in the log odds is that the decision boundaries that separate the two classes are linear. The logistic regression model is therefore not suitable for classification problems where the classes are not linearly separable. The Gaussian process extension of the logistic regression presented in Gaussian processes is an interesting nonlinear alternative, with the drawback of having a more complex interpretation and more demanding numerical computations for obtaining the posterior distribution.

Bayesian inference for logistic regression

Assume that the response observations $y_1, \ldots, y_n$ are independent conditional on the covariates. As in the regression case, it is com-

mon to assume that the covariates are known. Define $\theta(\boldsymbol{\beta}, \mathbf{x}) = 1/(1+\exp(-\mathbf{x}^\top \boldsymbol{\beta}))$ as the success probability for an observation with covariate vector $\mathbf{x}$. The likelihood function for the logistic regression is then a product of Bernoulli distributions

$$p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} \theta(\boldsymbol{\beta}, \mathbf{x}_i)^{y_i} (1 - \theta(\boldsymbol{\beta}, \mathbf{x}_i))^{1-y_i}, \tag{8.8}$$

just like the likelihood for Bernoulli trials in Chapter Single-parameter models. The difference is that the success probabilities $\theta(\boldsymbol{\beta}, \mathbf{x}_i)$ are not constant here across observations, but instead vary with the covariates $\mathbf{x}_i$ in a way determined by the logistic regression model.

Now that we have seen the connection to the likelihood for the Bernoulli model, let us write the likelihood for the logistic regression more compactly as

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{y_i} \left(\frac{1}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{1-y_i} \tag{8.9}$$

$$= \exp\Big(\sum_{i=1}^{n} y_i \mathbf{x}_i^\top \boldsymbol{\beta} \Big) \prod_{i=1}^{n} \left(\frac{1}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right). \tag{8.10}$$

The posterior distribution of $\boldsymbol{\beta}$ is then

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) p(\boldsymbol{\beta}), \tag{8.11}$$

for some prior distribution $p(\boldsymbol{\beta})$. Assuming for example a multivariate normal prior $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, we obtain the posterior

$$\exp\Big(\sum_{i=1}^{n} y_i \mathbf{x}_i^\top \boldsymbol{\beta} \Big) \prod_{i=1}^{n} \left(\frac{1}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right) \exp\Big(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}) \Big). \tag{8.12}$$

Unfortunately, the posterior in (8.12) is not a multivariate normal distribution, or any other known distribution. The posterior distribution for the logistic regression is intractable.

A natural approach is to use the posterior approximation in Chapter Normal posterior approximation. Figure 8.3 gives a complete code for obtaining the normal posterior approximation in Figure 7.3 for the logistic regression model with a multivariate normal prior for $\boldsymbol{\beta}$. Note that the log-likelihood of the logistic regression can be written

$$\log(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}) = \sum_{i=1}^{n} y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^{n} \log\left(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\right).$$

The code in Figure 8.3 uses the maximum likelihood estimate as the initial value $\boldsymbol{\beta}_0$ for $\boldsymbol{\beta}$. This is a good choice for fast convergence of the iterative optimization algorithm. A more rough estimate, or even

```julia
# 0. Loading packages
using Plots, Distributions, GLM, LinearAlgebra, Optim, ForwardDiff

# 1. Setting up the log posterior function
"""
    πlogisticreg(β, y, X, μ, Σ)

log posterior for the logistic regression model
    Pr(y=1|x) = 1/(1 + exp(-x'β))
with the prior
    β ~ N(μ,Σ).
"""
function πlogisticreg(β, y, X, μ, Σ)
    loglik = sum( y.*(X*β)  .- log.(1 .+ exp.(X*β)) )
    logprior = logpdf(MvNormal(μ, Σ), β)
    return(loglik + logprior)
end

# 2. Generate data from logistic regression with β = [1,-1,1,-1]
n = 100
p = 4
X = [ones(n,1) randn(n,p-1)]
β = [1,-1,1,-1]
probs = 1 ./ (1 .+ exp.(-X*β))
y = rand.(Bernoulli.(probs))

# 3. Set up prior
μ = zeros(p)
Σ = 10*I(p)

# 4. Initial value for the optimization
glmfit = glm(X, y, Bernoulli(), LogitLink()) # find MLE.
β₀ = coef(glmfit) # initial values from MLE.

# 5. Run optimizer with automatic differentiation to find mode and Hessian.
optres = maximize(β -> πlogisticreg(β, y, X, μ, Σ), β₀, autodiff = :forward)
βmode = Optim.maximizer(optres)

# 6. Compute Hessian to get posterior covariance matrix approximation
H(β) = ForwardDiff.hessian(β -> πlogisticreg(β, y, X, μ, Σ), β)
Ωᵦ = Symmetric(-inv(H(βmode))) # This is J^{-1}

# 7. Simulate from normal posterior approximation and compute odds ratios
βsim = rand(MvNormal(βmode,Ωᵦ), 10000)'
oddsratio = exp.(βsim) # 10000 × 4 matrix with draws of exp(βⱼ) in jth column.
```

Figure 8.3: Numerical optimization to find normal posterior approximation for the logistic regression model in Julia. The broadcasting operator in Julia is denoted by the dot (.) so that exp.(x) is the vector that applies the exponential function to each element of the vector x. Similarly, `a .- b` is the elementwise difference of the two vectors `a` and `b`

setting $\beta_0 = 0$, is often sufficient for convergence. The exceptions are when the log posterior is a complex surface with multiple modes. Note that the gradient vector

$$\frac{\partial \log[p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})p(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}}$$

and Hessian matrix

$$\frac{\partial^2 \log[p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})p(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}$$

are computed by automatic differentiation in Figure 8.3 using the Julia package `ForwardDiff.jl`. The gradient and Hessian for the logistic regression model are actually quite simple to derive and can be given as explicit arguments to the `maximize` function. We nevertheless use automatic differentiation in the code to illustrate the generality of the normal approximation approach in Figure 8.3 also when the gradient and Hessian are much more tedious to derive. Note also that we only need to implement a function that computes the log-likelihood and log prior, the rest is handled by the software.

The last two lines of code in Figure 8.3 illustrates the important point that the normal approximation is easy to simulate from, and the generated posterior draws can be used to compute the posterior distribution of any transformation of the parameters, exactly as we did in Chapter Multi-parameter models. Similarly, the posterior draws can be used to simulate from the predictive distribution, as we did in Chapter Prediction and Decision making. Given that the posterior approximation is accurate enough, the normal approximation obtained by numerical optimization followed by posterior simulation from the approximate posterior is clearly a general and highly useful approach to Bayesian learning, prediction and decision making.

Applied logistic regression - who survived the Titanic?

On April 15, 1912, the RMS Titanic sank on her maiden voyage after colliding with an iceberg. The catastrophe resulted in the death of 1502 persons among the 2224 persons onboard. The `titanic` dataset is a subset of the **titanic dataset** in the Kaggle repository[1], with missing values imputed by regression models with the other variables as covariates. The dataset consists of 887 passengers out of which 342 persons survived. Several variables such as age, ticket class and number of relatives are available to explain the binary response `Survived`. Table 8.1 gives a summary of the data. We will here model the survival probability with a logistic regression on an intercept and the three covariates: `age`, `sex` and `class`.

titanic dataset

[1] https://www.kaggle.com/c/titanic/

Let the prior be $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$. To determine suitable values for $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, let us derive the implied prior for the more interpretable

Table 8.1: Summary of the titanic data.

variable	description	data type	values	comment
`survived`	survived	binary	$\{0,1\}$	`survived=1` for survived
`class`	ticket class	ordinal	$\{1,2,3\}$	`class=1` for first class
`sex`	sex	binary	$\{0,1\}$	`sex=1` for females
`age`	age	continuous	$[0,\infty]$	range = $[0.42, 80]$
`sibling/spouse`	number aboard	counts	$\{0,1,\ldots\}$	range = $[0,8]$
`parent/child`	number aboard	counts	$\{0,1,\ldots\}$	range = $[0,6]$
`fare`	fare in $	continuous	$[0,\infty]$	range = $[14.45, 512.33]$

survival odds $\Pr(y = 1|\mathbf{x})/\Pr(y = 0|\mathbf{x})$, which we have seen is $\exp(\mathbf{x}^\top\boldsymbol{\beta})$ in the logistic regression. If $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then $\mathbf{x}^\top\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{x}^\top\boldsymbol{\mu}, \mathbf{x}^\top\boldsymbol{\Omega}\mathbf{x})$, which means that $\exp(\mathbf{x}^\top\boldsymbol{\mu})$ follows the log-normal distribution

$$\exp(\mathbf{x}^\top\boldsymbol{\beta}) \sim \mathrm{LN}(\mathbf{x}^\top\boldsymbol{\mu}, \mathbf{x}^\top\boldsymbol{\Omega}\mathbf{x}).$$

This means in particular that the prior median for the survival odds is $\exp(\mathbf{x}^\top\boldsymbol{\mu})$. We will here set

$$\boldsymbol{\mu} = \big(-1, -1/80, 1, 1\big)^\top.$$

The prior mean for the intercept $\mu_1 = -1$ was chosen so that survival probability of $\Pr(y = 1|\mathbf{x} = \mathbf{0}) \approx 0.269$ was deemed reasonable for newborn (`age=0`) boy (`sex=0`) not traveling in first class (`class=0`). The prior mean for the coefficient on age $\mu_2 = -1/80$ implies that the survival odds decrease with a multiplicative factor of $\exp(-1/80) \approx 0.988$ for each year so that the survival odds of for example an 80-year-old is roughly a third of a newborn's odds ($\exp(-80/80) \approx 0.368$). The prior means for `sex` and `class` are set so that both of these factors increase the survival odds by a factor $\exp(1) \approx 2.718$.

It remains to determine the prior variance around the mean. We will use

$$\boldsymbol{\Omega} = \begin{pmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 1/(80^2) & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

which assumes prior independence between the elements in $\boldsymbol{\beta}$, for simplicity. The prior variances are chosen so that the implied prior distributions for the survival odds of some selected ages, the variable `sex` and the variable `class` agree with my prior beliefs, see Figure 8.4.

Figure 8.5 shows the marginal posteriors for the elements in $\boldsymbol{\beta}$ from the normal approximation as dark blue lines. Note that these marginal posteriors are for the survival odds $\exp(\boldsymbol{\beta}_j)$, which follow a log-normal distribution when the posterior for $\boldsymbol{\beta}$ is approximated by a multivariate normal distribution. The histograms in Figure 8.5 are from 100,000 posterior draws using the Hamiltonian Monte Carlo (HMC) method presented later in Chapter Markov Chain Monte

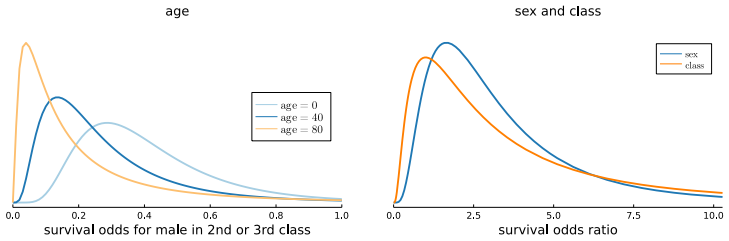


Figure 8.4: Implied prior distributions for the titanic data.

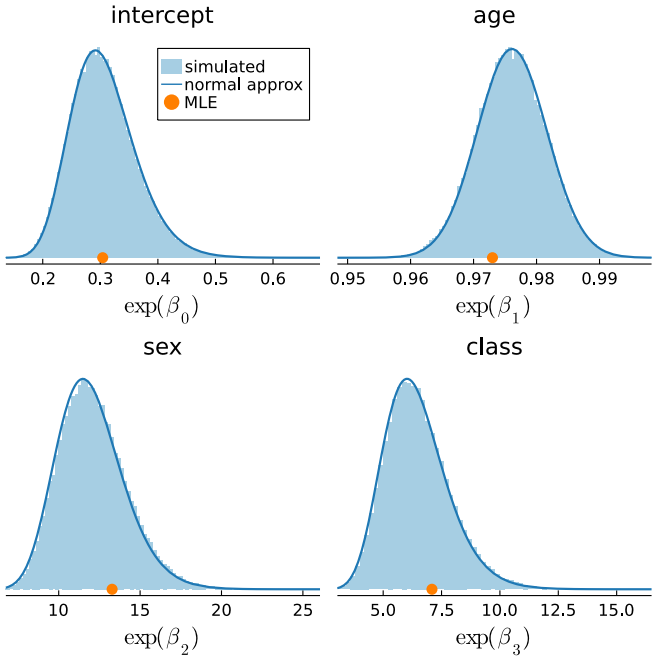


Figure 8.5: Marginal posterior distributions of the odds ratios from the normal approximate posterior for $\boldsymbol{\beta}$ compared to $100,000$ draws simulated from the true posterior using the Hamiltonian Monte Carlo method.

Carlo simulation. The histograms from HMC sampling can for practical purposes be taken to represent the exact posterior without approximation, and Figure 8.5 therefore shows that the normal approximation for $\boldsymbol{\beta}$, or equivalently, the log-normal approximation for $\exp(\beta_j)$ is extremely accurate here. The maximum likelihood (MLE) estimates are shown as reference.

To investigate how robust the posterior is to changes in the prior, Figure 8.6 compares the above results to those from a noninformative regularization prior $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, 10^2 I_p)$, where I_p is the $p \times p$ identity matrix; see Chapter Regularization for more on regularization priors. Since the dataset is only moderately large, the informative prior has some effect on the posterior, although not excessive.

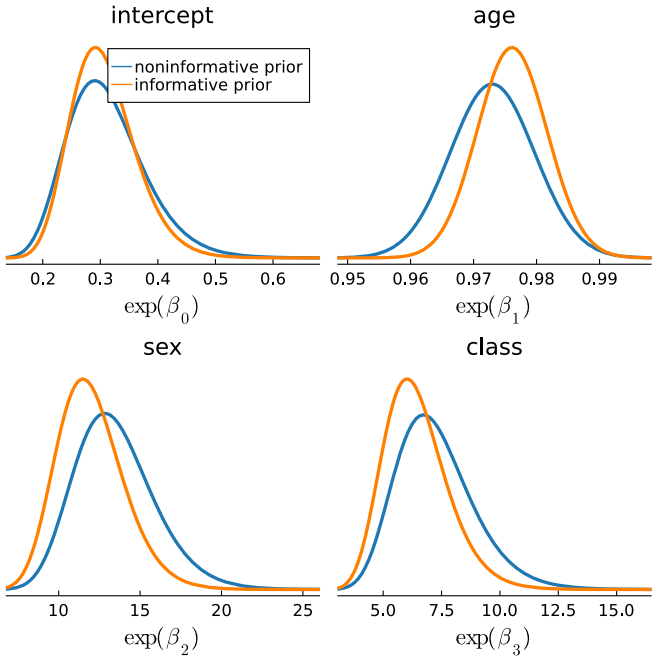


Figure 8.6: Marginal posterior distributions of the odds ratios from the normal approximate posterior for $\boldsymbol{\beta}$ using the informative prior. The marginal posteriors from the noninformative prior are given as a reference.

8.3 *Multi-class logistic regression*

The direct extension of the logistic regression model to the multi-class case is

$$\Pr(y = c|\mathbf{x}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_C) = \frac{\exp(\mathbf{x}^\top \boldsymbol{\beta}_c)}{\sum_{j=1}^{C} \exp(\mathbf{x}^\top \boldsymbol{\beta}_j)}, \tag{8.13}$$

where one immediately can see that $\sum_{c=1}^{C} \Pr(y = c|\mathbf{x}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_C) = 1$, as required. Note that each class has its own vector of regression coefficients, $\boldsymbol{\beta}_c, c = 1, \ldots, C$.

A problem with the model in (8.13) is that it is non-identified. A probabilistic model $p(\mathbf{x}|\boldsymbol{\theta})$ is said to be **non-identified** if there are multiple sets of parameter values $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q$ that imply identical probability distributions for the data. Non-identified models are problematic since the likelihood cannot discriminate between these different parameter values: for *any* dataset $\mathbf{x}$ from the data generating process we have $p(\mathbf{x}|\boldsymbol{\theta}_1) = \cdots = p(\mathbf{x}|\boldsymbol{\theta}_q)$. There can even be an infinite number of parameter values with identical $p(\mathbf{x}|\boldsymbol{\theta})$, for example all linear combinations $\mathbf{a}^\top\boldsymbol{\theta} = c$, for some vector $\mathbf{a}$ and constant c. The left graph in Figure 8.7 plots the log-likelihood function for the toy model $x_1, \ldots, x_n|\theta_1, \theta_2 \overset{\text{iid}}{\sim} \mathrm{N}(\theta_1 - \theta_2, 1)$, which is non-identified since for any pair (θ_1, θ_2) where $\theta_1 = \theta_2$ we obtain exactly the same probability distribution $\mathrm{N}(0, \sigma^2)$ for the data. Hence for *any* dataset $x_1, \ldots, x_n$ we have for example $p(x_1, \ldots, x_n|\theta_1 = 1, \theta_2 = 1) = p(x_1, \ldots, x_n|\theta_1 = 10, \theta_2 = 10)$.

non-identified

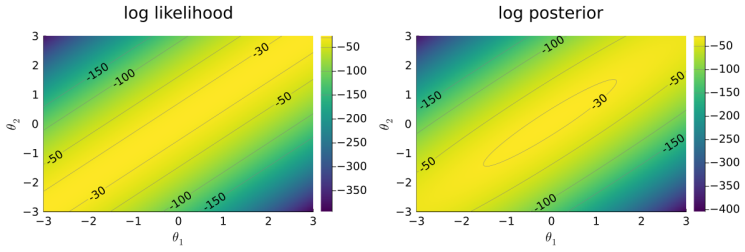


Figure 8.7: Illustrating non-identification in the model $x_1, \ldots, x_n|\theta_1, \theta_2 \overset{\text{iid}}{\sim} \mathrm{N}(\theta_1 - \theta_2, 1)$. The left graph plots the log-likelihood as a heatmap with overlayed contour lines. The log-likelihood attains the same value along each contour line. The log-likelihood cannot discriminate between the parameter combinations along a given line. The right graph shows how a prior "solves" the non-identification by combining the likelihood with a $\mathrm{N}(0, 1)$ prior for each parameter; the parameter combinations along the previous lines no longer have the same posterior density values.

