STK4021- Applied Bayesian Analysis - Chapters 1–8

Lecture notes by Dennis Christensen

November 16, 2023

Preamble

These notes will accompany the course STK4021 on Applied Bayesian Analysis for the autumn semester of 2023. They are largely self-contained, except with regards to solutions to exercises and coding examples, which will be covered in lectures. I advise you to read ahead in the notes before lectures - in that way the material will be more easily digestible. Most of the exercises are taken from Professor Nils Lid Hjort's lecture notes from previous years (available on the course website), or from relevant textbook sources (cited at the beginning of chapter if relevant). Although the lecture notes will cover all the material we need, it is of course possible to look for further reading and challenges in said sources.

If you have any questions about the material or would like to notify me about errors in the notes, please reach out via email: `dennis.christensen@ffi.no`

A note on notation

In previous statistics courses, your teachers have probably been consistent in using uppercase letters (like X, Y, Z) for random variables and lowercase letters (like x, y, z) for deterministic variables. In Bayesian statistics, the notation tends to be more sloppy, and we mostly use lowercase letters for everything, unless it is really important to separate between random and deterministic variables. Furthermore, introductory courses in statistics tend to have really precise notation with regards to probability density functions, writing for example $f_{X|Y=y}(x)$ for the density of X given that $Y = y$. In Bayesian statistics, we instead use a single letter π (or sometimes p) for every density, and let the argument of the function explain which density is referred to. For example, the density above would simply be written as $\pi(x \mid y)$. It will usually be clear from context which random variables are in play.

For common probability distributions, we will include an argument to indicate that we are referring to the density. For example, $y \sim \mathrm{N}(\mu, \sigma^2)$ will mean that y is normally distributed with mean μ and variance σ^2, but

$$\mathrm{N}(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

is the density of a $\mathrm{N}(\mu, \sigma^2)$ distribution, evaluated at y.

Contents

1 The Bayesian pipeline

1.1 Introduction

Prior and posterior distributions

This is a course on *Bayesian statistics*, which is one of the main paradigms of mathematical statistics. Loosely speaking, if we were to summarise the Bayesian way of thinking in a single phrase, it might be that "every unknown quantity is treated as a random variable." So, let $\theta \in \Theta$ be some unknown quantity that we care about, living in a parameter space Θ. We assign a probability distribution to θ, say $\pi(\theta)$. This is called the *prior* distribution of θ, since we choose it before having observed any data. It reflects our prior beliefs. In most cases, $\pi(\theta)$ is interpreted as a density, so that

$$\mathbb{P}(\theta \in A) = \int_A \pi(\theta)\,\mathrm{d}\theta.$$

We assume that the parameter θ governs some observable process from the real world. Thus, we observe data $y = (y_1, \ldots, y_n) \in \mathcal{Y}$, in some data space $\mathcal{Y}$. Here, each observed y_i may itself be multidimensional, and y is the collection of all the data. For example, if each $y_i \in \mathbb{R}^3$ is a three-dimensional vector, then $\mathcal{Y} = \mathbb{R}^{3\times n}$. We model the data y using the model's *likelihood function* $\pi(y \mid \theta)$. Note that we condition on θ, since θ is treated as an unknown parameter of the distribution $\pi(y \mid \theta)$. If the y_i are conditionally independent given θ, then we can write

$$\pi(y \mid \theta) = \prod_{i=1}^{n} \pi(y_i \mid \theta).$$

Having observed the data y, the key question for the Bayesian statistician is: *How have our beliefs about θ changed?* Since we have observed data, we have gained new information, and we thus need to update our beliefs about θ accordingly. That is, we need to find the distribution of θ *given the data*. Mathematically, we want to find $\pi(\theta \mid y)$, the *posterior* distribution. By Bayes' theorem,

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)}. \tag{1.1}$$

Let us break down (1.1) term by term so that we make sure we understand it fully. We have already seen two of the terms on the right hand side, namely the prior $\pi(\theta)$ and the likelihood $\pi(y \mid \theta)$. The term in the denominator is called the *marginal likelihood*, and can also be written as

$$\pi(y) = \int_\Theta \pi(y \mid \theta')\pi(\theta')\,\mathrm{d}\theta'. \tag{1.2}$$

Combining (1.1) and (1.2), we get

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\int_\Theta \pi(y \mid \theta')\pi(\theta')\,\mathrm{d}\theta'}, \tag{1.3}$$

which shows that the posterior distribution is entirely determined by the prior $\pi(\theta)$ and the likelihood $\pi(y \mid \theta)$. In other words, the inferential pipeline only depends on the choice of prior and the choice of likelihood. For this reason, when we refer to a *model* in Bayesian statistics, we mean the choice of a prior distribution and a likelihood for the data.

Exercise 1. Verify relation (1.2).

Posterior mean and variance

Once we have calculated the posterior distribution $\pi(\theta \mid y)$, we are usually interested in investigating the behaviour of θ in the posterior by, for example, calculating the posterior mean and variance. These are given by

$$\mathbb{E}[\theta \mid y] = \int_\Theta \theta \pi(\theta \mid y)\, \mathrm{d}\theta, \tag{1.4}$$

$$\mathrm{Var}(\theta \mid y) = \mathbb{E}[(\theta - \mathbb{E}[\theta \mid y])^2 \mid y] = \mathbb{E}[\theta^2 \mid y] - \mathbb{E}[\theta \mid y]^2. \tag{1.5}$$

Exercise 2. Verify relation (1.5).

Exercise 3 (Based on Nils Lid Hjort's exercises, #1)**.** This exercise illustrates the basic prior to posterior updating mechanism for Poisson data.

(a) First make sure that you are reasonably acquainted with the Gamma distribution. We say that $Z \sim \mathrm{Gamma}(a, b)$ if its density is

$$g(z; a, b) = \frac{b^a}{\Gamma(a)} z^{a-1} \exp(-bz) \quad \text{on } (0, \infty).$$

Here a and b are positive parameters. Show that

$$\mathbb{E}[Z] = \frac{a}{b}, \qquad \mathrm{Var}(Z) = \frac{a}{b^2} = \frac{\mathbb{E}[Z]}{b}.$$

In particular, low and high values of b signify high and low variability, respectively.

(b) Now suppose $y \mid \theta$ is a Poisson with parameter θ, and that θ has the prior distribution $\mathrm{Gamma}(a, b)$. Show that $\theta \mid y \sim \mathrm{Gamma}(a + y, b + 1)$.

(c) Then suppose there are repeated Poisson observations $y_1, \ldots, y_n$, being iid $\sim \mathrm{Poisson}(\theta)$ for given θ. Use the above result repeatedly, e.g. interpreting $\pi(\theta \mid y_1)$ as the new prior before observing y_2, etc., to show that

$$\theta \mid y_1, \ldots, y_n \sim \mathrm{Gamma}(a + y_1 + \cdots + y_n, b + n).$$

Also derive this result directly, i.e. without necessarily thinking about the data having emerged sequentially.

(d) Suppose the prior used is a rather flat $\mathrm{Gamma}(0.1, 0.1)$ and that the Poisson data are 6, 8, 7, 6, 7, 4, 11, 8, 6, 3. Plot the ten curves $\pi(\theta \mid y_1, \ldots, y_j)$ $(j = 1, \ldots, 10)$, along with the prior density $\pi(\theta)$. Also compute the ten Bayes estimates $\hat{\theta}_j = \mathbb{E}[\theta \mid y_1, \ldots, y_j]$ and the posterior standard deviations, for $j = 0, \ldots, 10$.

(e) The mathematics turned out to be rather uncomplicated in this situation, since the Gamma continuous density matches the Poisson discrete density so nicely. Suppose instead that the initial prior for θ is a uniform over $[0.5, 50]$. Try to compute posterior distributions, posterior means and posterior standard deviations also in this case, and compare with what you found above.