The multi-class logistic regression in (8.13) is non-identified since adding a vector $\mathbf{a}$ to all $\boldsymbol{\beta}_c$ does not affect the class probabilities:

$$\Pr(y = c|\mathbf{x}, \boldsymbol{\beta}_1 + \mathbf{a}, \ldots, \boldsymbol{\beta}_C + \mathbf{a}) = \frac{\exp(\mathbf{x}^\top(\boldsymbol{\beta}_c + \mathbf{a}))}{\sum_{j=1}^{C} \exp(\mathbf{x}^\top(\boldsymbol{\beta}_j + \mathbf{a}))} \tag{8.14}$$

$$= \frac{\exp(\mathbf{x}^\top\boldsymbol{\beta}_c)}{\sum_{j=1}^{C} \exp(\mathbf{x}^\top\boldsymbol{\beta}_j)}, \tag{8.15}$$

as $\mathbf{a}$ cancels out in the numerator and denominator. Hence, the likelihood attains exactly the same value for any $\mathbf{a}$, and the model is non-identified. Luckily, the model can be identified by setting $\boldsymbol{\beta}_c = \mathbf{0}$ for one of the classes. This class is then referred to as the **reference class**; we will set the last class to zero, i.e. $\boldsymbol{\beta}_C = \mathbf{0}$. The reason why this restriction identifies the model is that it makes sure that only $\mathbf{a} = \mathbf{0}$ is allowed, since any non-zero $\mathbf{a}$ violates the restriction $\boldsymbol{\beta}_C = \mathbf{0}$. Looking back now to the binary logistic regression we can understand why there is only one $\boldsymbol{\beta}$ in that model, despite there being two classes: we implicitly set the negative class to the reference class with zero regression coefficients.

reference class

Many machine learning libraries do not use zero restrictions to identify the multi-class model, and instead use a regularization prior

to identify the model. This is illustrated in the right graph in Figure 8.7 where the likelihood function is combined with a $\mathrm{N}(0,1)$ prior for each parameter. The contour curves are no longer lines since the prior is now adding information that helps to discriminate between parameter value pairs with identical likelihoods; the prior can be said to identify the model. Using a prior to cover up an identification problem in the model is not a great idea, but can be useful when it is difficult to impose identifying restrictions.

The role and interpretation of the multi-class parameters can be understood by the log odds comparing the classes pairwise:

$$\mathrm{LogOdds}_{c,k}(\mathbf{x}) \equiv \log \frac{\Pr(y=c|\mathbf{x})}{\Pr(y=k|\mathbf{x})} = \mathbf{x}^\top(\boldsymbol{\beta}_c - \boldsymbol{\beta}_k). \tag{8.16}$$

The log odds between any pair of classes is hence linear in the difference of the classes' parameter vectors. An increase in a covariate x_j with a larger coefficient in class c compared to class k would therefore increase the probability for class c *compared to* class k. Setting $\boldsymbol{\beta}_C = 0$ for identification makes it particularly easy to interpret the $\boldsymbol{\beta}_c$ coefficients by comparing them against the reference class

$$\mathrm{LogOdds}_{c,C}(\mathbf{x}) \equiv \log \frac{\Pr(y=c|\mathbf{x})}{\Pr(y=C|\mathbf{x})} = \mathbf{x}^\top\boldsymbol{\beta}_c.$$

8.4 Poisson regression and generalized linear models

The normal posterior approximation technique can be directly applied to many other interesting regression and classification models. As an example, the **Poisson regression** model for the count data y_i conditional on a vector of covariates $\mathbf{x}_i$ Poisson regression

$$y_i|\mathbf{x}_i \overset{\text{indep}}{\sim} \mathrm{Pois}(\exp(\mathbf{x}_i^\top\boldsymbol{\beta})).$$

Note that the Poisson mean $\mathbb{E}(y|\mathbf{x}) = \exp(\mathbf{x}^\top\boldsymbol{\beta})$ is modeled via the exponential function to make sure that the mean is always positive, as required in the Poisson distribution. Since $\log \mathbb{E}(y|\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\beta}$, the model in (8.4) is said to have a log *link function*.

Assuming for example a multivariate normal prior $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \tau^2 I_p)$, the posterior distribution is

$$p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) \propto \prod_{i=1}^{n} \frac{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})^{y_i} e^{-\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}}{y!} \exp\Big(-\frac{1}{2\tau^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\Big), \tag{8.17}$$

which is of unknown distributional form. Moreover, there is no other known prior that would make the posterior for Poisson regression tractable. However, the normal posterior approximation is easy, all

we need to do is to replace the unnormalized posterior for the logistic regression in Figure 8.3 with the logarithm of (8.17).

TODO! apply poisson regression to ebay data.

Generalized linear models (GLM) provide a generalization of the logistic regression for binary data and the Poisson regression for count data to any distribution in the exponential family. One version of the GLM family is

$$\begin{aligned} y_i|\mathbf{x}_i &\overset{\text{indep}}{\sim} \text{ExpFamily}(\mu_i) \\ g(\mu_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \end{aligned} \tag{8.18}$$

where we use a slightly different notation for the exponential family with the conditional mean $\mu = \mathbb{E}(y|\mathbf{x})$ as the argument instead of the parameter θ. A key assumption in GLMs is that the conditional mean transformed by the link function $g()$ is assumed to be a linear function of the covariates. The linear combination $\mathbf{x}_i^\top \boldsymbol{\beta}$ is termed the **linear predictor** in the GLM literature. The logistic regression is a GLM with the Bernoulli distribution as the exponential family member and the log odds as link function, since for a binary variable $\mathbb{E}(y|\mathbf{x}) = \Pr(y = 1|\mathbf{x})$. Poisson regression is a GLM with the Poisson distribution and the log link. Gaussian linear regression is a GLM with a Gaussian distribution and the identity function $g(\mu) = \mu$ as a link. The posterior distribution for $\boldsymbol{\beta}$ in GLMs are almost always intractable (the Gaussian linear regression is an exception), but the normal approximation method is simple to implement. General expressions for the gradient and Hessian matrix for GLMs are available in many textbooks, but automatic differentiation is an otherwise attractive alternative.

linear predictor

The GLM model in (8.18) can be further extended by replacing the linear predictor $\mathbf{x}_i^\top \boldsymbol{\beta}$ with a nonlinear function of the covariates. Polynomials or splines (see Regularization) are useful here, with the Gaussian processes (Gaussian processes) as an interesting flexible alternative. Functions that are nonlinear in both the covariates $\mathbf{x}$ and the parameters $\boldsymbol{\beta}$ can also be used for further expressiveness in the mean, e,g. deep neural networks. There is in principle nothing that stops us from using a normal posterior approximation for such models, but the approximation may be inaccurate and numerically costly in models with highly non-Gaussian high-dimensional posteriors.

8.5 *Bayesian discriminant analysis and Naive Bayes*

TODO! This section is very incomplete.

Logistic regression is a so called **discriminative model** where the class probabilities $\Pr(Y = y|\mathbf{x})$ are directly modelled using the

discriminative model

logistic function. This is in contrast to a **generative model** where the class probabilties are modeled more implicitly using Bayes' theorem

generative model

$$\Pr(Y = y|\mathbf{x}) \propto \Pr(\mathbf{x}|Y = y) \cdot \Pr(Y = y). \tag{8.19}$$

In generative models we obtain the aim $\Pr(Y = y|\mathbf{x})$ via modeling of the prior probability of the class $\Pr(Y = y)$ and by explicitly modeling the distribution of the covariates $\mathbf{x}$ for each class, $\Pr(\mathbf{x}|Y = y)$. An example of an generative classification model is Bayesian discriminant analysis presented in section Bayesian discriminant analysis and Naive Bayes, where also the two modelling approaches will be contrasted and further discussed.

By Bayes' theorem we have

$$\Pr(Y = y|\mathbf{x}) \propto \Pr(\mathbf{x}|Y = y) \cdot \Pr(Y = y), \tag{8.20}$$

where $\Pr(Y = y)$ is the prior probability of the class and $\Pr(\mathbf{x}|Y = y)$ is the **class-conditional distribution** of the covariates $\mathbf{x}$. The prior probability is usually relatively simple to determine, it can for example be computed as the fraction of observations in class c in the data, or by the Bayesian analysis of Bernoulli data with a Beta prior presented in section Bernoulli data. The class-conditional distributions $\Pr(\mathbf{x}|Y = y)$ for $y = 0$ and $y = 1$ are usually more difficult as they are the joint distributions of all covariates $\mathbf{x}$ for each of the two classes.

class-conditional distribution

If all covariates are continuous with values over the whole real line, perhaps after suitable transformations, a natural first model is a multivariate Gaussian model for each class

$$\mathbf{x}|y, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \quad y \in \{0, 1\}. \tag{8.21}$$

The parameters of the negative class $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ can be estimated from all the covariate observations in the negative class in the dataset, and the parameters of the positive class, $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$, can similarly be learned from the positive cases. This Gaussian model gives rise to the popular Quadratic Discriminant Analysis (QDA) procedure. If we restrict the two classes to have the same covariance matrix $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$ we obtain Linear Discriminant Analysis (LDA); see (Lindholm et al., 2022) for details and applications.

Defining $\omega_0 = \Pr(Y = 0)$ and $\omega_1 = \Pr(Y = 1)$, the unknown model parameters are $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$, ω_0 and ω_1. The predictive distribution is as usual obtained by integrating out the parameters with respect to the posterior distribution $p(\boldsymbol{\mu}_{\tilde{y}}, \boldsymbol{\Sigma}_{\tilde{y}}, \omega_{\tilde{y}}|\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ and $\mathbf{X}$ is the training data. Formally, we write

$$p(\tilde{y}|\tilde{\mathbf{x}}) \propto \int N(\tilde{\mathbf{x}}|\boldsymbol{\mu}_{\tilde{y}}, \boldsymbol{\Sigma}_{\tilde{y}}) \cdot \omega_{\tilde{y}} \cdot p(\boldsymbol{\mu}_{\tilde{y}}, \boldsymbol{\Sigma}_{\tilde{y}}, \omega_{\tilde{y}}|\mathbf{y}, \mathbf{X}) d\boldsymbol{\mu}_{\tilde{y}} d\boldsymbol{\Sigma}_{\tilde{y}} d\omega_{\tilde{y}}, \tag{8.22}$$

for $\tilde{y} = 0$ and $\tilde{y} = 1$. Here $N(\mathbf{x}|\mu, \Sigma)$ denotes the multivariate normal density function with mean μ and covariance matrix Σ evaluated at $\mathbf{x}$.

1. Give expression for $p(\tilde{y}|\tilde{\mathbf{x}})$.
2. Apply it to Palmer penguins data with Bill length and Flipper length as covariates. Plot decision boundaries.
3. Naive Bayes.

EXERCISES

1. Problem!
2. Next problem!

NOTEBOOKS

1. See the notebook Classification.

9 Gibbs sampling

Gibbs sampling is a general iterative method for simulating from complex multivariate distributions. It can in principle be applied to any multivariate distribution, not necessarily a Bayesian posterior distribution, but we will here consider the Bayesian application where the aim is to sample from a joint posterior distribution $p(\theta_1, \ldots, \theta_p|\mathbf{y})$.

Part of the appeal of Gibbs sampling is that we can often augment the data with additional auxiliary variables in a way that makes Gibbs sampling very easy to implement in a robust fashion. We will see several examples of this **data augmentation** approach in this chapter, for example when sampling from the posterior of the very useful class of mixture models, and also when we design a sampling algorithm for the probit regression model for binary response data.

9.1 The Gibbs sampling algorithm

Gibbs sampling simulates from a multivariate distribution by iterative simulating each parameter from its so called full conditional posterior distribution. The **full conditional posterior** for the parameter θ_j is

full conditional posterior

$$p(\theta_j|\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p, \mathbf{y}),$$

where we note that we condition on *all* other parameters except θ_j; this is the meaning of the word *full* conditional posterior. A common notation for the full conditional posterior is therefore $p(\theta_j|\boldsymbol{\theta}_{-j}, \mathbf{y})$ where $\boldsymbol{\theta}_{-j}$ is the vector of all parameters except θ_j.

Starting from a set of initial values $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}$, Gibbs sampling iterates over the parameters, simulating a new draw from the parameter's full conditional posterior, conditioned on the most recent draw available for the other parameters. The algorithm is illustrated in Algorithm 9.1 where parameters highlighted in orange indicates that the parameter has been updated in the current iteration of the algorithm.

Gibbs sampling is a member of the family of Markov Chain Monte

Gibbs sampling

Input: initial values $\theta_2^{(0)}, \ldots, \theta_p^{(0)}$
number of posterior draws m.
for i in $1{:}m$ **do**

$$\theta_1 \sim p\left(\theta_1 \mid \theta_2^{(i-1)}, \theta_3^{(i-1)}, \ldots, \theta_p^{(i-1)}, \mathbf{y}\right)$$
$$\theta_2 \sim p\left(\theta_2 \mid \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_p^{(i-1)}, \mathbf{y}\right)$$
$$\vdots$$
$$\theta_p \sim p\left(\theta_p \mid \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_p^{(i)}, \mathbf{y}\right)$$

end
Output: m autocorrelated draws for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ that converge in distribution to the joint posterior $p(\theta_1, \ldots, \theta_p|\mathbf{y})$.

Figure 9.1: Gibbs sampling algorithm for sampling from a joint posterior distribution $p(\theta_1, \ldots, \theta_p|\mathbf{y})$. Parameters highlighted in orange indicates that the parameter has been updated in the current iteration of the algorithm.

Carlo (MCMC) algorithms, where Markov Chains are used to simulate from multivariate distributions. We will explain MCMC more fully in the next chapter; for the moment, five related things about MCMC algorithms are important for understanding Gibbs sampling:

- samples from an MCMC algorithm are *autocorrelated* over the iterations, meaning that succesive draws of $\boldsymbol{\theta}$ are dependent on each other.
- simulations from MCMC can nevertheless be shown to *converge in distribution* to the target posterior distribution

$$\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)} \xrightarrow{d} p(\boldsymbol{\theta}|\mathbf{y}) \text{ as } m \to \infty$$

- the convergence to the target posterior happens *for any initial values* used to start up the sampling algorithm.
- a version of the central limit theorem for dependent variables can be used to establish that the sample mean of the draws can be well approximated by a normal distribution if we sample long enough. Informally, where $\bar{\boldsymbol{\theta}}_{1:m}$ is the sample mean of m MCMC draws:

$$\bar{\boldsymbol{\theta}}_{1:m} \overset{\text{approx}}{\sim} \mathrm{N}\left(\mathbb{E}(\boldsymbol{\theta}|\mathbf{y}), \frac{\mathbb{V}(\boldsymbol{\theta}|\mathbf{y})}{m}\right) \text{ for large } m$$

- MCMC sampling tends to be *less efficient* than iid sampling from the joint posterior which means that we have to sample more

draws to obtain the same accuracy in approximation the posterior with the samples.

As an example consider simulating from a bivariate normal distribution $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where ρ is the correlation between θ_1 and θ_2. In this case there is really no need for Gibbs sampling at all since a p-dimensional multivariate normal distribution $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by simulating a vector $z = (z_1, \ldots, z_p)^\top$ and with independent standard univariate normal variables and setting

$$\boldsymbol{\theta} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}, \tag{9.1}$$

where $\mathbf{L}$ is the $p \times p$ lower triangular Cholesky factor in the matrix decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$; see Appendix A.1 for details. We will nevertheless show how Gibbs sampling is implemented for the bivariate normal distribution as an illustration and compare its efficiency to direct iid sampling. Figure 9.2 gives the Gibbs sampling algorithm for a bivariate normal distribution target in pseudo code, and Figure 9.3 provides a Julia implementation of the general case with a multivariate normal distribution target.

Gibbs sampling from a bivariate normal

Input: initial value $\theta_2^{(0)}$
number of posterior draws m.

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

$$\theta_1^{(i)} \mid \theta_2 \sim \mathrm{N}\left(\mu_1 + \rho\tfrac{\sigma_1}{\sigma_2}(\theta_2^{(i-1)} - \mu_2),\, \sigma_1^2(1-\rho)^2\right)$$

$$\theta_2^{(i)} \mid \theta_1 \sim \mathrm{N}\left(\mu_2 + \rho\tfrac{\sigma_2}{\sigma_1}(\theta_1^{(i)} - \mu_1),\, \sigma_2^2(1-\rho)^2\right)$$

end

Output: m autocorrelated draws for $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$ that converge in distribution to the bivariate normal distribution $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Figure 9.2: Gibbs sampling algorithm for sampling from the bivariate normal distribution of $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$, i.e. $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Parameters highlighted in orange indicates that the parameter has been updated in the current iteration of the algorithm.