Solution.

(b) This exercise can be solved directly by applying Bayes' theorem directly, but we shall use a key trick known as the *normalisation trick*, or the *functional form trick*. Note that in Bayes theorem, the marginal likelihood $\pi(y)$ does not depend on θ. Thus, it merely serves as a normalisation constant, to make sure the right hand side integrates to unity. Therefore, if we recognise the functional form of the product $\pi(y \mid \theta)\pi(\theta)$, we can ignore the marginal likelihood $\pi(y)$ when deriving the posterior distribution. In this particular exercise, we have

$$\pi(y \mid \theta) = \frac{\theta^y \exp(-\theta)}{y!},$$

and so with the Gamma prior, we have, as a function of θ,

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta) \propto \theta^y \exp(-\theta) \times \theta^{a-1} \exp(-b\theta) = \theta^{a+y-1} \exp(-(b+1)\theta).$$

Here, the symbol $\propto$ means "is proportional to", and we have ignored all factors not depending on θ. Now, we see that our answer is of the same *functional form* as the Gamma density $g(\theta; a+y, b+1)$. Since we know $\pi(\theta \mid y)$ has to integrate to unity, we do not have to work out what the normalisation constant is. The normalisation *forces* $\pi(\theta \mid y) = g(\theta; a+y, b+1)$. We thus conclude that $\theta \mid y \sim \text{Gamma}(a+y, b+1)$. There is no other option.

The previous exercise introduced the key idea of the normalisation trick, the main idea behind which is to treat the marginal likelihood $\pi(y)$ as a normalisation constant. The trick only works if we are able to recognise the functional form of the product $\pi(y \mid \theta)\pi(\theta)$, which may not always be the case. However, when it works, the trick yields yet another interpretation of Bayes' theorem, namely

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta), \tag{1.6}$$

which is useful to keep in mind, in some cases. Phrased in words, we have

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}, \tag{1.7}$$

the simplest summary of the Bayesian inferential pipeline.

The predictive distribution

In many applications, we are not merely interested in the posterior distribution $\pi(\theta \mid y)$, but also in the *predictive* distribution of the next observation. Assume that, conditioned on θ, the observations $y_1, \ldots, y_n$ are iid (independent and identically distributed), and that we observe a new data point

y_{n+1} from the same model. What is the distribution of y_{n+1}, given the data $y = (y_1, \ldots, y_n)$? We have

$$\pi(y_{n+1} \mid y) = \int_\Theta \pi(y_{n+1} \mid \theta)\pi(\theta \mid y)\, \mathrm{d}\theta, \tag{1.8}$$

so, in principle, we can work out the predictive distribution if we know the posterior $\pi(\theta \mid y)$ and the single-observation likelihood $\pi(y_{n+1} \mid \theta)$.

Exercise 4. Verify relation (1.8). What happens here if $n = 0$ (so $y = \emptyset$)?

Exercise 5. The two statisticians Alice and Bob want to predict the probability of the sun rising tomorrow. Alice is a frequentist, whilst Bob is a Bayesian. They let $y_1, \ldots, y_n \in \{0, 1\}$ be binary variables indicating whether a sunrise occurred for each day $i = 1, \ldots, n$. Since the earth is 4.543×10^9 years old, we have that $n = 365 \times 4.543 \times 10^9 = 1.658195 \times 10^{12}$. Since the sun has risen every day, we have that $y_i = 1$ for each $i = 1, \ldots, n$. Both Alice and Bob assume that the sunrises are independent, and that the probability of any particular sunrise occurring on a given day is a Bernoulli random variable with parameter $\theta \in [0, 1]$, so that the probability mass function (pmf) for each y_i is $\pi(y_i \mid \theta) = \theta^{y_i}(1 - \theta)^{1-y_i}$.

(a) As a frequentist, Alice uses maximum likelihood theory to estimate θ. What does she find?

(b) Bob, on the other hand, is a Bayesian statistician, and so assigns a prior distribution to θ. As a simple choice, he chooses the uniform distribution, so $\pi(\theta) = 1$ for $0 \leq \theta \leq 1$, zero otherwise. Given the data, what is the posterior distribution $\pi(\theta \mid y)$?

(c) Calculate the posterior mean $\mathbb{E}[\theta \mid y]$ and the posterior variance $\mathrm{Var}(\theta \mid y)$. What happens as $n \to \infty$?

(d) What is the predictive probability that the sun will rise tomorrow, $\mathbb{P}(y_{n+1} = 1 \mid y)$?

(e) Verify the answer you got in point (c) via Monte Carlo simulations, but with a smaller value of n, say $n = 14$.

1.2 Bayesian decision theory

For a thorough treatment of this topic see Robert (2001, Chapter 2)

An important application of Bayesian analysis is *decision theory*, which is the study of how to take the optimal action based on given information. In addition to the setup we have established so far, we also need to consider different *actions*. Mathematically, we say that a *decision function* is a function $a : \mathcal{Y} \to \mathcal{A}$, mapping the data y to some action $a(y)$, living in a suitable action space $\mathcal{A}$. The action $a(y)$ could be a number (a statistical estimate, how much money we should gamble, etc), or a binary decision (implement policy versus do not implement policy), or generally something else entirely.

In order to determine how good a given action is, we choose a *loss function* $L : \Theta \times \mathcal{A} \to [0, \infty)$. This will return a low value if $a(y)$ is a good action, and a high value otherwise.

The loss function $L(\theta, a(y))$ evaluates the quality of the action $a(y)$ for a specific realisation of the data y. In order to evaluate the quality of the decision function a over all such realisations, we have the *frequentist risk*

$$R(a, \theta) = \int_{\mathcal{Y}} L(\theta, a(y))\pi(y \mid \theta)\, \mathrm{d}y.$$

However, as Bayesians, we can also marginalise out θ, yielding the *Bayes risk*

$$\mathrm{BR}(a) = \int_{\Theta} R(a, \theta)\pi(\theta)\, \mathrm{d}\theta = \int_{\Theta}\int_{\mathcal{Y}} L(\theta, a(y))\pi(y \mid \theta)\, \mathrm{d}y\, \pi(\theta)\, \mathrm{d}\theta. \tag{1.9}$$

Our goal as Bayesians is to minimise the Bayes risk. That is, to find the optimal decision function $a^\star$ satisfying

$$a^\star = \arg\min_a \mathrm{BR}(a).$$

In settings where a is an estimator, $a^\star$ is called the *Bayes estimator*. A priori, it seems difficult to actually compute $a^\star$, since we have to minimise the double integral (1.9). However, it turns out that we can work with a much simpler expression, which can be minimised pointwise. We define the *posterior expected loss* as

$$\rho(a \mid y) = \int_{\Theta} L(\theta, a(y))\pi(\theta \mid y)\, \mathrm{d}\theta. \tag{1.10}$$

We then have the following useful result.

Theorem 1.1. *The minimiser $a^\star$ of the Bayes risk is precisely the decision function which minimises the posterior expected loss for each y. That is,*

$$a^\star(y) = \arg\min_\alpha \int_{\Theta} L(\theta, \alpha)\pi(\theta \mid y)\, \mathrm{d}\theta. \tag{1.11}$$

The value of this theorem is that (1.11) is a much easier minimisation problem, as we shall see in examples and exercises.

The theorem also tells us that once we have chosen a prior, a likelihood and a loss function, the Bayes estimator is uniquely determined. So in the real world, if our policy makers wants us to calculate the optimal action to take, we need to know three pieces of information: what is their prior, what is their likelihood and what is their loss function? Once we know all three, there is a unique optimal action to take based on the observed data.

Exercise 6. Let $\theta \in \mathbb{R}$. Given the loss function $L(\theta, a) = (a - \theta)^2$, show that the Bayes estimator is the posterior mean, $\mathbb{E}[\theta \mid y]$.