Figure 9.4 plots the draws (points) from a bivariate normal target distribution (contours) for different correlations in the target distribution. Both iid sampling (left column) and Gibbs sampling (right

```
using Distributions, LinearAlgebra, InvertedIndices

function GibbsMvNormal(μ, Σ, m, θ₀)
    # Pre-computing stuff what doesn't change in the Gibbs sampler iterations.
    p = length(μ)
    σ = zeros(p)        # Conditional variances
    β = zeros(p-1,p)    # "Regression coefficients" for reg on other coordinates
    for j = 1:p
        β[:,j] = Σ[Not(j),Not(j)]\Σ[Not(j),j]
        σ[j] = √(Σ[j,j]-Σ[j,Not(j)]⋅β[:,j])
    end

    # Gibbs sampling iterations
    postDraws = zeros(m,p)
    θ = θ₀
    for i = 1:m
        for j = 1:p
            θ[j] = rand(Normal(μ[j] + β[:,j]⋅(θ[Not(j)]-μ[Not(j)]), σ[j]))
            postDraws[i,j] = θ[j]
        end
    end
    return postDraws
end

julia> gibbsDraws = GibbsMvNormal(μ = zeros(2), Σ = [1 0.7;0.7 1], m = 1000, θ₀ = zeros(2));
julia> cov(gibbsDraws)
2×2 Matrix{Float64}:
 1.00295   0.694255
 0.694255  0.998105
```

Figure 9.3: Gibbs sampling for a multivariate normal target distribution in Julia, including an example call of the function at the end. The `Not` function from the `InvertedIndices.jl` package selects all indices except the one in the argument, and $\cdot$ is the usual (dot) vector product.

column) are shown. It is clear that Gibbs sampling's coordinate-wise nature makes it explore the target distribution very slowly when the parameters are highly correlated. This is particularly clear in the case with $\rho = 0.99$ where the target distribution is strongly 'cigar shaped' and it takes a long time for Gibbs sampling to travel across the cigar. In contrast, iid sampling can of course freely move from one end of the cigar to the other from one iteration to the next since there is no dependence on the previous draw.

As mentioned above, the major disadvantage of Gibbs sampling (and MCMC more generally) is the draws are autocorrelated. This leads to less efficient estimates of, for example, the posterior mean and standard deviation, or more generally the whole posterior distribution. One way to quantify this loss of efficiency compared to iid sampling is to compare the variance of the sample mean $\bar{\theta}_{1:m}$ of the simulated draws, as a frequentist estimator of the posterior mean $\mathbb{E}(\theta|\mathbf{y})$. For iid sampling we immediately obtain the usual variance formula for sample mean:

$$\mathbb{V}_{\text{iid}}(\bar{\theta}_{1:m}) = \frac{\mathbb{V}(\theta|\mathbf{y})}{m}, \tag{9.2}$$

When the draws are autocorrelated the variance can for large m be approximated by:

$$\mathbb{V}_{\text{mcmc}}(\bar{\theta}_{1:m}) = \frac{\mathbb{V}(\theta|\mathbf{y})}{m}\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right), \tag{9.3}$$

where $\rho_k = \text{Corr}\left(\theta^{(i)}, \theta^{(i-k)}\right)$, i.e. the autocorrelation coefficient at lag k. The approximation is motivated by the fact that for a stationary

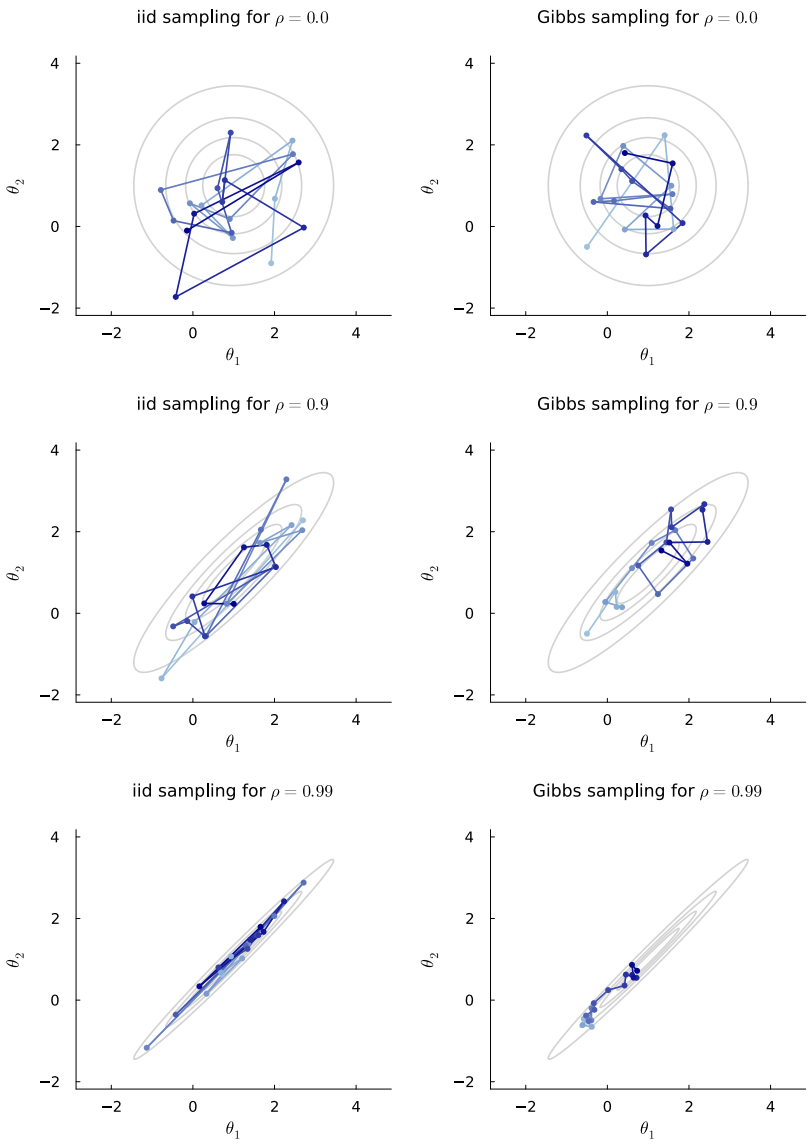


Figure 9.4: Comparing sampling paths of iid sampling (left column) vs Gibbs sampling (right column) for a bivariate normal target distribution in different correlations ρ (rows). The sampling path starts in the lightest blue and ends in the darkest blue.

process one can show that (Lindgren, 2012)

$$m \cdot \mathbb{V}_{\text{mcmc}}(\bar{\theta}_{1:m}) \xrightarrow{p} \mathbb{V}(\theta|\mathbf{y})\Big(1 + 2\sum_{k=1}^{\infty} \rho_k\Big) \text{ as } m \to \infty. \tag{9.4}$$

The loss of efficiency from having to use dependent sampling instead of iid sampling can now be quantified by the **inefficiency factor** (IF)

inefficiency factor

$$\text{IF} = \frac{\mathbb{V}_{\text{mcmc}}(\bar{\theta}_{1:m})}{\mathbb{V}_{\text{iid}}(\bar{\theta}_{1:m})} = 1 + 2\sum_{k=1}^{\infty} \rho_k. \tag{9.5}$$

The inefficiency factor measures how many times more draws are needed to achieve the same level of accuracy as iid sampling. For example, if the inefficiency factor is IF $= 10$, then we need ten times as many draws to achieve the same level of accuracy (sampling variance) as iid sampling. This allows us to define the **effective sample size** m_{eff} of a sampling algorithm that produces dependent draws as

effective sample size

$$m_{\text{eff}} = \frac{m}{\text{IF}}. \tag{9.6}$$

The effective sample size m_{eff} therefore represents the number of iid draws that would have the same sampling variance as the m dependent draws. A nominal sample size of $m = 10000$ draws from a sampling algorithm with IF $= 10$ is therefore equivalent to a sample with $m_{\text{eff}} = 1000$ iid draws.

Figure 9.5 uses simulation to explore the inefficiency factor (IF) for Gibbs sampling for a multivariate normal target density as the dimension of the multivariate normal increases. The covariance matrix of this target density is set to an equicorrelation matrix with correlation ρ and unit variance for all variables:

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}. \tag{9.7}$$

As expected the inefficiency factor is sharply increasing with the correlation coefficient for the larger ρ, and the inefficiency of Gibbs sampling is particularly severe in higher dimensions. For a 10-dimensional multivariate normal target density with $\rho = 0.8$ the inefficiency factor is around IF $= 30$ in the figure; hence we would need as much as 30 times as many draws to achieve the same level of accuracy as iid sampling.

Note that if the draws are negatively autocorrelated, i.e. $\rho_k < 0$, then we can have $m_{\text{eff}} > m$, meaning that the dependent draws are *more* efficient than iid sampling. This makes sense if one considers that negative autocorrelation means by definition that one draw

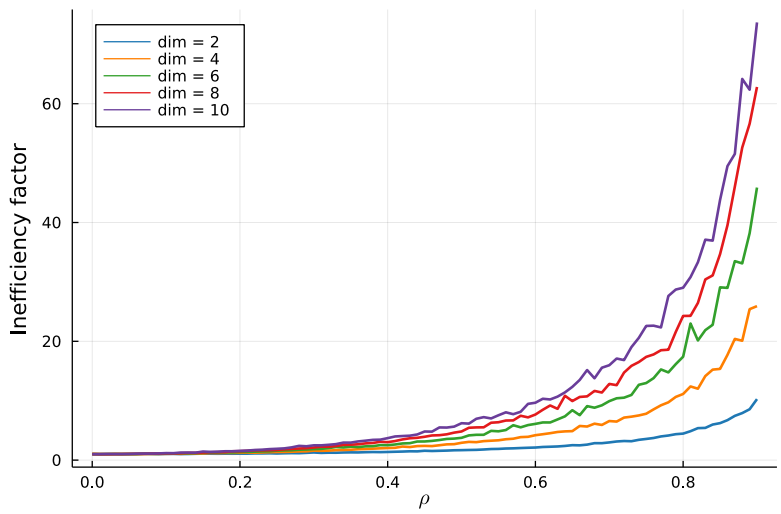


Figure 9.5: Inefficiency of Gibbs sampling for a multivariate normal target distribution in different dimensions with an equicorrelation matrix with correlation ρ.

above the mean tends to be followed by a draw below the mean. The average of draws that tends to alternate between above and below the mean will therefore be a very good estimator of the mean. This is referred to as antithetic sampling in the Monte Carlo integration literature. However, with Gibbs sampling and MCMC we typically get positive autocorrelation, and the effective sample size from Gibbs sampling and MCMC is therefore nearly always smaller than the nominal sample size.

We have now seen that correlation between parameters in the posterior can be devasting for the efficiency of Gibbs sampling. When possible one can consider **reparametrizing** the model parameters to reduce the posterior correlation. In the multivariate normal distribution one can rotate the coordinate system to the principal axes (see Appendix A.1) to acheive parameters that exactly uncorrelated. However, in more serious examples it can be difficult to find the appropriate reparametrization.

A more common technique to deal with the inefficiency of Gibbs sampling is to group correlated parameters in a so called **block Gibbs sampler**, and to sample the group/block jointly from its multivariate full conditional posterior. For example, consider a posterior with three parameters $p(\theta_1, \theta_2, \theta_3|\mathbf{y})$, where θ_1 and θ_2 are correlated and θ_3 is uncorrelated with the other two. We can then sample from the full conditional posterior by the following two-block Gibbs sampler:

block Gibbs sampler

$$\text{Block 1: } (\theta_1, \theta_2) \sim p(\theta_1, \theta_2|\theta_3, \mathbf{y})$$
$$\text{Block 2: } \theta_3 \sim p(\theta_3|\theta_1, \theta_2, \mathbf{y})$$

We are here sampling the two correlated parameters jointly and the

algorithm will be efficient since the two correlated dimensions we are traversing the cigar quickly in an iid fashion. This blocking technique can be applied in the same way with more than two blocks, and also with more than two parameters in a given block. However, we need to be able to sample from the full conditional posteriors of all blocks of parameters. While it is often that all univariate full conditional posteriors $p(\theta_1|\theta_2,\theta_3,\mathbf{y})$, $p(\theta_2|\theta_1,\theta_3,\mathbf{y})$ and $p(\theta_3|\theta_1,\theta_2,\mathbf{y})$ belong to easily sampled distributional families, the bivariate $p(\theta_1,\theta_2|\theta_3,\mathbf{y})$ may not be a recognizable, easily sampled, distribution. So, the reciepe for success is to group together any parameters that are highly correlated *and* for which we can sample from the full conditional posterior. We will see an example of this in the next section.

9.2 *Autoregressive processes*

We will here develop a Gibbs sampling algorithm for the posterior distribution $p(\mu,\phi_1,\ldots,\phi_p,\sigma|\mathbf{y}_{1:T})$ of the parameters in the **autoregressive model** of order p

autoregressive model

$$y_t = \mu + \phi_1(y_{t-1}-\mu) + \ldots + \phi_p(y_{t-p}-\mu) + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathrm{N}(0,\sigma^2), \tag{9.8}$$

where y_{t-k} is the kth **lagged value** of time series.

lagged value

Recall that the prior proposed in Chapter Priors is of the form

$$\begin{aligned} \mu &\sim \mathrm{N}(\mu_0,\tau_\mu^2), \\ \boldsymbol{\phi}|\sigma^2 &\sim \mathrm{N}\left(\boldsymbol{\mu}_0,\sigma^2\boldsymbol{\Omega}^{-1}\right) \\ \sigma^2 &\sim \mathrm{Inv}{-}\chi^2(\nu_0,\sigma_0^2), \end{aligned} \tag{9.9}$$

where $\Omega_0^{-1} = \mathrm{Diag}(\tau^2,\tau^2/2^2,\ldots,\tau^2/p^2)$ is a diagonal matrix with diagonal elements that decay with the lag length to encourage more shrinkage toward zero on the ϕ_k for longer lags. Note that the prior on $\boldsymbol{\phi}$ and σ is assumed to be independent of μ, and that the joint prior $p(\boldsymbol{\phi},\sigma)$ is exactly the conjugate prior for the linear Gaussian regression model in Chapter Regression.

The main complication with deriving the joint posterior distribution $p(\mu,\phi_1,\ldots,\phi_p,\sigma|\mathbf{y}_{1:T})$ in closed form is that the likelihood involves products of parameter pairs $\phi_k\mu$ for $k=1,\ldots,p$, and products of random variables are usually complicated. However, here is where Gibbs sampling comes to the rescue. Recall that Gibbs sampling only needs tractable posterior distributions for each parameter *conditional* on the other parameters. For example, once we condition on μ, the posterior for each ϕ_k and also for σ would be well known distributions, and it turns out that also the posterior for μ is a well known distribution, once we condition on all other model parame-

ters. Let us now derive the full conditional posterior distributions needed for a Gibbs sampling algorithm.

Conditional on μ let us rewrite the model in (9.8) as a homoscedastic Gaussian linear regression without intercept

$$\tilde{y}_t = \tilde{\mathbf{x}}_t^\top \boldsymbol{\phi} + \epsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2), \tag{9.10}$$

where $\tilde{y}_t = y_t - \mu$ and $\tilde{\mathbf{x}}_t = \big(y_{t-1} - \mu, \ldots, y_{t-p} - \mu\big)^\top$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$.

Given that the autoregressive model can be expressed as the linear regression in (9.10) conditional on μ, and the form of the prior in it is immediately clear that, once we condition on μ, we can sample from the joint posterior of $\boldsymbol{\phi}$ and σ^2 by using Figure 6.4 in Chapter Regression (where $\boldsymbol{\phi}$ now plays the role of the vector of regression coefficients $\boldsymbol{\beta}$).

The remaining question is then what the full conditional posterior $p(\mu|\boldsymbol{\phi}, \sigma^2, \mathbf{y})$ looks like. To derive that, note that the model in 9.8 can be rewritten by moving all the $\phi_k y_{t-k}$ terms to the left hand side to obtain

$$y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} = \mu(1 - \phi_1 - \ldots - \phi_p) + \varepsilon_t. \tag{9.11}$$

So by diving both sides by $1 - \phi_1 - \ldots - \phi_p$ we obtain

$$\check{y}_t = \mu + \tilde{\varepsilon}_t, \quad \tilde{\varepsilon}_t \overset{iid}{\sim} \text{N}(0, \tilde{\sigma}^2), \tag{9.12}$$

where

$$\check{y}_t = \frac{y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p}}{1 - \phi_1 - \ldots - \phi_p}, \tag{9.13}$$

and $\tilde{\sigma}^2 = \sigma^2/(1 - \phi_1 - \ldots - \phi_p)^2$. The model in (9.12) says that, conditional on $\boldsymbol{\phi}$ and σ^2, the transformed data $\check{\mathbf{y}}_{1:T}$ follows an iid Gaussian model with mean μ

$$\check{y}_t \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2). \tag{9.14}$$

Since we condition on σ^2 the conditional posterior of μ is then just the posterior for a mean in an iid Gaussian model with known variance, something we obtained already back in Chapter Single-parameter models

$$\mu|\mathbf{y}, \boldsymbol{\phi}, \sigma^2 \sim N(\mu_T, \tau_T^2),$$

where $\tau_T^{-2} = \frac{T}{\sigma^2} + \tau_\mu^{-2}$, $\mu_T = w\bar{\check{y}} + (1 - w)\mu_0$, $w = \frac{T}{\sigma^2}/(\frac{T}{\sigma^2} + \tau_\mu^{-2})$ and the little messy symbol $\bar{\check{y}}$ is the sample mean of the transformed response data $\check{\mathbf{y}}_{1:T}$. Note that the full conditional posterior distribution for μ is conditional on $\boldsymbol{\phi}$ and σ^2 via the transformed data $\check{\mathbf{y}}_{1:T}$, so the transformed data need to be re-computed every time a new draw of $\boldsymbol{\phi}$ and σ^2 is obtained.