Sometimes we come across exercises or problems which simply state "find the Bayes estimator". Technically speaking, this is not a well-defined question, since the Bayes estimator depends on the choice of loss function. However, in cases in which no loss function is specified, it is implicitly assumed that we are working with the quadratic loss, and so the Bayes estimator simply means the posterior mean.

Exercise 7. Now consider the loss function $L(\theta, a) = |a - \theta|$. What is the Bayes estimator?

Exercise 8. Prove Theorem 1.1. *Hint: Use Fubini's Theorem.*

Exercise 9 (Based on Nils Lid Hjort's exercises, #4)**.** A prototype normal mean model is the simple one with a single observation $y \sim \mathrm{N}(\theta, 1)$. We let the loss function be squared error, $L(\theta, a) = (a - \theta)^2$.

(a) Show that the maximum likelihood solution is simply $\theta^* = y$. Show that its risk function is $R(\theta^\star, \theta) = 1$, i.e. constant.

(b) Let θ have the prior $\mathrm{N}(0, \tau^2)$. Show that (θ, y) is binormal, and that $\theta \mid y \sim \mathrm{N}(\rho y, \rho)$, with $\rho = \tau^2/(\tau^2 + 1)$. In particular, $\hat{\theta}_B(y) = \rho y$ is the Bayes estimator.

(c) Find the risk function for the Bayes estimator, and identify where it is smaller than that of the maximum likelihood solution, and where it is larger. Comment on the situation where τ is small (and hence ρ), as well as on the case of τ being big (and hence ρ close to 1).

(d) Show that the minimal Bayes risk for the prior $\mathrm{N}(0, \tau^2)$ is $\rho = \tau^2/(\tau^2 + 1)$.

Exercise 10 (Based on Nils Lid Hjort's exercises, #5)**.** Let $y \mid \theta$ be a Poisson with mean parameter θ, which is is to be estimated with weighted squared error loss $L(\theta, t) = (t - \theta)^2/\theta$.

(a) Show that the maximum likelihood estimator is y itself, and that its risk function is the constant 1.

(b) Consider the prior distribution $\mathrm{Gamma}(a, b)$ for θ. Show that $\mathbb{E}[\theta] = a/b$ and that $\mathbb{E}[\theta^{-1}] = b/(a - 1)$ if $a > 1$, and infinite if $a \leq 1$.

(c) Show that $\theta \mid y$ is a $\mathrm{Gamma}(a + y, b + 1)$, from which follows

$$\mathbb{E}[\theta \mid y] = \frac{a + y}{b + 1}, \qquad \mathbb{E}[\theta^{-1} \mid y] = \frac{b + 1}{a - 1 + y}.$$

The latter formula holds if $a - 1 + y > 0$, which means for all y if $a \geq 1$, but care is needed if $a < 1$ and $y = 0$. Show that the Bayes solution is

$$\hat{\theta} = \frac{a - 1 + y}{b + 1} \quad \text{for all } y \geq 0,$$

provided $a \geq 1$, but that we need the more careful formula

$$\hat{\theta} = \begin{cases} (a - 1 + y)/(b + 1) & \text{if y} \geq 1, \\ 0 & \text{if y} = 0, \end{cases}$$

in the case of $a < 1$.

(d) Taking care of the simplest case $a > 1$ first, show that the minimal Bayes risk is $1/(b + 1)$ for the Gamma prior.

(e) Show that

$$\mathbb{E}[L(\theta, \hat{\theta}) \mid y] = \begin{cases} 1/(b+1) & \text{if y} \geq 1, \\ a/(b+1) & \text{if y} = 0. \end{cases}$$

Deduce from this a minimum Bayes risk formula also for the case of $a < 1$:

$$\frac{1}{b+1}\left\{1 - \left(\frac{b}{b+1}\right)^a\right\} + \frac{a}{b+1}\left(\frac{b}{b+1}\right)^a.$$

2 Choosing the prior distribution

Some of the remarks made here build on Nicholls (2023).

We have worked through quite a few examples of Bayesian inference, and have seen numerous examples of how the choice of prior distribution affects our analysis. A key question for Bayesian statisticians is how to choose a prior distribution. This section is meant to provide some general points to keep in mind for prior elicitation, as well as some standard recipes.

The question of choosing a prior is by no means an exact science, but there are some general points to keep in mind. Some key points are as follows.

1. **Domain expertise.** Sometimes we wish to model phenomena from the real world, where domain experts have well-founded beliefs about the parameters θ based on their knowledge of the subject. This should be incorporated into our prior distribution $\pi(\theta)$.

2. **Interpretation.** In some applications some function $g(\theta)$ of the parameters has a clear interpretation. Thus, our prior distribution for θ should yield realistic values of $g(\theta)$. Often, this step requires simulating lots of samples θ from the prior, calculating $g(\theta)$ for each one.

3. **Key hypothesis**. In many applications, we wish to investigate a specific hypothesis about the parameters θ, and so our prior should be non-informative with respect to this hypothesis. For example, if $\theta \in [0, 1]$, and the key hypothesis to investigate is whether $\theta > 0.99$, then a uniform prior $\theta \sim \text{Uniform}[0, 1]$ would be highly informative with respect to this hypothesis.

4. **Multiple priors.** To investigate how robust our results are with respect to the choice of prior, it is common to repeat the analysis with different choices of priors. We did this in exercise 3 (e). In general, different priors will yield the same posterior analysis once we have gathered enough data (by the Bernstein-von Mises theorem, which we will come back to later in the course), but if we have little data, the choice of prior may drastically affect our analysis.

2.1 Conjugate priors

In most of the examples we have seen so far, the prior and posterior distributions were of the same functional form. For example, in exercise 3, we started with a Gamma prior, and ended up with a Gamma posterior. This is an example of a *conjugate prior*. Mathematically speaking, a conjugate

prior is any prior distribution which is of the same functional form as the likelihood $\pi(y \mid \theta)$, thought of as a function of θ.

Exercise 11. This exercise introduces the binomial distribution, and looks at conjugate priors.

(a) Let $y = (y_1, \ldots, y_n)$, where all the y_i are iid Bernoulli variables with parameter θ, with density

$$f_\theta(y_i) = \theta^{y_i}(1-\theta)^{1-y_i},$$

for $y_i \in \{0, 1\}$. Verify that $\mathbb{E}[y_i] = \theta$ and that $\mathrm{Var}(y_i) = \theta(1-\theta)$.

(b) Write down the log-likelihood for the full data y and verify that the maximum likelihood estimator $\hat{\theta}$ for θ is the sample mean.

(c) Now let m be the number of observations with $y_i = 1$. In other words, $m = \sum_{i=1}^n y_i$, a sum of independent Bernoulli trials. We know that m follows the Binomial distribution, $m \mid \theta \sim \mathrm{Binomial}(\theta, \mathrm{n})$, with density

$$f_\theta(m) = \binom{n}{m}\theta^m(1-\theta)^{n-m},$$

for $m = 0, 1, \ldots, n$.

Verify that $\mathbb{E}[m] = \theta n$ and $\mathrm{Var}(m) = \theta(1-\theta)n$.

(d) Now verify that the conjugate prior for θ is the Beta distribution $\mathrm{Beta}(\mathrm{a}, \mathrm{b})$, with density

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},$$

where $a, b > 0$. Show also that

$$\mathbb{E}[\theta] = \frac{a}{a+b}, \qquad \mathrm{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}.$$

(e) (Based on Bishop (2006, chapter 2)) Although the Beta prior and the binomial likelihood clearly share the same functional form as a function of θ, it is less obvious where the normalisation constant in the Beta distribution comes from. In this exercise, we verify that this normalisation constant is correct. We need to show that

$$\int_0^1 s^{a-1}(1-s)^{b-1}\,\mathrm{d}s = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

where we recall the definition of the Gamma function,

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\,\mathrm{d}x.$$

Now, first note that

$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1}\exp(-x)\,\mathrm{d}x \int_0^\infty y^{b-1}\exp(-y)\,\mathrm{d}y$$
$$= \int_0^\infty x^{a-1}\left(\int_0^\infty y^{b-1}\exp(-(x+y))\,\mathrm{d}y\right)\mathrm{d}x.$$

Prove the normalisation by first substituting $t = y + x$ (holding x fixed), then change the order of integration between x and t, and finally make the substitution $x = ts$ (holding t fixed).

(f) Find the posterior distribution $\pi(\theta \mid m)$. Show that we can write

$$\mathbb{E}[\theta \mid m] = \lambda \mathbb{E}[\theta] + (1 - \lambda)\hat{\theta}$$

for some $0 \leq \lambda \leq 1$ which you should identify. What happens to $\mathbb{E}[\theta \mid m]$ as the number of observations goes to infinity?

Exercise 12 (Based on Nils Lid Hjort's exercises, # 13)**.** The Beta-binomial model, with a Beta distribution for the binomial probability parameter, is on the 'Nice List' where the Bayesian machinery works particularly well: Prior elicitation is easy, as is the updating mechanism. This exercise concerns the generalisation to the Dirichlet-multinomial model, which is certainly also on the Nice List and indeed in broad and frequent use for a number of statistical analyses.

(a) Let $(y_1, \ldots, y_m)$ be the count vector associated with n independent experiments having m different outcomes $A_1, \ldots, A_m$. In other words, y_j is the number of events of type A_j, for $j = 1, \ldots, m$. Show that if the vector of $\mathbb{P}(A_j) = p_j$ is constant across the n independent experiments, then the probability distribution governing the count data is

$$f(y_1, \ldots, y_m) = \frac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m},$$

for $y_1 \geq 0, \ldots, y_m \geq 0, y_1 + \cdots + y_m = n$. This is the multinomial model. Explain how it generalises the binomial model.

(b) Show that

$$\mathbb{E}\, Y_j = np_j, \quad \operatorname{Var} Y_j = np_j(1 - p_j), \quad \operatorname{cov}(Y_j, Y_k) = -np_j p_k \text{ for } j \neq k.$$

(c) Now define the Dirichlet distribution over m cells with parameters $(a_1, \ldots, a_m)$ as having probability density

$$\pi(p_1, \ldots, p_{m-1}) = \frac{\Gamma(a_1 + \cdots + a_m)}{\Gamma(a_1) \cdots \Gamma(a_m)} p_1^{a_1 - 1} \cdots p_{m-1}^{a_{m-1} - 1} (1 - p_1 - \cdots - p_{m-1})^{a_m - 1},$$

over the simplex where each $p_j \geq 0$ and $p_1 + \cdots + p_{m-1} \leq 1$. Of course we may choose to write this as

$$\pi(p_1, \ldots, p_{m-1}) \propto p_1^{a_1 - 1} \cdots p_{m-1}^{a_{m-1} - 1} p_m^{a_m - 1},$$

with $p_m = 1 - p_1 - \cdots - p_{m-1}$; the point is however that there are only $m - 1$ unknown parameters in the model as one knows the mth once one learns the values of the other $m - 1$. Show that the marginals are Beta distributed,

$$p_j \sim \text{Beta}(a_j, a - a_j) \quad \text{where } a = a_1 + \cdots + a_m.$$

(d) Infer from this that

$$\mathbb{E}\, p_j = p_{0,j} \qquad \mathrm{Var}\, p_j = \frac{1}{a+1} p_{0,j}(1-p_{0,j}),$$

in terms of $a_j = a p_{0,j}$. Show also that

$$\mathrm{cov}(p_j, p_k) = -\frac{1}{a+1} p_{0,j} p_{0,k} \quad \text{for } j \neq k.$$

For the 'flat Dirichlet', with parameters $(1, \ldots, 1)$ and prior density $(m-1)!$ over the simplex, find the means, variances, covariances.

(e) Now for the basic Bayesian updating result. When $(p_1, \ldots, p_m)$ has a $\mathrm{Dir}(a_1, \ldots, a_m)$ prior, then, given the multinomial data $y = (y_1, \ldots, y_m)$, show that

$$(p_1, \ldots, p_m) \mid y \sim \mathrm{Dir}(a_1 + y_1, \ldots, a_m + y_m).$$

Give formulae for the posterior means, variances, and covariances. In particular, explain why

$$\hat{p}_j = \frac{a_j + y_j}{a + n}$$

is a natural Bayes estimate of the unknown p_j. Also find an expression for the posterior standard deviation of the p_j.

(f) In order to carry out easy and flexible Bayesian inference for $p_1, \ldots, p_m$ given observed counts $y_1, \ldots, y_m$, one needs a recipe for simulating from the Dirichlet distribution. One such is as follows: Let $X_1, \ldots, X_m$ be independent with $X_j \sim \mathrm{Gamma}(a_j, 1)$ for $j = 1, \ldots, m$. Then the ratios

$$Z_1 = \frac{X_1}{X_1 + \cdots + X_m}, \ldots, Z_m = \frac{X_m}{X_1 + \cdots + X_m}$$

are in fact $\mathrm{Dir}(a_1, \ldots, a_m)$. Try to show this from the transformation law for probability distributions: If X has density $f(x)$, and $Z = h(X)$ is a one-to-one transformation with inverse $X = h^{-1}(Z)$, then the density of Z is

$$g(z) = f(h^{-1}(z)) \left| \frac{\partial h^{-1}(z)}{\partial z} \right|$$

(featuring the determinant of the Jacobian of the transformation). Use in fact this theorem to find the joint distribution of $(Z_1, \ldots, Z_{m-1}, S)$, where $S = X_1 + \cdots + X_m$ (one discovers that the Dirichlet vector of Z_j is independent of their sum S).

(g) The Dirichlet distribution has a nice 'collapsibility' property: If say $(p_1, \ldots, p_8)$ is $\mathrm{Dir}(a_1, \ldots, a_8)$, show that then the collapsed vector $(p_1 + p_2, p_3 + p_4 + p_5, p_6, p_7 + p_8)$ is $\mathrm{Dir}(a_1 + a_2, a_3 + a_4 + a_5, a_6, a_7 + a_8)$.

Exercise 13 (Based on Nils Lid Hjort's exercises, # 14)**. Gott würfelt nicht**, but I do so, on demand. I throw a certain moderately strange-looking die 30 times and have counts $(2, 5, 3, 7, 5, 8)$ of outcomes 1, 2, 3, 4, 5, 6.

(a) Use either of the priors

- 'flat', $\mathrm{Dir}(1,1,1,1,1,1)$,
- 'symmetric but more confident', $\mathrm{Dir}(3,3,3,3,3,3)$,
- 'unwilling to guess', $\mathrm{Dir}(0.1,0.1,0.1,0.1,0.1,0.1)$

for the probabilities $(p_1,\ldots,p_6)$ to assess the posterior distribution of each of the following quantities:

$$\begin{aligned}\rho &= p_6/p_1,\\ \alpha &= (1/6)\sum_{j=1}^{6}(p_j-1/6)^2,\\ \beta &= (1/6)\sum_{j=1}^{6}|p_j-1/6|,\\ \gamma &= (p_4p_5p_6)^{1/3}/(p_1p_2p_3)^{1/3}.\end{aligned}$$

(b) The above priors are slightly artificial in this context, since they do not allow the explicit possibility that the die in question is plain boring utterly simply a correct one, i.e. that $p=p_0=(1/6,\ldots,1/6)$. The priors used hence do not give us the possibility to admit that ok, then, perhaps $\rho=1$, $\alpha=0$, $\beta=0$, $\gamma=1$, after all. This motivates using a mixture prior which allows a positive chance for $p=p_0$. Please therefore redo the Bayesian analysis above, with the same $(2,5,3,7,5,8)$ data, for the prior $\frac{1}{2}\delta(p_0)+\frac{1}{2}\mathrm{Dir}(1,1,1,1,1,1)$. Here $\delta(p_0)$ is the 'degenerate prior' that puts unit point mass at position p_0. Compute in particular the posterior probability that $p=p_0$, and display the posterior distributions of ρ,α,β,γ.