Gibbs sampling for AR processes

Input: initial value $\mu^{(0)}$
number of posterior draws m.
for i in $1{:}m$ **do**
  for t in $1{:}T$ **do**
    $\tilde{y}_t = y_t - \mu^{(i-1)}$ and
    $\tilde{\mathbf{x}}_t = \left(y_{t-1} - \mu^{(i-1)}, \ldots, y_{t-p} - \mu^{(i-1)}\right)^\top$
  end
  Set up $\tilde{\mathbf{X}} = \left(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_T\right)^\top$ and $\tilde{\mathbf{y}} = \left(\tilde{y}_1, \ldots, \tilde{y}_T\right)^\top$
  Sample $(\sigma^2)^{(i)} \mid \mu^{(i-1)}, \tilde{\mathbf{y}}, \tilde{\mathbf{X}} \sim \text{Inv}{-}\chi^2\left(\nu_T, \sigma_T^2\right)$
  Sample $\boldsymbol{\phi}^{(i)} \mid (\sigma^2)^{(i)}, \mu^{(i-1)}, \tilde{\mathbf{y}}, \tilde{\mathbf{X}} \sim \text{N}\left(\boldsymbol{\mu}_\phi, \sigma^2 \boldsymbol{\Omega}_\phi^{-1}\right)$
  for t in $1{:}T$ **do**
    $\check{y}_t = \frac{y_t - \phi_1^{(i)} y_{t-1} - \ldots - \phi_p^{(i)} y_{t-p}}{1 - \phi_1^{(i)} - \ldots - \phi_p^{(i)}}$
  end
  Sample $\mu^{(i)} \mid \boldsymbol{\phi}^{(i)}, (\sigma^2)^{(i)}, \check{\mathbf{y}} \sim \text{N}\left(\mu_T, \tau_T^2\right)$
end
Output: m autocorrelated draws from the joint posterior $p(\boldsymbol{\phi}, \sigma, \mu | \mathbf{y}_{1:T})$.

Figure 9.6: Pseudo code for Gibbs sampling from the joint posterior distribution $p(\boldsymbol{\phi}, \sigma, \mu | \mathbf{y}_{1:T})$ in an autoregressive model. The two loops over the data points would be replaced by fast vectorized operations in a real implementation in a non-compiled language.

10 Markov Chain Monte Carlo simulation

10.1 Markov Chain Monte Carlo

10.2 Hamiltonian Monte Carlo

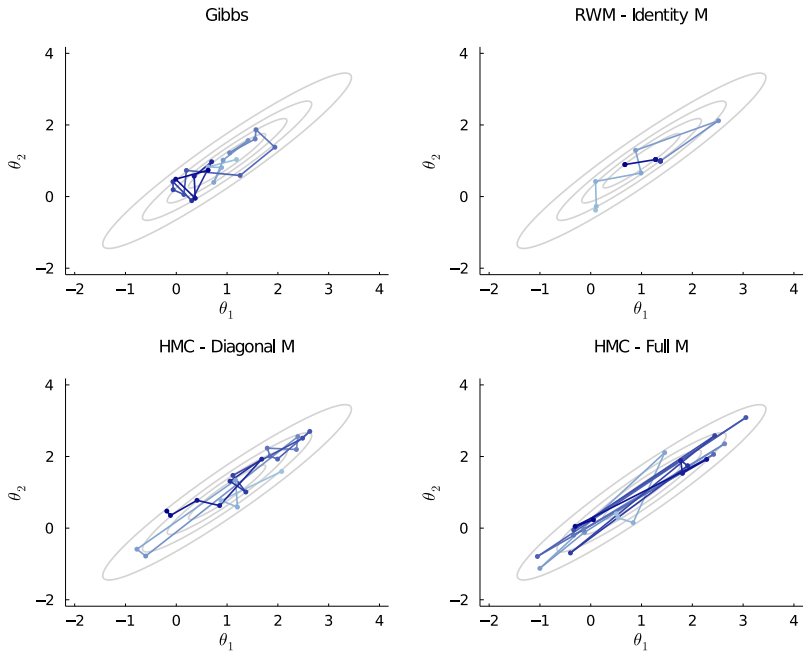


Figure 10.1: Comparing simulation paths of four algorithms for sampling from a bivariate normal target with $\boldsymbol{\mu} = (1,1)^\top$, unit variances and correlation $\rho = 0.95$. The four compared algorithms are: i) Gibbs sampling, ii) random walk Metropolis with identity scaling, iii) HMC-NUTS with diagonal mass matrix and iv) HMC-NUTS with full mass matrix.

10.3 Probabilistic programming frameworks

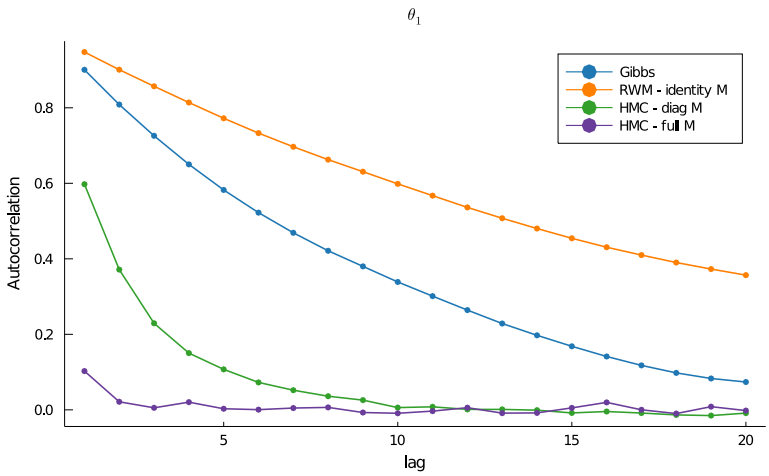


Figure 10.2: Comparing autocorrelation functions from four posterior sampling algorithms for sampling from a bivariate normal target with $\boldsymbol{\mu} = (1,1)^\top$, unit variances and correlation $\rho = 0.95$. The four compared algorithms are: i) Gibbs sampling, ii) random walk Metropolis with identity scaling, iii) HMC-NUTS with diagonal mass matrix and iv) HMC-NUTS with full mass matrix.

```
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 10.3: Turing.jl code for the iid Bernoulli model with a Beta prior.

```
using Turing, StatsPlots, Random
ScaledInverseChiSq(ν,τ^2) = InverseGamma(ν/2,ν*τ^2/2) # Inv-χ^2 distribution

# Setting up the Turing model:
@model function iidnormal(x, μ_0, κ_0, ν_0, σ^2_0)
    σ^2 ~ ScaledInverseChiSq(ν_0, σ^2_0)
    θ ~ Normal(μ_0,σ^2/κ_0)  # prior
    n = length(x)  # number of observations
    for i in 1:n
        x[i] ~ Normal(θ, √σ^2) # model
    end
end

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

# Set up the prior
μ_0 = 20; κ_0 = 1; ν_0 = 5; σ^2_0 = 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, μ_0, κ_0, ν_0, σ^2_0), NUTS(α), niter,
  discard_initial = nburn)