2.1.1 The multivariate Gaussian distribution

This section largely builds on Bishop (2006, Chapter 2).

A key object of study in statistics is the multivariate Gaussian distribution. As we shall see, this distribution has some very nice properties, from a Bayesian point of view. The p-dimensional Gaussian distribution has density

$$\pi(x;\mu,\Sigma)=\frac{1}{(2\pi)^{p/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right\}, \tag{2.1}$$

where $\mu\in\mathbb{R}^p$ and Σ is a $p\times p$ symmetric positive definite matrix. We write $x\sim\mathrm{N}(\mu,\Sigma)$ if x has this density.

Exercise 14.

(a) Show that Σ being symmetric is not really a restriction, in the sense that if someone proposes a quadratic function $(x-\mu)^\top A^{-1}(x-\mu)$, where A is not necessarily symmetric, there exists a symmetric matrix Σ such that

$$(x-\mu)^\top A^{-1}(x-\mu)=(x-\mu)^\top\Sigma^{-1}(x-\mu).$$

(b) Let A be an $n \times n$ matrix. Show that $x^\top Ax = 0$ for all $x \in \mathbb{R}^n$ if and only if $A + A^\top = 0$.

(c) Use the result from (b) to show that the matrix Σ you found in part (a) is in fact unique.

Exercise 15. Using the spectral theorem for real symmetric matrices, show that we can write

$$\Sigma = \sum_{i=1}^{p} \lambda_i u_i u_i^\top,$$

where $\lambda_i \in \mathbb{R}$ are the eigenvalues of Σ and $u_i \in \mathbb{R}^p$ are the corresponding eigenvectors, with $u_i^\top u_j = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

Explain why $\lambda_i \neq 0$ for all i.

Exercise 16.

(a) Define the quadratic function in the exponential by

$$\Delta^2 = (x - \mu)^\top \Sigma^{-1} (x - \mu).$$

Show that we can write

$$\Delta^2 = \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i},$$

where

$$y_i = u_i^\top (x - \mu).$$

(b) The coordinate transformation above may be written as

$$y = U^\top (x - \mu),$$

where U is the $p \times p$ matrix whose columns are given by u_i. Use the transformation law to write down the density of y, and show that the y_i are independent Gaussian random variables. Deduce from this that the normalisation constant in (2.1) is correct.

(c) By making the transformation of variables $z = x - \mu$, show from first principles that $\mathbb{E}[x] = \mu$.

(d) Show that we can write

$$z = \sum_{i=1}^{p} y_i u_i.$$

(e) Use (d) to show that

$$\mathbb{E}[xx^\top] = \mu\mu^\top + \Sigma.$$

Deduce from this that

$$\mathrm{cov}(x) = \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^\top] = \Sigma.$$

Conditional Gaussian distributions

We shall now prove two very important properties of the Gaussian distribution, namely that if two vectors x_a and x_b are jointly Gaussian, then the conditional $x_a \mid x_b$ and the marginals x_a and x_b are also Gaussian. We start by introducing some notation, let

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$$

be a vector whose first k components are x_a and whose last $p-k$ components are x_b. Similarly, introduce

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}.$$

Note that $\Sigma_{aa}^\top = \Sigma_{aa}, \Sigma_{bb}^\top = \Sigma_{bb}$ and $\Sigma_{ab}^\top = \Sigma_{ba}$ since Σ is symmetric.
It turns out to be useful to decompose the inverse of Σ in a similar way, so let $\Lambda = \Sigma^{-1}$, and decompose Λ as

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}.$$

Note that in general, $\Sigma_{aa}^{-1} \neq \Lambda_{aa}$, and so on. The inverse Λ of the covariance matrix Σ is called the *precision* matrix, and is sometimes easier to work with than Σ.
Now, the quadratic function Δ^2 decomposes as

$$-\frac{1}{2}\Delta^2 = -\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) = -\frac{1}{2}\begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}^\top \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix} =$$
$$-\frac{1}{2}(x_a-\mu_a)^\top \Lambda_{aa}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^\top \Lambda_{ab}(x_b-\mu_b) - \frac{1}{2}(x_b-\mu_b)^\top \Lambda_{ba}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^\top \Lambda_{bb}(x_b-\mu_b).$$

To prove that the conditional distribution of $x_a \mid x_b$ is Gaussian, we need to show that $\pi(x_a \mid x_b)$ is of the functional form as a Gaussian density. That is, we need to show that

$$\pi(x_a \mid x_b) \propto \exp\left\{\text{quadratic function in } x_a\right\}.$$

Equivalently, we need to show that

$$\log \pi(x_a \mid x_b) = \left\{\text{quadratic function in } x_a\right\} + \text{constant},$$

where the constant does not depend on x_a. Now, $\log \pi(x_a \mid x_b) = \log \pi(x) - \log \pi(x_b)$, so, since $\log \pi(x_b)$ does not depend on x_a, we need to write Δ^2 as a quadratic in x_a, plus a constant. We do this by completing the square. From the decomposition above, we can write

$$-\frac{1}{2}\Delta^2 = -\frac{1}{2}x_a^\top \Lambda_{aa} x_a + x_a^\top \left\{\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\right\} + \text{constant}.$$

Completing the square, we get

$$-\frac{1}{2}\Delta^2 = -\frac{1}{2}(x_a - (\left\{\mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)\right\})^\top \Lambda_{aa}(x_a - (\left\{\mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)\right\}) + \text{constant}.$$

Now, we have shown that, as function of x_a

$$\begin{aligned}\pi(x_a \mid x_b) &\propto \pi(x) \\ &\propto \exp\left\{-\frac{1}{2}(x_a - (\{\mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)\})^\top \Lambda_{aa}(x_a - (\{\mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)\})\right\},\end{aligned}$$

and so by functional form,

$$x_a \mid x_b \sim \mathrm{N}(\mu_{a|b}, \Sigma_{a|b}),$$

where

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b), \quad \Sigma_{a|b} = \Lambda_{aa}^{-1}.$$

Exercise 17. The Woodbury matrix identity says that the inverse of a block matrix can be calculated as

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix},$$

where

$$M = (A - BD^{-1}C)^{-1}.$$

Use this identity to express Λ_{aa} and Λ_{ab} in terms of $\Sigma_{aa}, \Sigma_{ab}, \Sigma_{ba}$ and Σ_{bb}. Finally, deduce that

$$\begin{aligned}\mu_{a|b} &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \\ \Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}.\end{aligned}$$

Marginal Gaussian distributions

Our next is to show that the marginal distribution x_a is also Gaussian, and to derive its mean vector and covariance matrix. Now, we want to marginalise

$$\pi(x_a) = \int \pi(x_a, x_b)\,\mathrm{d}x_b.$$

In order to evaluate this integral, we write Δ^2 as a quadratic function in x_b. We do this by completing the square, just like in the previous chapter. We have that

$$-\frac{1}{2}\Delta^2 = -\frac{1}{2}(x_b - \mu_{b|a})^\top \Lambda_{bb}(x_b - \mu_{b|a}) + \frac{1}{2}\mu_{b|a}^\top \Lambda_{bb}\mu_{b|a} - \frac{1}{2}(x_a - \mu_a)^\top \Lambda_{aa}(x_a - \mu_a) + (x_a - \mu_a)^\top \Lambda_{ab}\mu_b.$$

Thus, when we integrate, the terms in the exponential which do not depend on x_b factor out. Thus, we are left to evaluate

$$\int \exp\left\{-\frac{1}{2}(x_b - \mu_{b|a})^\top \Lambda_{bb}(x_b - \mu_{b|a})\right\}\,\mathrm{d}x_b,$$

but this is just a non-normalised Gaussian integral, which we know is just proportional to the determinant of the covariance matrix. Therefore, it simply integrates out to a constant, independent of x_a.