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

Figure 10.4: Turing.jl 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)
```

Figure 10.5: Rstan code for the iid normal model with a conjugate prior.

11 *Variational inference*

12 Regularization

12.1 Model complexity and overfitting

In statistical analyses one typically has to decide on the degree of complexity of the statistical model, particularly when the focus is on prediction and decision making. Models with a small number of parameters, for example linear regression and classification models, have the advantage that they are simple to interpret and they run less risk of **overfitting** the data. A model overfits the data when the fitted model is more complex than the underlying data generating process. An overfitted model looses track of the general tendencies in the data and tries too hard to capture individual observations, and will there generalize poorly to new data points that were not included in the fitting.

overfitting

To see the effects of overfitting, let us consider fitting regression models to the `mtcars` data from R. We will use miles per gallon `mpg` as the response and the horsepower `hp` of the cars as a single explanatory variable. To have a numerically stable solution we scale the `hp` by dividing by 100 so that `hp` measures hundreds of horsepowers. Figure X plots the data and the fit from a linear regression model. The linear model is clearly too simple to capture the non-linear relationship between `mpg` and `hp`; the linear model is **underfitting** the data.

underfitting

A Gaussian polynomial regression model

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

is fitted to the `mtcars` data from R. We will use miles per gallon `mpg` as the response and the horsepower `hp` of the cars as a single explanatory variable. To have a numerically stable solution we scale the `hp` by dividing by 100 so that `hp` measures hundreds of horsepowers. Figure 12.3 plots the data and the fit from a linear regression model.

TODO: Discuss alternative scenario with p»n, linear model but many variables.

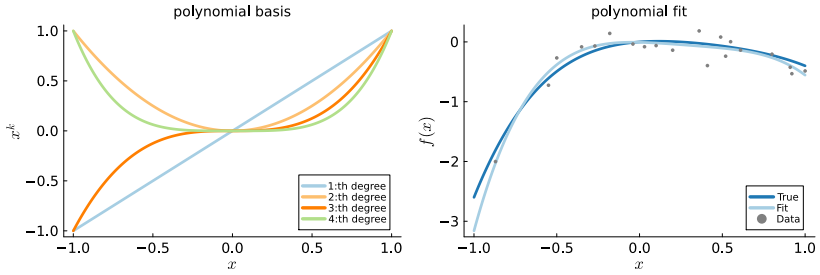


Figure 12.1: Polynomial regression.

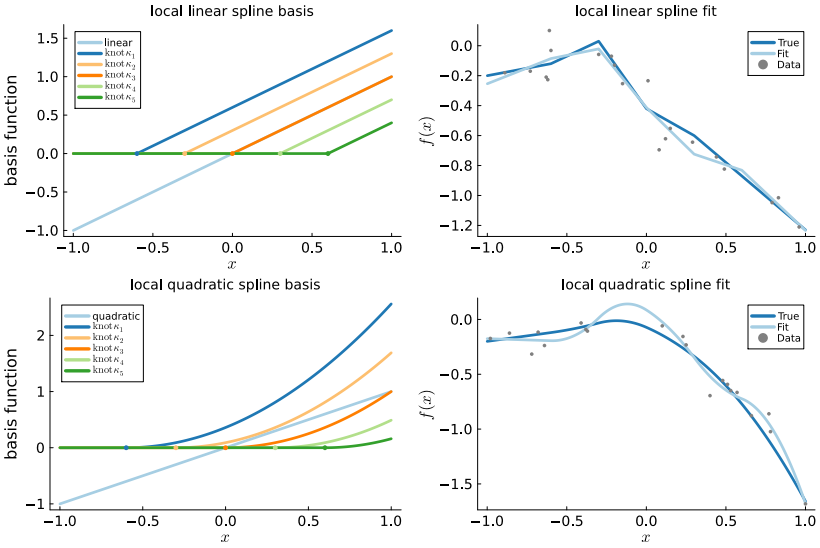


Figure 12.2: Spline regression.

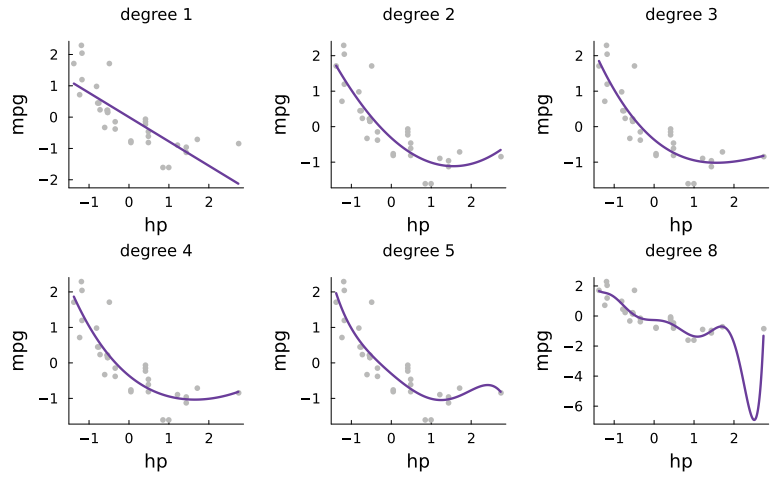


Figure 12.3: Fitting a polynomial regression to the `mtcars` data.

12.2 L2-regularization and Ridge regression

The overfitting problem with high order polynomial regression has its roots in that the regression coefficients β_j are allowed take on any values, including values that may be very large (in absolute value). This causes the fitted polynomial function to be very wiggly, and have a large variance from sample to sample. There is of course a simple solution to this overfitting problem: use a low order polynomial, or even a linear model (first order polynomial). This is somewhat extreme however since it corresponds to setting the β_j for higher order terms exactly to zero. An obvious drawback with this approach is that we may easily underfit when the data generating process actually requires a more flexible model.

The traditional non-Bayesian way out of this dilemma is to use a flexible model e.g. a high order polynomial, but **penalizing** large values of the regression coefficients in the fitting procedure. This has the effect of encouraging the fitting method to produce estimates that imply a smoother model fit. One of the most commonly used regularization method is **L2-regularization** which modifies the MLE estimate by adding a quadratic penalty $\|\boldsymbol{\beta}\|_2^2 = \boldsymbol{\beta}^\top\boldsymbol{\beta}$ to the log-likelihood, where $\|\boldsymbol{\beta}\|_2 = (\boldsymbol{\beta}^\top\boldsymbol{\beta})^{1/2}$ is the L_2 norm, i.e the usual Euclidean length. For regression, the penalty is usually added to $-2\log(p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}))$, which in the Gaussian case is just the residual sum of squares

penalizing

L2-regularization

$$-2\log p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \tag{12.1}$$

The L2-regularized estimate therefore minimizes

$$Q_\lambda(\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}, \tag{12.2}$$

where $\lambda > 0$ is a regularization parameter that determines the degree of penalization, and needs to be set by the user, or estimated by, for example, cross-validation. The first order condition for a minimum gives p equations to solve for the p unknown β_j:

$$\frac{\partial Q_\lambda(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = \mathbf{0}, \tag{12.3}$$

which has the following **ridge regression** estimator as solution

ridge regression

$$\hat{\boldsymbol{\beta}}_{L_2} = \left(\mathbf{X}^\top\mathbf{X} + \lambda I_p\right)^{-1}\mathbf{X}^\top\mathbf{y}, \tag{12.4}$$

where the subscript L_2 signifies the L2-norm used in the regularization.

The Ridge estimator has three important effects. First, due to the added λ on the diagonal of $\mathbf{X}^\top\mathbf{X}$ before taking the matrix inverse in (12.4), the Ridge regression gets improved numerical stability, it

solves the multicollinearity problem. It even becomes possible to use more covariates than the number of observations! This case is often called the $p > n$ case (or $p \gg n$ when the problem is really sparse) since p is typically used to denote the number of covariates and n the number of observations. In such cases, the least squares and maximum likelihood estimator of $\boldsymbol{\beta}$ do not exist since $\mathbf{X}^\top\mathbf{X}$ is non-singular. Adding λ to the diagonal fixes this as $\mathbf{X}^\top\mathbf{X} + \lambda I_p$ is invertible and we can compute $\hat{\boldsymbol{\beta}}_{L_2}$. Second, since the L2-penalty introduces a cost for having large elements in $\boldsymbol{\beta}$, the Ridge estimator gives estimates that are shrunken toward zero. This is most easily seen in the case of orthogonal covariates where $\mathbf{X}^\top\mathbf{X} = I_p$ since then the Ridge estimator becomes $\hat{\boldsymbol{\beta}}_{L_2} = c\hat{\boldsymbol{\beta}}$, where $c = 1/(1+\lambda)$ is the **shrinkage factor** and $\hat{\boldsymbol{\beta}}$ is the least squares solution. As $\lambda \to 0$ we have $\hat{\boldsymbol{\beta}}_{L_2} \to \hat{\boldsymbol{\beta}}$ whereas as $\lambda \to \infty$ the Ridge estimator gets shrunk all the way to zero. Third, the shrinkage is the same for all elements of $\boldsymbol{\beta}$ when $\mathbf{X}^\top\mathbf{X} = I_p$. In the more general case where the covariates are not orthogonal, the shrinkage of Ridge estimator is a little elaborate, but the end result is that $\hat{\boldsymbol{\beta}}_{L_2}$ applies more shrinkage along the dimensions of $\hat{\boldsymbol{\beta}}$ given by the eigenvectors of $\mathbf{X}^\top\mathbf{X}$ with the smallest eigenvalues; see Appendix 18.4 for a definition of eigenvectors and eigenvalues. Nevertheless, since Ridge regression only has one λ to control the shrinkage, its shrinkage is one-dimensional and can therefore be restrictive in some applications. We return to this point below, after learning how Ridge regression can be interpreted from a Bayesian point of view.

shrinkage factor

Ridge regression is an independent Gaussian prior

The prior

$$\beta_j|\sigma^2 \overset{\text{iid}}{\sim} N(\mathbf{0}, \sigma^2/\lambda) \tag{12.5}$$
$$\sigma^2 \sim \text{Inv}{-}\chi^2(\nu_0, \sigma_0^2) \tag{12.6}$$

for the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,\ \varepsilon \sim N(\mathbf{0}, \sigma^2 I_n) \tag{12.7}$$

implies a posterior mean for $\boldsymbol{\beta}$ equal to the Ridge estimator

$$\hat{\boldsymbol{\beta}}_{L_2} = \left(\mathbf{X}^\top\mathbf{X} + \lambda I_p\right)^{-1}\mathbf{X}^\top\mathbf{y}. \tag{12.8}$$

Figure 12.4: Ridge is Bayes with particular Gaussian prior.

Consider now a Bayesian approach using the following prior

$$\beta_j|\sigma^2 \overset{\text{iid}}{\sim} N(0, \sigma^2/\lambda), \tag{12.9}$$

where we initially assume σ^2 to be known for simplicity. Note that this prior is a special case of the $\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\mu}_0, \sigma^2\Omega_0^{-1})$ prior used in Chapter 5 on regression with $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\Omega_0 = \lambda I_p$, where the simple structure for Ω_0 comes from the iid assumption in (12.9). It then follows from Figure 5.3 that the posterior is Gaussian with a posterior mean equal to

$$\boldsymbol{\mu}_n = \Omega_n^{-1}(\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} + \Omega_0\boldsymbol{\mu}_0) = (\mathbf{X}^\top\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^\top\mathbf{y}, \tag{12.10}$$

which is exactly the Ridge estimator in (12.4). This is also the marginal posterior mean of $\boldsymbol{\beta}$ when σ^2 is unknown following a $\sigma^2 \sim \text{Inv}{-}\chi^2(\nu_0, \sigma_0^2)$ prior (see Figure 5.3), we have the result in Figure 12.4.

This Bayesian characterization of the Ridge regression estimator gives an interesting interpretation of the otherwise rather arbitrarily defined L2-penalty: Ridge regression comes from prior beliefs that the elements of $\boldsymbol{\beta}$ are Gaussian, independent and with the same prior precision.

To see the implication of the Gaussian assumption consider fitting a regression model with a large number of covariates, but where only a handful of those covariates actually have a sizeable effect on the response variable, i.e. most β_j are zero or very small (noise covariates), but a small number of the β_j are non-zero and potentially large (signal covariates). In such sparse situations we would like to have a method that estimates the β_j for noise covariates close to zero while preserving the β_j for noise covariates unshrunk. This will typically not happen with the Ridge estimator where in order for the estimator to be able to shrink all the unimportant β close to zero, it will also have to shrink the effects of the signal covariates. The end result will be a compromise where the effect of the noise covariates are not sufficiently shrunk toward zero and the effect of the signal covariates are shrunk more than one would like to. From a Bayesian point of view, this effect is entirely caused by the homoscedastic Gaussian prior. Why? Well, a Gaussian distribution has thin tails, it doesn't generate any outliers. So by saying that you believe all regression coefficients to be $N(0, \sigma^2/\lambda)$ a priori you are in effect saying: "I believe all the regression coefficients to be roughly of the same size. In particular, I do not think that many of them are close to zero while a few of them are very large". So, unless the data are really informative, your prior will affect the results and lead to over-shrinkage of the signals and under-shrinkage of the noise.

You may have wondered how exactly the Ridge estimator solves the multicollinearity problem, particularly in the $p > n$ case with more covariates than observations, where we need to estimate more regression coefficients than we have data points. How is it possible to separate out the effects of different covariates when we do not have

enough information in the data? The answer is that we are using extra information in addition to that coming from the data: the prior. Statistics is never a magic wand, you need enough information - data, prior or both - to make inferences.

The regularization parameter λ is important for the model fit. This is a prior hyperparameter that should be determined subjectively by the user. When it is too demanding to set a particular value for λ, the user can just put a prior on λ and estimate it along with $\boldsymbol{\beta}$ and σ^2. This is a hierarchical prior (see Chapter 4) where the joint prior for all unknown parameters $p(\boldsymbol{\beta}, \sigma^2, \lambda)$ can be decomposed as $p(\boldsymbol{\beta}|\sigma^2, \lambda)p(\sigma^2|\lambda)p(\lambda)$. For example,

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, \lambda &\sim N(\mathbf{0}, (\sigma^2/\lambda)I_p) \\ \sigma^2 &\sim \text{Inv}{-}\chi^2(\tau_0^2, \nu_0) \\ \lambda^{-1} &\sim \text{Inv}{-}\chi^2(\omega_0, \psi_0^2). \end{aligned} \tag{12.11}$$

The reason why we put a prior on λ^{-1} instead of λ is because λ is a precision parameter, i.e. the prior variance of each β_j is λ^{-1}. The scaled Inv$-\chi^2$ prior will be shown to be the conjugate prior for λ^{-1} *conditional on* $\boldsymbol{\beta}$ and σ^2. This will make it possible to set up a Gibbs sampling algorithm to sample the joint posterior $p(\boldsymbol{\beta}|\sigma^2, \lambda^{-1}|\mathbf{y}, \mathbf{X})$, from which we can simply invert each draw of λ^{-1} to obtain the marginal posterior of λ.

With the hierarchical prior in (12.11) the user now needs to specify the "best guess" ψ_0^2 for λ^{-1} and the degree of freedom ω_0 to determine the precision in that guess. It therefore seems that we have just pushed the problem one step down the hierarchy. However, the posterior of $\boldsymbol{\beta}$ is typically far less sensitive to ω_0 and ψ_0^2 than it is to λ.

To derive the full conditional for λ we first use Bayes' theorem to obtain

$$p(\lambda|\boldsymbol{\beta}, \sigma^2, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \lambda)p(\lambda|\boldsymbol{\beta}, \sigma^2) \tag{12.12}$$

However, conditional on $\boldsymbol{\beta}$ the distribution of the data $\mathbf{y}$ no longer depends on λ. This is because λ only enters via the prior for $\boldsymbol{\beta}$, it has no direct connection to the data. So the first factor $p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \lambda)$ in (12.12) does not depend on λ and can be absorbed in the proportionality constant. Using Bayes' theorem one more time, we can write

$$p(\lambda|\boldsymbol{\beta}, \sigma^2, \mathbf{y}) \propto p(\lambda|\boldsymbol{\beta}, \sigma^2) \propto p(\boldsymbol{\beta}|\sigma^2, \lambda)p(\lambda|\sigma^2). \tag{12.13}$$

By assumption, the prior for λ does not depend on σ^2 so we can write the full conditional posterior of λ as

$$p(\lambda|\boldsymbol{\beta}, \sigma^2, \mathbf{y}) \propto p(\boldsymbol{\beta}|\sigma^2, \lambda)p(\lambda). \tag{12.14}$$

This looks a little strange since the likelihood part $p(\boldsymbol{\beta}|\sigma^2, \lambda)$ does not involve the data $\mathbf{y}$ at all, and $\boldsymbol{\beta}$ seems to play the role of the data in the full conditional posterior of λ. This is entirely natural however since λ is a hyperparameter in the prior for $\boldsymbol{\beta}$ and λ does not figure in the likelihood for $\mathbf{y}$; so *conditional* on $\boldsymbol{\beta}$, these regression coefficients act like data for λ. In the *marginal* posterior $p(\lambda|\mathbf{y})$ the data $\mathbf{y}$ does however inform us about λ. One way to think about this is that $\mathbf{y}$ informs us about $\boldsymbol{\beta}$, and $\boldsymbol{\beta}$ informs us about λ.

Now, inserting $\beta_j|\sigma^2, \lambda \overset{\text{iid}}{\sim} N(\mathbf{0}, \sigma^2/\lambda)$ and $\lambda^{-1} \sim \text{Inv}{-}\chi^2(\omega_0, \psi_0^2)$ in (12.14) we get

$$
\begin{aligned}
p\left(\lambda^{-1}|\beta, \sigma^2, \mathbf{y}\right) &\propto p\left(\beta|\sigma^2, \lambda^{-1}\right) p\left(\lambda^{-1}\right) \\
&\propto \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2/\lambda}} \exp\left(-\frac{\beta_i^2}{2\sigma^2/\lambda}\right) \cdot \lambda^{\omega_0/2+1} \exp\left(-\lambda\frac{\omega_0\psi_0^2}{2}\right) \\
&\propto \lambda^{p/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{p}(\beta_i/\sigma)^2\right) \cdot \lambda^{\omega_0/2+1} \exp\left(-\lambda\frac{\omega_0\psi_0^2}{2}\right) \\
&\propto \lambda^{(p+\omega_0)/2+1} \exp\left(-\lambda\left(\frac{\sum_{i=1}^{p}(\beta_i/\sigma)^2 + \omega_0\psi_0^2}{2}\right)\right)
\end{aligned}
$$

which can be recognized as

$$
\lambda^{-1}|\boldsymbol{\beta}, \sigma^2, \mathbf{y} \sim \text{Inv}{-}\chi^2\left(\omega_0 + p, \frac{\sum_{i=1}^{p}(\beta_i/\sigma)^2 + \omega_0\psi_0^2}{\omega_0 + p}\right). \tag{12.15}
$$