To evaluate the distribution of x_a, we look at the remaining terms in the quadratic Δ^2 after having integrated out x_b, which are

$$\rho = \frac{1}{2}\mu_{b|a}^\top \Lambda_{bb}\mu_{b|a} - \frac{1}{2}(x_a - \mu_a)^\top \Lambda_{aa}(x_a - \mu_a) + (x_a - \mu_a)^\top \Lambda_{ab}\mu_b.$$

We now complete the square like before. We have

$$\begin{aligned}
\rho &= \frac{1}{2}(\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a))^\top \Lambda_{bb}(\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)) \\
&\qquad - \frac{1}{2}(x_a - \mu_a)^\top \Lambda_{aa}(x_a - \mu_a) + (x_a - \mu_a)^\top \Lambda_{ab}\mu_b \\
&= -\frac{1}{2}x_a^\top(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})x_a + x_a^\top(\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b - \Lambda_{ab}(\mu_b + \Lambda_{bb}^{-1}\Lambda_{ba}\mu_a) + \text{constant} \\
&= -\frac{1}{2}x_a^\top(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})x_a + x_a^\top(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a + \text{constant} \\
&= -\frac{1}{2}(x_a - \mu_a)^\top(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})(x_a - \mu_a) + \text{constant}.
\end{aligned}$$

So, by functional form, we conclude that $x_a \sim \mathrm{N}(\mu_a, \Sigma_a)$, where

$$\Sigma_a = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}.$$

Exercise 18. Use the Woodbury identity to show that in fact

$$\Sigma_a = \Sigma_{aa}.$$

Bayes' theorem for Gaussian variables

Suppose we have a Gaussian model for a single p-dimensional vector y of the form

$$y \mid \theta \sim \mathrm{N}(A\theta + b, L^{-1}), \tag{2.2}$$

so given θ, y has mean $A\theta + b$ and precision matrix L. Such models are commonly called *linear Gaussian models*, due to the linear dependence on θ in the mean.

Now impose a prior distribution on θ. It turns out that the conjugate prior for the Gaussian is also Gaussian, so

$$\theta \sim \mathrm{N}(\mu, \Lambda^{-1})$$

in the prior. Note that we specify a precision matrix rather than a covariance matrix, as this makes the analysis somewhat simpler. To show that the Gaussian prior is indeed conjugate, we first need the joint distribution of both y and θ. To simplify notation, let

$$z = \begin{pmatrix} \theta \\ y \end{pmatrix}.$$

As usual, we will derive the distribution of z by studying the quadratics in the exponents. We have

$$\begin{aligned}
\log \pi(z) &= \log \pi(\theta) + \log \pi(y \mid \theta) \\
&= -\frac{1}{2}(\theta - \mu)^\top \Lambda(\theta - \mu) - \frac{1}{2}(y - A\theta - b)^\top L(y - A\theta - b) + \text{constant}
\end{aligned}$$

Let us first identity the precision matrix by gathering all second order terms,

$$\text{2nd order terms} = -\frac{1}{2}\theta^\top \Lambda\theta - \frac{1}{2}y^\top Ly + y^\top LA\theta - \frac{1}{2}\theta^\top A^\top LA\theta$$
$$= -\frac{1}{2}\begin{pmatrix}\theta\\ y\end{pmatrix}^\top \begin{pmatrix}\Lambda + A^\top LA & -A^\top L\\ -LA & L\end{pmatrix}\begin{pmatrix}\theta\\ y\end{pmatrix} = -\frac{1}{2}z^\top Rz.$$

From this, we see that z has precision matrix R.

Exercise 19. Using Woodbury's matrix identity, show that the covariance matrix of z is

$$\text{cov}(z) = R^{-1} = \begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^\top\\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^\top\end{pmatrix}.$$

Next, we determine the mean vector of z by gathering linear terms and completing the square. We have

$$\text{1st order terms} = \theta^\top\Lambda\mu + y^\top Lb - \theta^\top A^\top Lb = \begin{pmatrix}\theta\\ y\end{pmatrix}^\top \begin{pmatrix}\Lambda\mu - A^\top Lb\\ Lb.\end{pmatrix}$$

Hence, completing the square, we get

$$\mathbb{E}[z] = R^{-1}\begin{pmatrix}\Lambda\mu - A^\top Lb\\ Lb\end{pmatrix} = \begin{pmatrix}\mu\\ A\mu + b,\end{pmatrix}$$

a very nice result, which makes intuitive sense.

Having worked out the expression for the mean and covariance of $z = (\theta, y)^\top$, we can now write down the joint distribution.

$$\begin{pmatrix}\theta\\ y\end{pmatrix} \sim \text{N}\left(\begin{pmatrix}\mu\\ A\mu + b\end{pmatrix}, \begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^\top\\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^\top\end{pmatrix}\right).$$

Exercise 20. Using the previous results for conditional and marginal Gaussian distributions, show that the posterior distribution is

$$\theta \mid y \sim \text{N}\left(\Sigma\{A^\top L(y-b) + \Lambda\mu\}, \Sigma\right), \tag{2.3}$$

where

$$\Sigma = (\Lambda + A^\top LA)^{-1}. \tag{2.4}$$

Show further that the marginal distribution of y is

$$y \sim \text{N}\left(A\mu + b, L^{-1} + A\Lambda^{-1}A^\top\right). \tag{2.5}$$

Exercise 21 (Based on Nils Lid Hjort's exercises, # 9)**.**

(a) How tall is Professor Hjort? Assume that the heights of Norwegian men above the age of twenty follows the normal distribution $\text{N}(\xi, \sigma^2)$, with $\xi = 180$ cm and $\sigma = 9$ cm. Thus, if you have *not yet seen* or bothered to notice this particular aspect of Professor Hjort and his lectures, your point estimate of his height ought to be $\xi = 180$ and a 95% prediction

interval for his height would be $\xi \pm 1.96\,\sigma$, or $[162.4, 197.6]$. – Assume now that you learn that his four brothers are actually 195 cm, 207 cm, 196 cm, 200 cm tall, and furthermore that correlations between brothers' heights in the population of Norwegian men is equal to $\rho = 0.80$. Use this information about his four brothers (still assuming that you have not noticed Professor Hjort's height) to revise your initial point estimate of Professor Hjort's height. Is he a five-percent statistical outlier in his family (i.e. outside the 95% prediction interval)?

(b) Assume Professor Hjort has n brothers (rather than merely four) and that you're learning their heights $X_1, \ldots, X_n$. What is the conditional distribution of Professor Hjort's height X_0, given this information? Represent this as a $\mathrm{N}(\xi_n, \sigma_n^2)$ distribution, with proper formulae for its parameters. How small is σ_n for a large number of brothers? (The point here is partly that even if you observe and measure his 99 brothers, there's still a limit to how much you can infer about Professor Hjort.)

Hint from Dennis: Use the Sherman-Morrison formula, which says that for any $n \times n$ matrix A and vectors u, v, we have

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}.$$

2.2 Empirical Bayes

We have gotten used to the idea of assigning a prior distribution to the unknown parameters θ. However, when we do so, we usually introduce a new set of parameters. Think back to exercise 3, where we had Poisson data y with parameter θ, and we assigned a prior $\mathrm{Gamma}(a, b)$ to θ, with its own parameters a and b.

The parameters of the prior distribution are sometimes called *hyperparameters*, as they are 'parameters governing the distribution of the parameters'. We usually have to make a choice of these parameters, perhaps guided by the remarks listed in the beginning of Section 2.