This shows that the scaled Inv$-\chi^2$ distribution is indeed the conditionally conjugate prior for λ^{-1}. Note how the conditional posterior for λ^{-1} is determined by the *variability* in the normalized regression coefficients β_i/σ for $i = 1, \ldots, p$. If the data suggests a large variability in the normalized β_j then the posterior for λ^{-1} will concentrate on large values, and the posterior for λ will concentrate on small values, i.e. only mild shrinkage of the β_j toward zero. Since λ^{-1} is essentially learned from the variance of the β_j coefficients, inference for λ will be imprecise when p is small.

Conditional on λ, the joint posterior $p(\boldsymbol{\beta}, \sigma^2|\lambda, \mathbf{y})$ for $\boldsymbol{\beta}$ and σ^2 is directly given by Figure 5.3 from Chapter 5 with $\Omega_0 = \lambda I_p$. We can therefore set up a two-block Gibbs sampler with $\boldsymbol{\beta}$ and σ^2 in one block and λ in the other. This is summarized in Figure 12.5.

12.3 *L1-regularization and the Lasso estimator*

Gibbs sampling linear regression - L2 regularization prior

The posterior for the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,\ \varepsilon \sim N(\mathbf{0}, \sigma^2 I_n), \tag{12.16}$$

with hierarchical L2 regularization prior

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, \lambda &\sim N(\mathbf{0}, (\sigma^2/\lambda) I_p) \\ \sigma^2 &\sim \text{Inv}{-}\chi^2(\tau_0^2, \nu_0) \\ \lambda^{-1} &\sim \text{Inv}{-}\chi^2(\omega_0, \psi_0^2). \end{aligned}$$

can be sampled by a two-block Gibbs sampler:

$$\begin{aligned} \text{Block1}: \ \boldsymbol{\beta}|\sigma^2, \lambda, \mathbf{y} &\sim N\big(\hat{\boldsymbol{\beta}}_{L_2}, \sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda I_p)^{-1}\big) \\ \sigma^2|\lambda, \mathbf{y} &\sim \text{Inv}{-}\chi^2(\tau_n^2, \nu_n) \end{aligned}$$

$$\text{Block2}: \ \lambda^{-1}|\boldsymbol{\beta}, \sigma^2, \mathbf{y} \sim \text{Inv}{-}\chi^2(\omega_n, \psi_n^2),$$

where
$\hat{\boldsymbol{\beta}}_{L_2} = \left(\mathbf{X}^\top\mathbf{X} + \lambda I_p\right)^{-1}\mathbf{X}^\top\mathbf{y}$ is the Ridge estimator,
ν_n and τ_n^2 are given in Figure 5.3,
$\omega_n = \omega_0 + p$ and $\omega_n\psi_n^2 = \sum_{i=1}^{p}(\beta_i/\sigma)^2 + \omega_0\psi_0^2$.

Figure 12.5: Gibbs sampling for the linear regression model with a L2 regularization prior.

Lasso regression is an independent Laplace prior

The prior $\beta_j|\sigma^2 \overset{\text{iid}}{\sim} \text{Laplace}(0, \sigma^2/\lambda)$ for the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \varepsilon \sim N(\mathbf{0}, \sigma^2 I_n) \tag{12.17}$$

implies a posterior mode for $\boldsymbol{\beta}$ equal to the Lasso estimator.

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 12.8: Lasso is Bayes with particular Laplace prior.

Figure 12.6: Laplace distribution

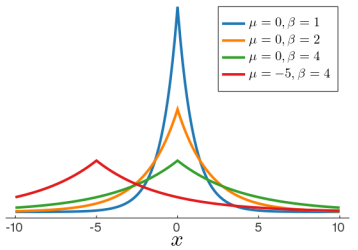


Figure 12.7: Some Laplace distributions.

12.4 Global-local regularization and Horseshoe

Both the L1 (Laplace) and L2 (Gaussian) regularization priors are *global* regularizers that penalize all coefficients equally using a single hyperparameter λ that acts on all element of $\boldsymbol{\beta}$. The **horseshoe prior** is instead a so called **global-local shrinkage prior** containing a global shrinkage parameter τ that acts on all p regression coefficients $\boldsymbol{\beta}$, but also local shrinkage parameters λ_j for $j = 1, \ldots, p$ that can modulate the global shrinkage on for β_j. The horseshoe prior is in the following hierarchical form

horseshoe prior

global-local shrinkage prior

$$\beta_j|\lambda_j^2, \tau^2, \sigma^2 \sim N\left(0, \sigma^2\tau^2\lambda_j^2\right) \tag{12.18}$$
$$\sigma^2 \sim \text{Inv}{-}\chi^2(\nu_0, \tau_0^2) \tag{12.19}$$
$$\lambda_j \sim C^+(0,1) \tag{12.20}$$
$$\tau \sim C^+(0,1) \tag{12.21}$$

where $C^+(0,1)$ is the half-Cauchy distribution with location parameter 0 and scale parameter 1, i.e. a Cauchy distribution (student-t distribution with one degree of freedom) truncated to have strictly positive support. As before the intercept β_0 is assigned a separate prior, usually non-informative, since it rarely makes sense to shrink the intercept.

When $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}_p$, posterior mean for $\boldsymbol{\beta}$ satisfies approximately

$$\mu_{n,j} \approx (1-\phi_j)\hat{\boldsymbol{\beta}}_j, \text{ where } \phi_j = \frac{1}{1+(n/\sigma^2)\tau^2\lambda_j^2}$$

is the local shrinkage factor for β_j and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is the least squares estimator.

A Gibbs sampler can be obtained by writing the half Cauchy distribution as continuous mixture (Makalic and Schmidt, 2015)

$$X \sim C^+(0,1) \iff X^2|Y \sim \text{Inv}{-}\chi^2(1, 2/Y) \text{ and } Y \sim \text{Inv}{-}\chi^2(1,2),$$

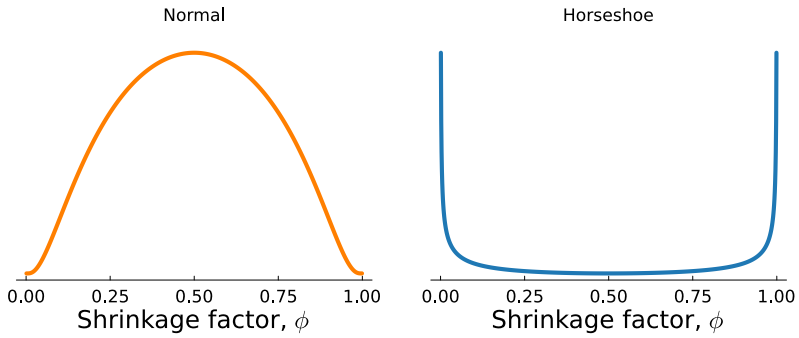


Figure 12.9: Comparing the implied prior on the shrinkage parameter in the normal and horseshoe priors.

meaning that the density function of $X \sim C^+(0,1)$ can be expressed as

$$p(x) = \int_0^\infty p(x|y)p(y)dy.$$

The horseshoe prior can then be written as

$$\begin{aligned}
\boldsymbol{\beta}|\lambda_1,\ldots,\lambda_p,\tau^2,\sigma^2 &\sim N\left(0,\sigma^2\tau^2\boldsymbol{\Lambda}\right), \text{ where } \boldsymbol{\Lambda} = \text{Diag}(\lambda_1,\ldots,\lambda_p) \\
\sigma^2 &\sim \text{Inv}-\chi^2(\nu_0,\tau_0^2) \\
\lambda_j^2|\nu_j &\sim \text{Inv}-\chi^2(1,2/\nu_j) \\
\tau^2|\xi &\sim \text{Inv}-\chi^2(1,2/\xi) \\
\nu_1,\ldots,\nu_p,\xi &\overset{\text{iid}}{\sim} \text{Inv}-\chi^2(1,2)
\end{aligned}$$

It should be immediately clear that $p(\boldsymbol{\beta},\sigma^2|\lambda_1,\ldots,\lambda_p,\tau)$ is exactly like the posterior for the linear regression in Chapter Regression with $\boldsymbol{\Omega}_0^{-1} = \tau^2\boldsymbol{\Lambda}$. Recall the L2-regularization case where the posterior for the hyperparameter λ given $\boldsymbol{\beta}$ and σ^2 was only determined from the prior $p(\boldsymbol{\beta},\sigma^2|\lambda)$ without the likelihood for the data entering all. The same thing happens here and one can show that the full conditional posteriors for the hyperparameters in the Horsehoe prior are (Makalic and Schmidt, 2015):

$$\lambda_j^2|\nu_j,\tau,\boldsymbol{\beta},\sigma \sim \text{Inv}-\chi^2\left(2,\frac{1}{\nu_j}+\frac{1}{2}\left(\frac{\beta_j}{\sigma\tau}\right)^2\right) \tag{12.22}$$

$$\tau^2|\xi,\lambda_1,\ldots,\lambda_p,\boldsymbol{\beta},\sigma \sim \left(p+1,\frac{\frac{2}{\xi}+\sum_{j=1}^p\left(\frac{\beta_j}{\sigma\lambda_j}\right)^2}{p+1}\right) \tag{12.23}$$

$$\nu_j|\lambda_j \overset{\text{iid}}{\sim} \text{Inv}-\chi^2(2,1+1/\lambda_j^2) \tag{12.24}$$

$$\xi|\tau \overset{\text{iid}}{\sim} \text{Inv}-\chi^2(2,1+1/\tau^2), \tag{12.25}$$

where we have only written out explicitly the conditioning on the parameters that appear in each full conditional distribution.

13 Model comparison

13.1 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 in 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 is 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

$$\Pr(M = M_k|\mathbf{x}) \propto p(\mathbf{x}|M_k) \cdot \Pr(M_k), \tag{13.1}$$

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 13.1 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 13.1: Example of prior-to-posterior updating of model probabilities.

The likelihood contribution to (13.1), $p(\mathbf{x}|M_k)$, does not condition on the parameters $\boldsymbol{\theta}_k$ in model M_k; the parameters have been marginalized out and

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\boldsymbol{\theta}_k, M_k) p(\theta_k|M_k) d\theta_k, \tag{13.2}$$

is therefore usually called the **marginal likelihood**. The alternative

marginal likelihood

name **evidence** is often used in machine learning. It is important to note that the parameters are integrated out by the *prior* and that the marginal likelihood is the prior expected likelihood function:

evidence

$$p(\mathbf{x}|M_k) = \mathbb{E}_{\theta_k}\left(p(\mathbf{x}|\boldsymbol{\theta}_k, M_k)\right). \tag{13.3}$$

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

$$B_{12}(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}. \tag{13.4}$$

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.

13.2 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:

$$p(\mathbf{x}|M_1) = \int \prod_{i=1}^{n} N(x_i|\theta, \sigma^2)\mathrm{N}(\theta|\mu_0, \sigma^2/\kappa_0)d\theta. \tag{13.5}$$

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

$$p(\mathbf{x}|M_1, \sigma^2) = c(\sigma^2, s^2) N\big(\bar{x}|\mu_0, \sigma^2(1/n + 1/\kappa_0)\big), \tag{13.6}$$

and the Bayes factor for a given σ^2 is

$$\mathrm{BF}_{01}(\mathbf{x}, \sigma^2) = \frac{p(\mathbf{x}|M_0, \sigma^2)}{p(\mathbf{x}|M_1, \sigma^2)} = \frac{\mathrm{N}(\bar{x}|\mu_0, \sigma^2/n)}{\mathrm{N}\big(\bar{x}|\mu_0, \sigma^2(1/n + 1/\kappa_0)\big)}. \tag{13.7}$$

The expression in (13.7) 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 13.1 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 region 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 13.2 illustrates how the prior predictive density assigns increasingly lower density to the observed $\bar{x} = 15.99$ when κ_0 decreases.

Figure 13.3 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 and 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 $\kappa_0 \in (0.01, 3)$. This is also clear from Figure 13.2 if we move the purple data point to $\bar{x} = 12$.

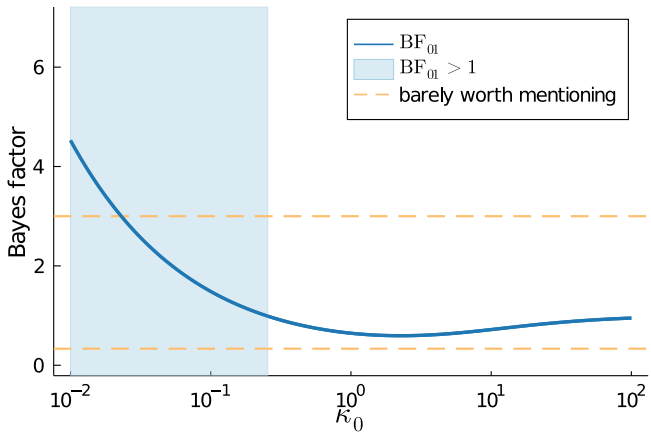


Figure 13.1: Bayes factor for the internet speed data with known variance $\sigma^2 = 5^2$. The graph plots the Bayes factor BF_{01} as a function of the prior sample size κ_0 in log-scale. The shaded region shows the values for κ_0 where $\text{BF}_{01} > 1$, i.e. where there is support in favor of the null model. The limits for "barely worth mentioning" in the Jeffreys scale of evidence are marked out as horizontal orange dashed lines.

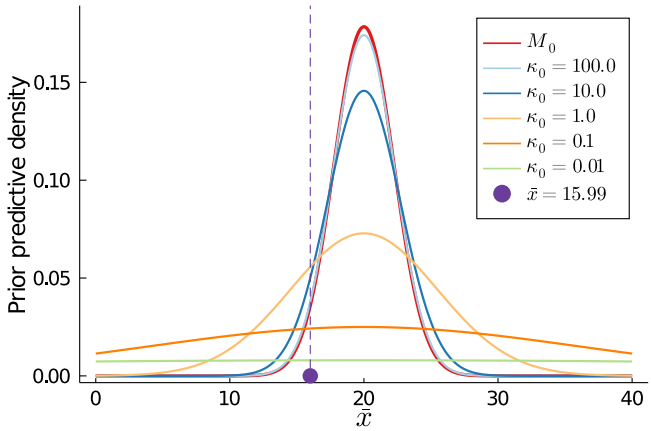


Figure 13.2: Internet speed data with known variance $\sigma^2 = 5^2$. Prior predictive densities for $\bar{x}$ in the models M_0 and M_1 for different values of the prior hyperparameter κ_0. The realized data of $\bar{x} = 15.99$ is shown as a purple dot with dashed line.

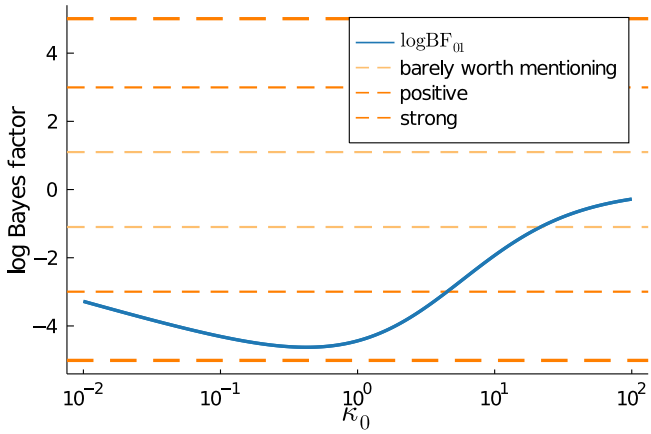


Figure 13.3: Log Bayes factor for the internet speed data with $\bar{x}$ artificially set to $\bar{x} = 12$ instead of the actually observed $\bar{x} = 15.99$. The graph plots the log Bayes factor BF_{01} as function of the prior sample size κ_0 in log-scale. The limits for Jeffreys' scale of evidence (in logs) are marked out as 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 as 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 13.6 shows the results from this experiment repeated four times to also see the sampling variation. The graph to the left in Figure 13.6 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 13.6 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:

$$\Pr(\mathrm{M}_k^\star | \mathbf{x}) \xrightarrow{p} 1 \text{ as } n \to \infty, \tag{13.8}$$

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 13.4: The Geometric distribution.

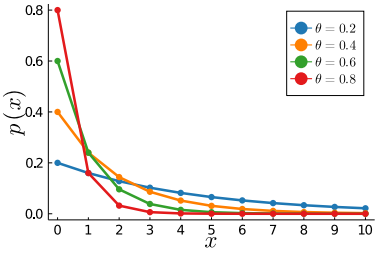


Figure 13.5: Some Geometric distributions.

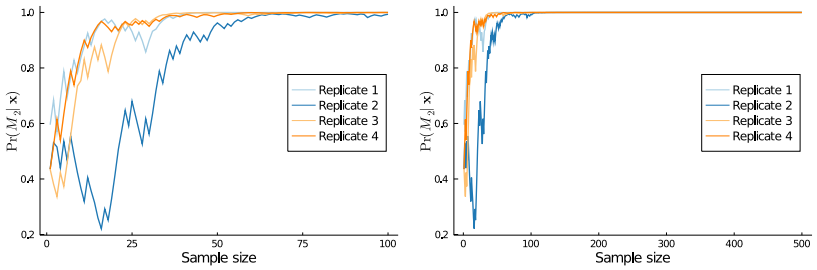


Figure 13.6: Asymptotic behavior of posterior model probabilities in $\mathcal{M}$-closed when comparing the models:
M_1: Geo(θ_1), $\theta_1 \sim$ Beta$(10, 10)$
M_2: Pois(θ_2), $\theta_2 \sim$ Gamma$(10, 10)$.
The graphs show the evolution of the posterior probability for the Poisson model as the sample 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:

$$p(x) = \binom{x+r-1}{x}(1-\theta)^r\theta^x, \text{ for } x = 0, 1, 2, \ldots \tag{13.9}$$

The negative binomial in Single-parameter models was the total number of trials until a certain number of successes. The negative binomial in (13.9) instead counts the number of successes x before the rth failure occurs.

Figure 13.9 (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 13.9 (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

$\mathcal{M}$-open

negative binomial distribution

Negative binomial distribution

$X \sim \text{NegBin}(r, \theta)$
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 13.7: The Negative binomial distribution.

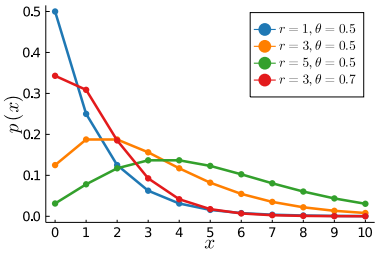


Figure 13.8: Some Negative binomial distributions.

asymptotically. The asymptotic tendency seen in Figure 13.9 can be proved to hold quite generally in that

$$\Pr(\mathrm{M}_k^\star|\mathbf{x}) \xrightarrow{p} 1 \text{ as } n \to \infty, \tag{13.10}$$

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

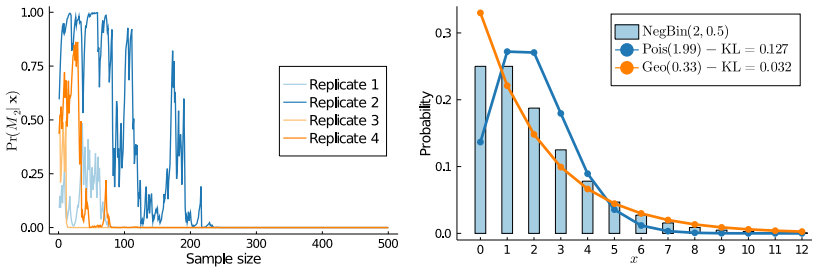


Figure 13.9: Asymptotic behavior of posterior model probabilities in $\mathcal{M}$-open when comparing the models:
M_1: Geo(θ_1), $\theta_1 \sim$ Beta$(10, 10)$
M_2: Pois(θ_2), $\theta_2 \sim$ Gamma$(10, 10)$.
The left graph shows the evolution of the posterior probability for the Poisson model as the sample size increases. Each line corresponds to a replication of the experiment. The data are generated from the iid 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 regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 I_n), \tag{13.11}$$

is given by

$$p(\mathbf{y}|\mathbf{X}) = \iint p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) p(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2. \tag{13.12}$$

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

$$\mathbf{y}|\mathbf{X} \sim t_{\nu_0}\left(\mathbf{X}\boldsymbol{\mu}_0, \sigma_0^2(\boldsymbol{I}_n + \mathbf{X}\Omega_0^{-1}\mathbf{X}^\top)\right). \tag{13.13}$$

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

13.3 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

$$\widehat{p(\mathbf{x})} = \frac{1}{m}\sum_{i=1}^{m} p(\mathbf{x}|\boldsymbol{\theta}^{(i)}), \tag{13.14}$$

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

Unfortunately, the simple Monte Carlo estimator in (13.14) usually has disastrously large variance and is rarely used in practice. The problem with the estimator in (13.14) 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

$$\ln \hat{p}(\boldsymbol{x}) = \ln p(\boldsymbol{x}|\hat{\theta}) + \ln p(\hat{\theta}) + (1/2)\ln\left|J^{-1}_{\boldsymbol{x},\hat{\theta}}\right| + (p/2)\ln(2\pi), \tag{13.15}$$

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:

$$J_{\hat{\boldsymbol{\theta}},\boldsymbol{x}} = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{13.16}$$

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:

$$\begin{aligned}
\ln p(\boldsymbol{x}|\theta)p(\theta) &= (\alpha+s-1)\ln\theta + (\beta+f-1)\ln(1-\theta) \\
\frac{\partial \ln p(\boldsymbol{x}|\theta)p(\theta)}{\partial\theta} &= \frac{\alpha+s-1}{\theta} - \frac{\beta+f-1}{1-\theta} \\
\frac{\partial^2 \ln p(\boldsymbol{x}|\theta)p(\theta)}{\partial\theta^2} &= -\frac{\alpha+s-1}{\theta^2} - \frac{\beta+f-1}{(1-\theta)^2}
\end{aligned}$$

Solving $\partial \ln p(\boldsymbol{x}|\theta)p(\theta)/\partial\theta = 0$ 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 $J^{-1}_{\boldsymbol{x},\hat{\theta}} = sf/n^3 = 0.024$. The Laplace approximation of the log marginal likelihood in (13.15) is therefore

$$\ln \hat{p}(\boldsymbol{x}) = 6\ln(0.6) + 4\ln(0.4) + (1/2)\ln(0.024) + (1/2)\ln(2\pi) \approx -7.676,$$

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 $\theta = 0.5$. 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.

13.4 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 the 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:

$$p(x_1,\ldots,x_n) = p(x_1)p(x_2|x_1)\cdots p(x_1|x_1,x_2,\ldots,x_{n-1}) \qquad (13.17)$$

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

In a time series context where the observations have a natural ordering in time, the factor $p(x_i|x_1,...,x_{i-1})$ in the decomposition in (13.17) 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 (13.17) can be done in many different ways by ordering the observations differently; we return this interpretation later in this section.

The decomposition in (13.17) 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 (13.17) 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: $x_1,\ldots,x_n|\theta \overset{\text{iid}}{\sim} \text{N}(\theta,\sigma^2)$ 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 (13.17) is

$$x_i|x_1,\ldots,x_{i-1} \sim \text{N}\left(\mu_{i-1},\sigma^2\left(1+\frac{1}{i-1+\kappa_0}\right)\right), \tag{13.18}$$

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 $\text{N}(20,5^2)$ 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 13.10 plots the log of the marginal likelihood decomposition

$$\log p(x_1,\ldots,x_n) = \log p(x_1) + \log p(x_2|x_1) + \ldots + \log p(x_1|x_1,\ldots,x_{n-1}), \tag{13.19}$$

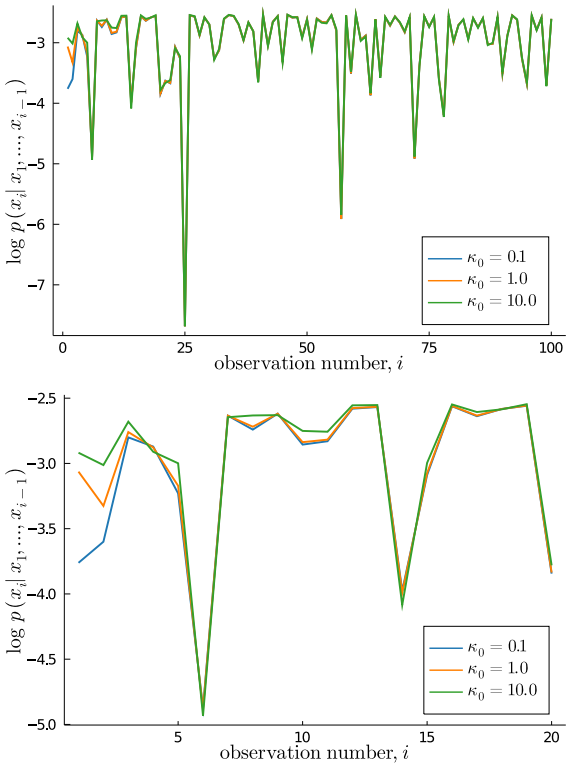


Figure 13.10: Decomposition of the log marginal likelihood for the simulated internet speed data with $n = 100$ observations for three different values for the prior sample size κ_0. The bottom graph zooms in on the gray shaded region with the first 20 observations.

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 in the decomposition (13.19). The lower graph in Figure 13.10 makes this more visible by zooming in on the 20 first observations. This comes as no surprise since it is clear from (13.18) 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 to reduce the prior sensitivity while still remaining close to the marginal likelihood is therefore to discard the first terms in (13.19). This is the **log predictive score** (**LPS**):

log predictive score

$$\text{LPS} = \sum_{i=i^\star+1}^{n} \log p(x_i|x_1,\ldots,x_{i-1}). \tag{13.20}$$

The LPS is effectively using the first $i^\star$ observations to train the prior $p(\boldsymbol{\theta})$ into an 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 (13.20) 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 13.11 plots the LPS as a function of the training fraction $f = i^\star/n$ for each of the three κ_0 values. 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 13.11 with training fraction $f = 0$ is the original log marginal likelihood where the prior (κ_0) has a substantial effect on the LPS. Already with a training fraction of 15% is the LPS insensitive to $\kappa_0 \in [0.1, 10]$.

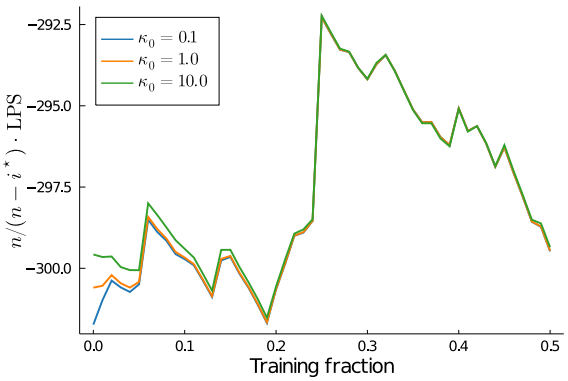


Figure 13.11: Log predictive score (scaled) as a function of the training fraction for the simulated internet speed data with $n = 100$ observations.

The LPS in (13.20) 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 evaluate the predictive performance on the $K-1$ folds left out. This is repeated K times, each time with a new fold as the training fold. Table 13.2 illustrates the data partitioning. Note that this is different from the usual cross-validation used in machine learning where instead $K-1$ folds would be used for training and the single remaining 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 methods that aim to estimate the generalization performance of the model are discussed in the next section.

K-fold cross-validated LPS

	n data observations				
	$1, 2, \ldots, n-1, n$				
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 13.2: 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.

13.5 *Bayesian estimators of generalization performance*

Leave-one-out and cross-validation. WAIC

14 *Variable selection*

15 Gaussian processes

15.1 Gaussian processes

16 Interaction models

16.1 Surface splines

.

16.2 Bayesian regression trees

17 Mixture models

Finite mixtures

Mixtures of regressions

Latent Dirichlet allocation

Infinite mixtures

18 Dynamic models and sequential inference

18.1 Dynamic models

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

18.2 Bayesian filtering and smoothing

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

18.3 Sequential Monte Carlo

Basic particle filter

18.4 Sequential decision making

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

Bibliography

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

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

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

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Fanaee-T, H. and Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15.

Fox, J. and Weisberg, S. (2019). *An R companion to applied regression*. Sage publications.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*, volume 3rd edition. CRC press.

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

Harville, D. A. (1998). Matrix algebra from a statistician's perspective.

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

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

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

Lindgren, G. (2012). *Stationary stochastic processes: theory and applications*. CRC Press.

Lindholm, A., Wahlström, N., Lindsten, F., and Schön, T. B. (2022). *Machine Learning: A First Course for Engineers and Scientists*. Cambridge University Press.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Mardia, K., Kent, J., and Bibby, J. (1979). Multivariate analysis, 1979.

Migon, H. S., Gamerman, D., and Louzada, F. (2014). *Statistical inference: an integrated approach*. CRC press.

Sundberg, R. (2019). *Statistical modelling by exponential families*, volume 12. Cambridge University Press.

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

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

Appendix: Some Mathematical results

A.1 Some linear algebra

This section summarizes some selected results from matrix algebra and multivariate analysis. The results are mostly given without proof, and the reader is referred to for example Harville (1998) for an extensive account or Appendix A in Mardia et al. (1979) for a more condensed treatment. The starred sections are not strictly required for understanding the material in this book, but are widely used results that every statistician should know about.

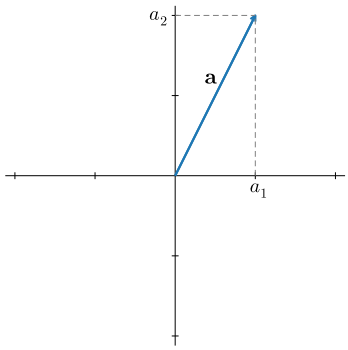


Figure A.11: Geometric illustration of the vector $\mathbf{a} = (a_1, a_2)^\top$.

Vectors, matrices and their products

Let

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

be a vector with p elements. We always define vectors as *column* vectors. A vector can be turned into a row vector by the **vector transpose** $\mathbf{a}^\top = (a_1, a_2, \ldots, a_p)$.

vector transpose

The **dot product** of two vectors $\mathbf{a}$ and $\mathbf{b}$ with the same number elements is defined as

$$\mathbf{a}^\top \mathbf{b} = \sum_{j=1}^{p} a_j b_j,$$

which is often written as $\mathbf{a} \cdot \mathbf{b}$. Two vectors $\mathbf{a}$ and $\mathbf{b}$ are **orthogonal** (perpendicular) to each other if and only if $\mathbf{a} \cdot \mathbf{b} = 0$; see Figure A.1.

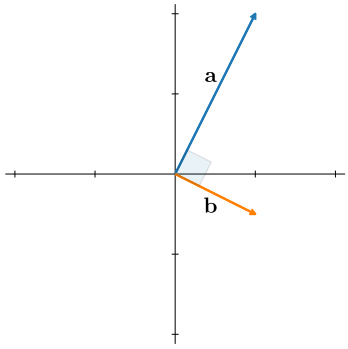


Figure A.12: Geometric illustration of two orthogonal vectors $\mathbf{a}$ and $\mathbf{b}$.

dot product

orthogonal

The *Euclidean length*, or L_2**-norm**, of a vector is defined as

$$\|\mathbf{a}\|_2 = (\mathbf{a}^\top \mathbf{a})^{1/2} = \Big(\sum_{j=1}^{p} a_j^2 \Big)^{1/2}.$$

L_2-norm

Another common norm is the L_1**-norm**

$$\|\mathbf{a}\|_1 = \sum_{j=1}^{p} |a_j|.$$

L_1-norm

Let $\mathbf{A}$ be a $p \times r$ matrix, i.e. and matrix with p rows and r columns:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pr} \end{pmatrix}.$$

The **identity matrix** $\mathbf{I}_p$ is the $p \times p$ matrix

identity matrix

$$\mathbf{I}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which pays the role of 1 in the world of matrices so that $\mathbf{A}\mathbf{I}_p = \mathbf{I}_p\mathbf{A} = \mathbf{A}$ for any $p \times p$ matrix $\mathbf{A}$.

The **matrix-vector product** of an $p \times r$ matrix $\mathbf{A}$ and r-element vector $\mathbf{b} = (b_1, b_2, \ldots, b_r)^\top$ is

matrix-vector product

$$\mathbf{A}\mathbf{b} = \begin{pmatrix} \sum_{j=1}^r a_{1j} b_j \\ \sum_{j=1}^r a_{2j} b_j \\ \vdots \\ \sum_{j=1}^r a_{pj} b_j \end{pmatrix}.$$

Defining $\mathbf{a}_i^\top$ to be the ith row of $\mathbf{A}$ we can write

$$\mathbf{A}\mathbf{b} = \begin{pmatrix} \mathbf{a}_1^\top \mathbf{b} \\ \mathbf{a}_2^\top \mathbf{b} \\ \vdots \\ \mathbf{a}_p^\top \mathbf{b} \end{pmatrix},$$

where $\mathbf{a}_i^\top \mathbf{b} = \sum_{j=1}^r a_{ij} b_j$ is a simple vector (dot) product.

Similarly, the **matrix-matrix product** of the $p \times q$ matrix $\mathbf{A}$ and the $q \times r$ matrix $\mathbf{B}$ is defined as

matrix-matrix product

$$\mathbf{A}\mathbf{B} = \begin{pmatrix} \mathbf{a}_1^\top \mathbf{b}_1 & \mathbf{a}_1^\top \mathbf{b}_2 & \cdots & \mathbf{a}_1^\top \mathbf{b}_r \\ \mathbf{a}_2^\top \mathbf{b}_1 & \mathbf{a}_2^\top \mathbf{b}_2 & \cdots & \mathbf{a}_2^\top \mathbf{b}_r \\ \vdots & & & \\ \mathbf{a}_p^\top \mathbf{b}_1 & \mathbf{a}_p^\top \mathbf{b}_2 & \cdots & \mathbf{a}_p^\top \mathbf{b}_r \end{pmatrix}.$$

Note the the number of columns in $\mathbf{A}$ must equal the number of rows in $\mathbf{B}$ and the end result of the product is a matrix with dimensions $p \times r$. We use the terminology that $\mathbf{A}$ *pre-multiplies* $\mathbf{B}$ in the product $\mathbf{AB}$, or, equivalently, that $\mathbf{B}$ *post-multiplies* $\mathbf{A}$.

The **matrix transpose** of $p \times r$ matrix $\mathbf{A}$, denoted by $\mathbf{A}^\top$, is the

matrix transpose

$r \times p$ matrix where the ith column is the i row of $\mathbf{A}$ Let $\mathbf{A}$ be a matrix with p rows and r columns

$$\mathbf{A}^\top = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{1p} \\ a_{12} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{1p} & \cdots & a_{rp} \end{pmatrix}.$$

Determinant and inverse matrix

The **determinant** of a square 2×2 matrix $\mathbf{A}$ is the scalar (i.e. single number) determinant

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \tag{1}$$

and for a 3×3 matrix

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{array}{l} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ -a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}, \end{array} \tag{2}$$

and increasingly more complex expressions for higher dimensional matrices. The exact expressions are less important here however. It is enough to remember that a determinant of a matrix $\mathbf{A}$ is a scalar that represent the *volume* of the matrix, in the sense that the absolute value of the determinant of $\mathbf{A}$ is the volume of a parallelepiped formed by the columns of $\mathbf{A}$; see Figure A.1 for an illustration. We will most often see the determinant of a covariance matrix Σ for a random vector $\mathbf{x}$, where $|\Sigma|$ can then be taken as a measure of *total variance* of $\mathbf{x}$.

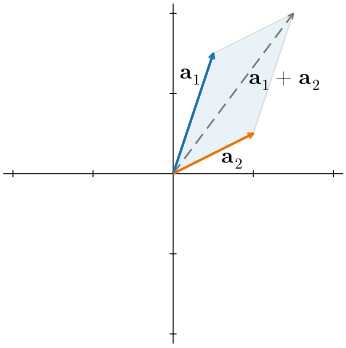


Figure A.13: Geometric illustration of the determinant as the area of the parallelogram formed by the 2×2 matrix $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2)$.

Some rules of determinants are worth noting. First, $|c\mathbf{A}| = c^p|\mathbf{A}|$ for any scalar c and $p \times p$ matrix $\mathbf{A}$. Second, the determinant of a diagonal matrix is just the product of the diagonal elements

$$\begin{vmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{pp} \end{vmatrix} = a_{11}a_{22}\cdots a_{pp}.$$

The same is true for a lower diagonal matrix, i.e. a matrix where all the elements above the diagonal are zero, but some elements on the diagonal and/or below the diagonal may be non-zero. Finally, for the product of two square matrices $\mathbf{A}$ and $\mathbf{B}$ we have

$$|\mathbf{AB}| = |\mathbf{A}| \cdot |\mathbf{B}|. \tag{3}$$

The same type of result holds for a product of three matrices $|\mathbf{ABC}| = |\mathbf{A}| \cdot |\mathbf{B}| \cdot |\mathbf{C}|$ and so on.

The **matrix inverse** of a square $p \times p$ matrix $\mathbf{A}$ is the matrix $\mathbf{A}^{-1}$ such that

matrix inverse

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = I_p. \tag{4}$$

Not every square matrix has an inverse, but when it exists it is unique. A sufficient and necessary condition for a square matrix $\mathbf{A}$ to have an inverse is that its column are linearly independent, i.e. that $\sum_{j=1}^{p} \alpha_j \mathbf{a}_j = \mathbf{0}$ only for $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$, where $\mathbf{a}_j$ is the jth column of $\mathbf{A}$ and $\mathbf{0}$ is the zero vector. Invertible matrices are also called non-singular. Here are two useful rules for inverses:

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$$

and if both $\mathbf{A}$ and $\mathbf{B}$ are invertible then

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1},$$

where you should note the reverse order of the matrices. The same type of result holds for a product of three matrices $(\mathbf{A}\mathbf{B}\mathbf{C})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$.

The **matrix trace** of a matrix $\mathbf{A}$ is simply the sum of its diagonal elements

matrix trace

$$\mathrm{tr}(\mathbf{A}) = \sum_{j=1} a_{jj}. \tag{5}$$

The trace has the following circular property

$$\mathrm{tr}(\mathbf{A}\mathbf{B}\mathbf{C}) = \mathrm{tr}(\mathbf{C}\mathbf{A}\mathbf{B}) = \mathrm{tr}(\mathbf{B}\mathbf{C}\mathbf{A}), \tag{6}$$

for any square matrices $\mathbf{A}$,$\mathbf{B}$ and $\mathbf{C}$ with the same dimensions.

*Partitioned matrices**

Consider a *partitioned matrix* of dimensions $p \times p$

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \tag{7}$$

where $\mathbf{A}_{11}$ is of dimensions $p_1 \times p_1$, $\mathbf{A}_{22}$ is of dimensions $p_2 \times p_2$, $\mathbf{A}_{12}$ and $\mathbf{A}_{21}$ are of dimensions $p_1 \times p_2$ and $p_2 \times p_1$ respectively. Hence, $p = p_1 + p_2$. The determinant can be then be expressed

$$|\mathbf{A}| = |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}| = |\mathbf{A}_{22}||\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}|.$$

and the inverse

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}^{(11)} & -\mathbf{A}^{(11)}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{(11)} & \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1} \end{pmatrix},$$

where $\mathbf{A}^{(11)} = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1}$.

*Linear transformation, eigendecomposition and principal components**

Consider a linear transformation $\mathbf{y} = \mathbf{m} + \mathbf{A}\mathbf{x}$ from $\mathbf{x}$ to $\mathbf{y}$, where $\mathbf{y}$ and $\mathbf{m}$ are p-dimensional vectors, $\mathbf{x}$ is an q-dimensional vector, and $\mathbf{A}$ is a $p \times q$ matrix. If $\mathbf{x}$ is a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then

$$\mathbb{E}(\mathbf{y}) = \mathbf{m} + \mathbf{A}\boldsymbol{\mu} \tag{8}$$
$$\mathbb{V}(\mathbf{y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top \tag{9}$$

Let $p = 1$ so that $\mathbf{A} = \mathbf{a}^\top$ is a r-dimensional row vector. Then $y = m + \mathbf{a}^\top\mathbf{x} = m + \sum_{i=1}^r a_i x_i$ is a scalar, and $\mathbb{V}(y) = \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}$. Since we require a variance to be positive we must require that the covariance matrix $\boldsymbol{\Sigma}$ satisfies $\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a} > 0$ for all $\mathbf{a} \neq \mathbf{0}$. We say that $\boldsymbol{\Sigma}$ must be **positive definite**. A matrix $\boldsymbol{\Sigma}$ is positive definite if and only if $|\boldsymbol{\Sigma}| > 0$. If we allow that the variance can also be exactly zero, then we require $\boldsymbol{\Sigma}$ to be positive semidefinite, sometimes abbreviated by psd or p.s.d.

positive definite

An **eigenvector** $\mathbf{v}$ of an invertible matrix $\mathbf{A}$ is a vector that keeps its direction when transformed by $\mathbf{A}$, i.e.

eigenvector

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v},$$

where λ is the **eigenvalue** associated with the eigenvector $\mathbf{v}$. Note how the transformation only leads to a scaling of $\mathbf{v}$ by λ, but the direction of the vector remains the same. A non-singular $p \times p$ matrix $\mathbf{A}$ has p linearly independent eigenvectors, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ each associated with its own eigenvalue $\lambda_1, \lambda_2, \ldots, \lambda_p$. Eigenvectors are normalized to have unit length, i.e. $\mathbf{v}_j^\top\mathbf{v}_j = 1$ for $j = 1, \ldots, p$ and to be orthogonal to each other, i.e. $\mathbf{v}_i^\top\mathbf{v}_j = 0$ for $i \neq j$. We can therefore collect all eigenvectors into a $p \times p$ *orthonormal* matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$ with the property $\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$; note that the inverse of an orthonormal matrix is simply its transpose. We can now write

eigenvalue

$$\mathbf{A}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}, \tag{10}$$