Another technique for choosing the hyperparameters is that of *empirical Bayes*, which instead estimates the hyperparameters directly from the data by maximising the marginal likelihood $\pi(y)$, as a function of the hyperparameters. Note that this approach slightly breaks with the pure Bayesian philisophy, which says that we should choose the prior distribution *before* we observe any data. Nevertheless, empirical Bayes is a popular technique amongst many statisticians and machine learning theorists, and has a surfeit of interesting and nice properties. In this section, we shall look at some basic examples of the technique. However, it will later show up as a useful tool when we look at Bayesian regression and classification.

Exercise 22.

(a) Go back to Exercise 3, with the Poisson-Gamma model. Show that the marginal likelihood $\pi(y)$ is given by

$$\pi(y) = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + n\bar{y})}{(b+n)^{a+n\bar{y}}} \prod_{i=1}^{n} \frac{1}{y_i!},$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$.

(b) Suppose that we know the value of a, say a_0. Show that the maximum (marginal) likelihood estimate $\hat{b}$ for b is given by

$$\hat{b} = \frac{a}{\bar{y}}.$$

Let $a_0 = 0.1$ like before and let y be the Poisson data from Exercise 3. What is $\hat{b}$ in this case?

(c) Suppose that we want to use a fully empirical Bayesian approach and maximise $\pi(y)$ as a function of both a and b (so $a = a_0$ is no longer fixed). What goes wrong?

Exercise 23 (Based on Nils Lid Hjort's exercises, # 24)**.** This exercise investigates the phenomenon known as the *Stein effect*, and its connections with empirical Bayes. Suppose there is an ensemble of parameters $\theta_1, \ldots, \theta_k$ to estimate, where these are thought to be not unreasonably dissimilar, and where it may make sense to think about them as having arisen from a distribution of parameter values. In such cases various empirical Bayes constructions will often be successful, in the sense that they lead to 'joint estimation' procedures that typically perform better than using 'separate estimation'. What *is and remains* surprising is that for certain situations of the above type, there are empirical Bayes methods that *always and uniformly* improve upon the 'separate estimation' procedures, i.e. even when the underlying parameters are widely dissimilar. This phenomenon is loosely referred to as 'the Stein effect', or even 'the Stein paradox', from influential papers by Charles Stein in 1956 and later. Even *The Scientific American* have had papers on this for a wider audience. The paradox in question is that when needing to estimate apples, oranges, bananas, then it is counterintuitively possible to do better by calling in information about oranges and bananas to estimate apples, etc.

The present exercise looks into one of these models where reasonably clean proofs may be given for the type of universal risk dominance of certain procedures over the standard ones. Let $Y_i \sim \mathrm{N}(\theta_i, 1)$ be independent for $i = 1, \ldots, k$, where the aim is to estimate each of the θ_i with a combined loss function

$$L(\theta, \hat{\theta}) = k^{-1} \sum_{i=1}^{k} (\hat{\theta}_i - \theta_i)^2.$$

The ensuing risk for the $\hat{\theta}$ procedure is

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta L(\theta, \hat{\theta}) = k^{-1} \sum_{i=1}^{k} \mathbb{E}_\theta (\hat{\theta}_i - \theta_i)^2.$$

This may again be represented as the average variance plus the average squared bias (as a function of the position in parameter space). Note that $\hat{\theta}_i$ for θ_i ought to be allowed to depend on all the data, not merely Y_i.

(a) The standard estimator here is simply using Y_i for θ_i, for $i = 1, \ldots, k$; Y_i is after all the least squares estimator, the maximum likelihood estimator, the best unbiased estimator, it is admissible, etc. Show that its risk function is simply 1, constant across the parameter space.

The challenge is to find an estimator which has risk function smaller than 1 everywhere in the parameter space.

(b) For a single $Y \sim \mathrm{N}(\theta, 1)$, show that under very mild conditions on the function $b(y)$, one has

$$\mathbb{E}_\theta (Y - \theta) b(Y) = \mathbb{E}_\theta b'(Y),$$

where $b'(y) = \mathrm{d}b(y)/\mathrm{d}y$. (Hint: Use integration by parts.) Check with e.g. $b(Y) = Y$ and $b(Y) = Y^2$ to get a feeling for how the identity works.

(c) Using the same technique, generalise the above to

$$\mathbb{E}_\theta (Y_i - \theta_i) b_i(Y) = \mathbb{E}_\theta b_{i,i}(Y),$$

where $b_{i,j}(y) = \partial b_i(y)/\partial y_j$.

(d) Consider a general competitor to Y of the form $\hat\theta_i = Y_i - b_i(Y)$. Show that

$$\mathbb{E}_\theta \{ (Y_i - b_i(Y) - \theta_i)^2 - (Y_i - \theta_i)^2 \} = \mathbb{E}_\theta \{ b_i(Y)^2 - 2 b_{i,i}(Y) \}$$

and hence that

$$R(\hat\theta, \theta) = R(Y, \theta) + E_\theta D(Y) = 1 + \mathbb{E}_\theta D(Y),$$

where

$$D(y) = k^{-1} \sum_{i=1}^k \{ b_i(y)^2 - 2 b_{i,i}(y) \}.$$

If in particular we manage to find $b_i(y)$ functions for which $D(y) < 0$ for all y, then $\hat\theta$ is a uniform improvement over the standard estimator Y. It turns out to be impossible to find such functions for $k = 1$ or $k = 2$, but indeed possible for $k \geq 3$.

(e) Try $b_i(y) = c y_i / \|y\|^2$, with $\|y\|^2$ being the squared Euclidean norm $\sum_{i=1}^k y_i^2$, corresponding to

$$\hat\theta = y - b(y) = \left(1 - \frac{c}{\|y\|^2} \right) y.$$

Show that

$$D(y) = \frac{1}{k} \frac{1}{\|y\|^2} \{ c^2 - 2c(k-2) \},$$

and that this is indeed negative for an interval of c values, provided the dimension is $k \geq 3$. Indeed demonstrate that the best value is $c_0 = k - 2$ and that the consequent risk function can be expressed as

$$R(\hat\theta, \theta) = 1 - \frac{(k-2)^2}{k} \mathbb{E}_\theta \frac{1}{\|Y\|^2} = 1 - \frac{k-2}{k} \mathbb{E} \frac{k-2}{\chi^2_k(\|\theta\|^2)}.$$

Here $\chi^2_k(\lambda)$ is the excentric chi-squared distribution with k degrees of freedom and excentre parameter λ.

(f) The arguments above led to the estimator

$$\hat{\theta}_i = \left(1 - \frac{k-2}{\|y\|^2}\right) y_i \quad \text{for } i = 1, \ldots, k,$$

which is a version of the Stein estimator. A useful modification is to truncate the shrinking factor $1-(k-2)/\|y\|^2$ to zero in the case of this being negative, i.e. $\|y\|^2 \leq k-2$. We write this as

$$\hat{\theta}_{\text{Stein}} = \left(1 - \frac{k-2}{\|y\|^2}\right)^+ y, \quad \text{where } x^+ = \max(0, x).$$

Prove that this modification actually improves the performance further. (It remains easier to work directly with $\hat{\theta}$, though, e.g. regarding risk functions.)

(g) Show that the greatest risk reduction for $\hat{\theta}$ takes place at zero, with $R(\hat{\theta}, 0) = 2/k$. For a few values of k, say $k = 5, 10, 100$, compute and display the risk functions for Y and $\hat{\theta}$, as functions of $\|\theta\|$. Do the same with the $\hat{\theta}_{\text{Stein}}$ estimator (for which you may use simulations to compute the risk).