where $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \ldots, \lambda_p)$ is a diagonal matrix of eigenvalues. We therefore obtain the **spectral decomposition** of the invertible matrix $\mathbf{A}$ by post-multiplying both sides of (10) with $\mathbf{V}^\top$ (since $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$)

spectral decomposition

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top. \tag{11}$$

The spectral decomposition gives us connection between the determinant and inverse of a matrix and its eigenvalues and eigenvectors. The determinant can be written

$$|\mathbf{A}| = |\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top| = |\mathbf{V}||\boldsymbol{\Lambda}||\mathbf{V}^\top| = |\boldsymbol{\Lambda}||\mathbf{V}\mathbf{V}^\top| = \prod_{j=1}^p \lambda_j,$$

since the determinant of a diagonal matrix is the product of its diagonal elements and $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$ so $|\mathbf{V}\mathbf{V}^\top| = 1$. Given that a matrix is positive definite if its determinant is non-zero, this shows that a matrix is positive definite if and only if all of its eigenvalues are positive.

Since the inverse of an orthonormal matrix is its transpose, we can use the product rule for inverses to express the inverse of $\mathbf{A}$ as

$$\mathbf{A}^{-1} = (\mathbf{V}^\top)^{-1}\mathbf{\Lambda}^{-1}\mathbf{V}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^\top,$$

and $\mathbf{\Lambda}^{-1} = \text{Diag}(1/\lambda_1, \ldots, 1/\lambda_p)$. There are more general decompositions of matrices, also for non-square and non-invertible matrices, the most famous being the singular value decomposition Harville (1998).

Finally, using the circular property of the trace in (6), we see that the trace of matrix is the sum of its eigenvalues

$$\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top) = \text{tr}(\mathbf{V}^\top\mathbf{V}\mathbf{\Lambda}) = \text{tr}(\mathbf{\Lambda}) = \sum_{j=1}^{p} \lambda_j.$$

Consider now the spectral value decomposition $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ on a covariance matrix $\mathbf{\Sigma}$ of a random vector $\mathbf{x}$. The transformation $\mathbf{y} = \mathbf{V}^\top\mathbf{x}$ has an interesting covariance matrix

$$\mathbb{V}(\mathbf{y}) = \mathbf{V}^\top\mathbf{\Sigma}\mathbf{V} = \mathbf{V}^\top(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top)\mathbf{V} = \mathbf{\Lambda}. \tag{12}$$

Hence, the new variables in $y_j = \mathbf{v}_j^\top\mathbf{x}$ for $j = 1, \ldots, p$ are uncorrelated and have the eigenvalues as variances: $\mathbb{V}(y_j) = \lambda_j$. These variables are called the **principal components** of $\mathbf{x}$. If we order the eigenvalues in descending order $\lambda_1 \geq \cdots \geq \lambda_p$ then the first principal component $y_1 = \mathbf{v}_1^\top\mathbf{x}$ is the linear combination of the variables in $\mathbf{x}$ with maximal variance, the second principal component $y_2 = \mathbf{v}_2^\top\mathbf{x}$ is the linear combination with maximal variance subject to being uncorrelated with y_1 and so on. Replacing a possibly high-dimensional correlated $\mathbf{x}$ with the $r < p$ largest principal components is therefore a useful way to compress the data while retaining most of the variance. Figure A.14 illustrates the transformation of sampled data into uncorrelated principal components.

principal components

*Matrix powers and the Cholesky decomposition**

The spectral decomposition is useful for defining powers of a matrix. Let $\mathbf{A}$ be a square non-singular matrix with spectral decomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$. Then since V is orthonormal we have

$$\mathbf{A}^2 = \mathbf{A}\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^\top,$$

where $\mathbf{\Lambda}^2 = \text{Diag}(\lambda_1^2, \ldots, \lambda_p^2)$. Continuing by multiplying with additional $\mathbf{A}$ factors we have for any positive integer k the **matrix power**

matrix power

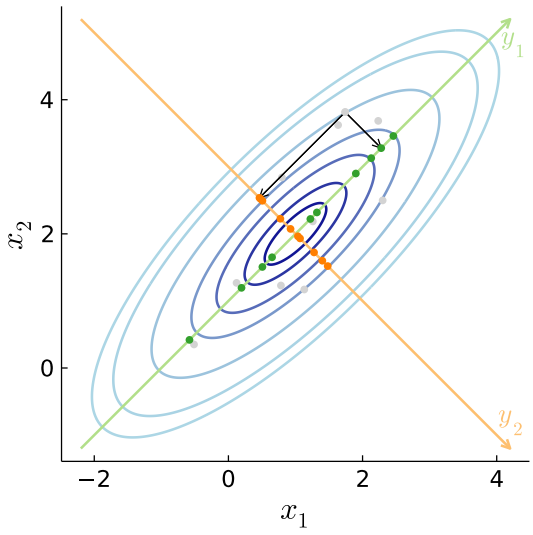


Figure A.14: Illustration of principal components from data points sampled from a multivariate normal distribution with mean $\boldsymbol{\mu} = (1,2)^\top$ and correlation $\rho = 0.8$. The sampled data points are shown in light gray and their projections onto the first principal components axis (y_1) are shown as green points and as orange points when projected against the second principal component axis (y_2); this projection is illustrated by arrows for one of the data points. The larger variability of the green points along the y_1 axis compared to the variability of the orange points along the y_2 is reflected in the eigenvalues $\lambda_1 = 1.8 > \lambda_2 = 0.2$.

$$\mathbf{A}^k = \mathbf{V}\boldsymbol{\Lambda}^k\mathbf{V}^\top.$$

We can extend this to any power k, not necessarily a positive integer, and in particular to $k = 1/2$ to define a **matrix square root** $\mathbf{A}^{1/2} = \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^\top$ with the property $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$. This construction can be used to simulate $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by

matrix square root

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}, \tag{13}$$

where $\mathbf{z}$ is a p-dimensional vector with independent standard normal variables. Since linear transformations of normal variables are normal, $\mathbf{x}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbb{V}(\mathbf{x}) = \boldsymbol{\Sigma}^{1/2}\mathbb{V}(\mathbf{z})\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}\mathbf{I}_p\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$ as required. The spectral decomposition is just one way of defining a matrix square root. Another commonly used matrix square root is the **Cholesky decomposition**

Cholesky decomposition

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\top, \tag{14}$$

where

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ l_{31} & l_{32} & l_{33} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ l_{p1} & l_{p2} & \ldots & l_{p,p-1} & l_{pp} \end{pmatrix}$$

is a lower triangular matrix. The Cholesky square root can of course equally well be used for multivariate normal simulation: if $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ then $\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where again $\mathbf{z}$ is a p-dimensional vector with independent standard normal variables. The Cholesky

decomposition makes it possible to compute the multivariate normal density cheaply since

$$\begin{aligned} p(\mathbf{x}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big) \\ &= (2\pi)^{-p/2}|\mathbf{L}\mathbf{L}^\top|^{-1/2} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top (\mathbf{L}\mathbf{L}^\top)^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big) \\ &= (2\pi)^{-p/2}|\mathbf{L}|^{-1} \exp\Big(-\frac{1}{2}\mathbf{y}^\top\mathbf{y}\Big), \end{aligned} \tag{15}$$

where $\mathbf{y} = \mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ and $|\mathbf{L}| = \prod_{j=1}^{p} l_{jj}$ since $\mathbf{L}$ is lower triangular. We can compute $\mathbf{y} = \mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ without explicitly inverting $\mathbf{L}$ by solving the system of equations $\mathbf{L}\mathbf{y} = \mathbf{x}-\boldsymbol{\mu}$ for $\mathbf{y}$. Since $\mathbf{L}$ is lower triangular this can be solved quickly using forward/backward substitution. Note that we have used several of the above mentioned results for determinants and inverses in (15), so verifying this derivation is a useful exercise.

*Vector differentiation**

Let $f(\mathbf{x})$ be a scalar valued function of an p-dimensional vector $\mathbf{x}$. The gradient of $f(\mathbf{x})$ with respect to $\mathbf{x}$ is the p-dimensional vector with partial derivatives

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_p} f(\mathbf{x}) \end{pmatrix}$$

The gradient is sometimes written $\nabla_{\mathbf{x}} f(\mathbf{x})$. For a linear function $f(\mathbf{x}) = \mathbf{a}^\top\mathbf{x}$ for some p-dimensional vector $\mathbf{a}$ the gradient is easily seen to be

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{a}^\top\mathbf{x} = \mathbf{a},$$

matching up with the one-dimensional case $\frac{d}{dx}ax = a$. For a quadratic function $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}$ for some square matrix $\mathbf{A}$, often called a quadratic form, we have the gradient

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x},$$

which also matches the one-dimensional case $\frac{d}{dx}ax^2 = 2ax$.

Consider now a *multi-output* function $\mathbf{y} = \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x}))^\top$ with p-dimensional output $\mathbf{y}$ and q-dimensional input $\mathbf{x}$. The $p \times q$ matrix of partial derivatives

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \frac{\partial}{\partial x_2} f_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_q} f_1(\mathbf{x}) \\ \vdots & & & \\ \frac{\partial}{\partial x_1} f_p(\mathbf{x}) & \frac{\partial}{\partial x_2} f_p(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_q} f_p(\mathbf{x}) \end{pmatrix}.$$

is called the **Jacobian matrix**. For a linear multi-output function $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ we have $\frac{\partial}{\partial \mathbf{x}}\mathbf{A}\mathbf{x} = \mathbf{A}$.

Jacobian matrix

Recall that the **chain rule** for differentiation of the function composition $f(x) = g(h(x))$ is the product of the so called outer and inner derivatives: $\frac{d}{dx}f(x) = \frac{d}{dz}g(z)\frac{d}{dx}h(x)$. The chain rule for a multi-dimensional function composition $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $h : \mathbb{R}^p \to \mathbb{R}^q$ and $g : \mathbb{R}^q \to \mathbb{R}$, is similar

chain rule

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \Big(\frac{\partial}{\partial \mathbf{x}} h(\mathbf{x})\Big)^\top \frac{\partial}{\partial \mathbf{z}} g(\mathbf{z}),$$

where $\mathbf{z} = h(\mathbf{x})$ is in general a mapping $\mathbf{x} \to \mathbf{z}$ from $\mathbb{R}^p$ to $\mathbb{R}^q$, so that $\frac{\partial}{\partial \mathbf{x}} h(\mathbf{x})$ is a $q \times p$ Jacobian matrix when both $p > 1$ and $q > 1$.

As an example on how to use the above rules for differentiation, consider deriving the least squares estimator in linear regression obtained by minimizing the residual sum of squares

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{e}(\boldsymbol{\beta})^\top \mathbf{e}(\boldsymbol{\beta}),$$

where $\mathbf{e}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is the vector of residuals. The least squares estimate is therefore the solution to $\frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \mathbf{0}$ where

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \Big(\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{e}(\boldsymbol{\beta})\Big)^\top \frac{\partial}{\partial \mathbf{e}} \mathbf{e}^\top \mathbf{e} = \big(\frac{\partial}{\partial \boldsymbol{\beta}}(-\mathbf{X}\boldsymbol{\beta})\big)^\top 2\mathbf{e} = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Hence the least squares estimator is the solution to $\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}$. If the columns of $\mathbf{X}$ are linearly independent then the inverse $\big(\mathbf{X}^\top \mathbf{X}\big)^{-1}$ exist and we can multiply both sides with it to get the least squares solution $\hat{\boldsymbol{\beta}} = \big(\mathbf{X}^\top \mathbf{X}\big)^{-1} \mathbf{X}^\top \mathbf{y}$.

A.2 Taylor approximation

The Taylor approximation is a tailored polynomial approximation of a function $f(x)$. The **Taylor series** of an infinitely differentiable function $f(x)$ is

Taylor series

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x - a)^k, \tag{16}$$

where $f^{(k)}(a)$ is the kth derivative of f evaluated in the point $x = a$.

The classical example of a Taylor series is that of the exponential function. The derivatives of the exponential function $f(x) = e^x$ are the exponential function itself, i.e. $f^{(k)}(x) = e^x$ for all k. The Taylor series expansion of the exponential function around $x = 0$ is therefore

$$\begin{aligned} e^x &= e^0 + \frac{1}{1!}e^0(x-0) + \frac{1}{2!}e^0(x-0)^2 + \frac{1}{3!}e^0(x-0)^3 + \dots \\ &= 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \\ &= \sum_{k=0}^{\infty} \frac{x^k}{k!}. \end{aligned}$$

A **Taylor approximation** of $f(x)$ uses only a small number of terms in the Taylor series

Taylor approximation

$$f(x) \approx \sum_{k=0}^{K} \frac{f^{(k)}(a)}{k!}(x-a)^k, \tag{17}$$

for some finite and typically small K. Figure A.21 shows how the Taylor approximation of e^x improves as higher order polynomial terms are included in the approximation. Taylor's theorem can be used to bound the approximation error of a kth order Taylor approximation using the $(k+1)$th derivative of the function.

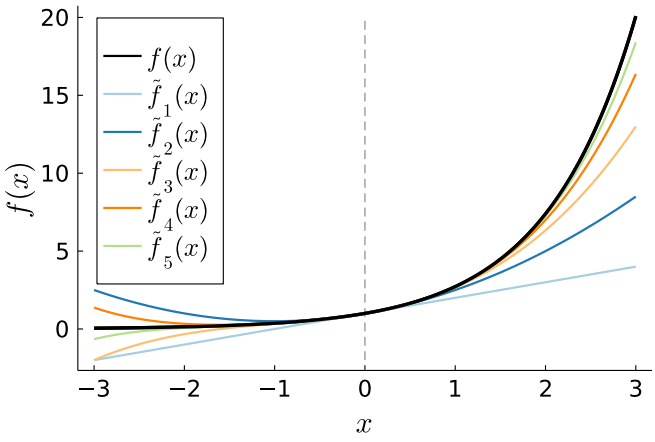


Figure A.21: Taylor approximation of the exponential function for different polynomial orders.

The Taylor expansion is a local approximation around the expansion point $x = a$, and the approximation is most accurate in a neighborhood around a. This point is illustrated in Figure A.22 where the function $\log(1+x)$ is well approximated only in the neighborhood around the expansion point $x = 0$.

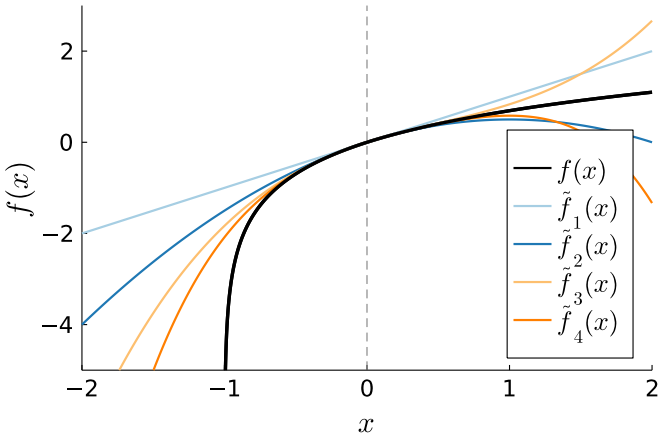


Figure A.22: Taylor approximation of $\log(1+x)$ around $x = 0$ for different approximation orders.

There is a multi-dimensional version of the Taylor approximation for functions $f(\mathbf{x}) = f(x_1, \ldots, x_d)$ of several variables. We will only

make use of the first and second order versions. The second order Taylor approximation of the function $f(\mathbf{x})$ around the point $\mathbf{x} = \mathbf{a}$ is

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{a}}(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{x}-\mathbf{a})^\top \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^\top}|_{\mathbf{x}=\mathbf{a}}(\mathbf{x}-\mathbf{a}),$$

where

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_d}\right),$$

is the **gradient** row vector with partial derivatives of $f(\mathbf{x})$ with respect to each of the input coordinates $x_1, \ldots, x_d$. The notation $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{a}}$ means that this vector of derivatives is evaluated in the point $\mathbf{x} = \mathbf{a}$. The $d \times d$ matrix $\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^\top}$ is the **Hessian** matrix

gradient

Hessian

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^\top} = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_d} \\ \vdots & & \ddots & \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_2} & & \frac{\partial^2 f(\mathbf{x})}{\partial x_d^2}, \end{pmatrix}$$

with second derivatives $\frac{\partial^2 f(\mathbf{x})}{\partial x_j^2}$ and cross-derivatives $\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k}$.

To see the multidimensional Taylor approximation in action, consider the following two-dimensional function

$$f(x_1, x_2) = \exp(x_1)\sin(x_2).$$

To compute a second order Taylor approximation around $\mathbf{x} = (0,0)^\top$ we need to compute the gradient vector and Hessian matrix. The gradient vector is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \Big(\exp(x_1)\sin(x_2), \exp(x_1)\cos(x_2)\Big),$$

which evaluates to $(0,1)$ at $\mathbf{x} = (0,0)^\top$. The Hessian matrix is

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^\top} = \begin{pmatrix} \exp(x_1)\sin(x_2) & \exp(x_1)\cos(x_2) \\ \exp(x_1)\cos(x_2) & -\exp(x_1)\sin(x_2) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

at $\mathbf{x} = (0,0)^\top$. The second order Taylor approximation is therefore

$$f(x_1, x_2) \approx 0 + (0,1)(x_1, x_2)^\top + \frac{1}{2}(x_1, x_2)^\top \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} (x_1, x_2) = x_2 + 2x_1x_2.$$

Figure A.23 plots the second order Taylor approximation of $\exp(x_1)\sin(x_2)$.

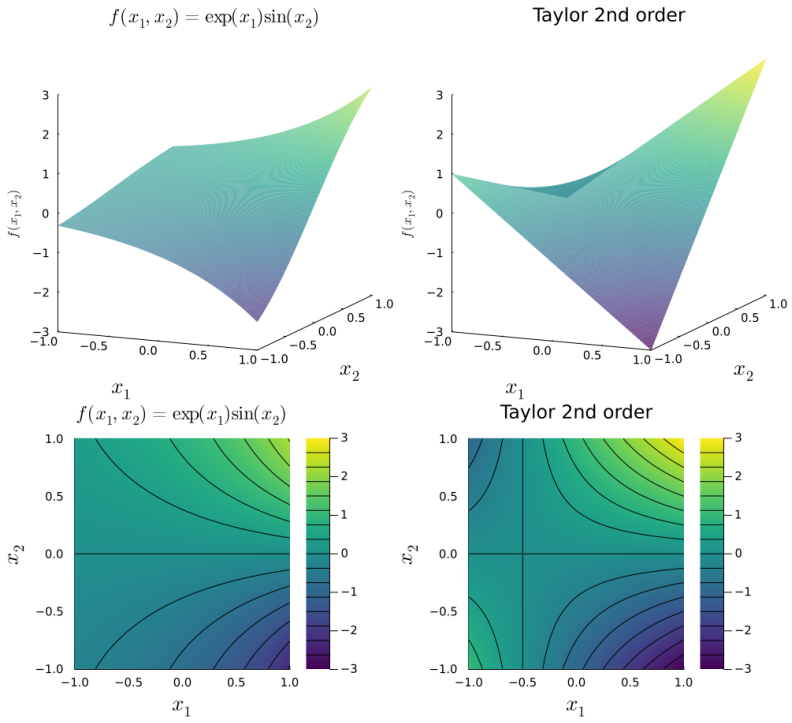


Figure A.23: Taylor approximation of $f(x_1, x_2) = \exp(x_1)\sin(x_2)$ around $\mathbf{x} = (0, 0)$. The graphs in the first row show function surface plots and the second row displays corresponding heatmaps and contours of the functions.

Index