(h) Now make the empirical Bayes connection, as follows. Start with the prior that takes $\theta_1, \ldots, \theta_k$ independent from the $\text{N}(0, \tau^2)$, and show that the Bayes estimator takes the form

$$\theta_i^* = \theta_i^*(\rho) = \rho y_i, \quad \text{with } \rho = \frac{\tau^2}{\tau^2 + 1}.$$

(Hint: use the results from Exercise 20.) Show that the marginal distribution of $y_1, \ldots, y_k$ is that of independent $\text{N}(0, 1+\tau^2)$ components, with maximum (marginal) likelihood estimate $\hat{\tau}^2 = (W-1)^+$, where $W = k^{-1} \sum_{i=1}^k y_i^2$. This invites

$$\hat{\rho} = \frac{(W-1)^+}{W} = \left(1 - \frac{k}{\|y\|^2}\right)^+,$$

or versions close to this, for the empirical Bayes estimator

$$\hat{\theta}_{i,\text{EB}} = \theta_i^*(\hat{\rho}) = \hat{\rho} y_i.$$

The Stein type estimator above can accordingly be viewed as an empirical Bayes construction. Note that $\hat{\theta}_{i,\text{EB}}$ can be motivated and constructed without any direct concern or calculations for the risk functions per se.

2.3 The Jeffreys prior

As Bayesian statisticians, if we genuinely know nothing about the prior parameter θ, we would like to be *non-informative*. That is, we would like to choose a prior distribution $\pi(\theta)$ which does not assume anything too strongly about θ. As pointed out in the beginning of Section 2, this often means that we want to be non-informative with respect to some key hypothesis. However, in many problems, there is no key hypothesis that we wish to test, and so we want a more general notion of a non-informative prior. One possibility is that we insist that the prior distribution

should be invariant under a change of coordinates for the parameters θ. That is, if statisticians A and B use two different parametrisation θ and ϕ of the same probability distribution, then for all (measurable) subsets $S \subseteq \Theta$,

$$\int_S \pi_\theta(\theta)\,\mathrm{d}\theta = \int_{\phi(S)} \pi_\phi(\phi)\,\mathrm{d}\phi. \tag{2.6}$$

In order to find such the prior satisfying this property, we first need to introduce the *Fisher information*. Let Y be a single observation from the density $f_\theta(y)$. Then the Fisher information $\mathcal{I}(\theta)$ is defined as

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(y)\right] = -\int_{\mathcal{Y}}\left\{\frac{\partial^2}{\partial\theta^2}\log f_\theta(y)\right\} f_\theta(y)\,\mathrm{d}y.$$

The *Jeffreys prior* is defined to be proportional to the square root of the Fisher information,

$$\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}.$$

One downside to the Jeffreys prior is that it does not always exist. For some model choices, the integral $\int \sqrt{\mathcal{I}(\theta)}\,\mathrm{d}\theta = \infty$ is not bounded, and so we cannot normalise the function $\sqrt{\mathcal{I}(\theta)}$ to get a valid density. In such cases, we say that $\pi(\theta)$ is an *improper prior*. We cannot plot its density, we cannot simulate data from it, etc. However, even though the prior does not exist, the posterior $\pi(\theta \mid y)$ often *does* exist. Therefore, posterior inference is still possible, even though we started with an improper prior.

Exercise 24. Let $Y \sim \text{Bernoulli}(\theta)$, so Y has the mass function

$$f_\theta(y) = \theta^y(1-\theta)^{1-y}, \quad \text{for } y = 0, 1.$$

Find the Fisher information $\mathcal{I}(\theta)$ and show that $\theta \sim \text{Beta}(1/2, 1/2)$ is the Jeffreys prior.

Exercise 25. Let $Y \sim \text{N}(\theta, \sigma^2)$, where $\sigma > 0$ is a known parameter. Find the Jeffreys prior for θ and show that it is improper.

Exercise 26. In this exercise, we prove the invariance property of the Jeffreys prior in the case where θ is one-dimensional.

Let $f_\theta(y)$ be a density with parameter $\theta \in \mathbb{R}$. Now, let $\phi = \phi(\theta)$ be a different parametrisation, and let $g_\phi(y)$ be the reparametrised density, so

$$g_{\phi(\theta)}(y) = f_\theta(y).$$

(a) Let $\ell_\phi(y) = \log g_\phi(y)$. Use the chain rule to show that

$$\frac{\partial^2\ell_\phi}{\partial\theta^2} = \frac{\partial^2\ell_\phi}{\partial\phi^2}\left(\frac{\partial\phi}{\partial\theta}\right)^2 + \frac{\partial\ell_\phi}{\partial\phi}\frac{\partial^2\phi}{\partial\theta^2}.$$

(b) The first derivative $\partial\ell_\phi/\partial\phi$ is called the *score*. Show that under mild conditions on g_ϕ, it has zero expectation,

$$\mathbb{E}_\phi\left[\frac{\partial\ell_\phi(y)}{\partial\phi}\right] = \int\left\{\frac{\partial\ell_\phi(y)}{\partial\phi}\right\} g_\phi(y)\,\mathrm{d}y = 0.$$

(c) Let $\mathcal{I}_\theta$ and $\mathcal{I}_\phi$ denote the Fisher informations with respect to the θ and ϕ parametrisations, respectively. Show that

$$\mathcal{I}_\theta(\theta) = \mathcal{I}_\phi(\phi) \left(\frac{\partial \phi}{\partial \theta} \right)^2.$$

(d) Finally, show that for any (measurable) subset $S \subseteq \Theta$,

$$\int_S \sqrt{\mathcal{I}_\theta(\theta)} \, \mathrm{d}\theta = \int_{\phi(S)} \sqrt{\mathcal{I}_\phi(\phi)} \, \mathrm{d}\phi.$$

3 The Laplace Approximation (Lazy Bayes)

In the previous section, we saw various ways to define prior distributions in Bayesian problems. We started by investigating conjugate priors, where the prior density has the same functional form as the likelihood. These are computationally useful since they allow us to easily compute the resulting posterior distributions. However, if you impose something different as your prior distribution – like the Jeffreys prior – you will find if rather difficult (or even impossible) to derive the posterior distribution in closed form. The difficulty arises from the *marginal likelihood*,

$$\pi(y) = \int \pi(y \mid \theta') \pi(\theta') \, \mathrm{d}\theta'.$$

This integral is usually not tractable, meaning that we cannot evaluate the normalisation constant for the posterior distribution.

So why did conjugate priors work so well? It all boils down to the functional form trick. When the prior and the posterior are of the same functional form, then so is the posterior, since

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

Hence, we only had to inspect the behaviour of the parameters of the posterior distribution, bypassing the marginal likelihood altogether. The key point is that if you choose anything *different* than a conjugate prior, this trick does not work in general.

Later in the course, we shall see the standard way of dealing with this problem, namely the use of Markov chain Monte Carlo (MCMC) methods. These methods allow us to approximately sample from the posterior distribution, even when the marginal likelihood cannot be evaluated. In this chapter, however, we shall consider a simpler approach which works surprisingly well in many cases, called the Laplace approximation (or sometimes, 'lazy Bayes'). Suppose we have a density $p(x)$ with an unknown normalisation constant Z, so $p(x) = f(x)/Z$, where we only know the function f. In the Bayesian setting we care about, p would be the posterior $\pi(\theta \mid y)$, Z would be the marginal likelihood $\pi(y)$ and f would be the product of the prior and the likelihood, $\pi(\theta)\pi(y \mid \theta)$. Let x_0 be the *mode* of the distribution p. That is,

$$x_0 = \arg\max_x p(x) = \arg\max_x f(x),$$

the point at which p (or equivalently, f) attains its maximum value. The idea of the Laplace approximation is to Taylor expand the funciton $\log f(x)$ about the mode x_0 and only include terms up to second order. Doing so, we have

$$\log f(x) \approx \log f(x_0) - \frac{1}{2}A(x - x_0)^2,$$

where

$$A = -\left.\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log f(x)\right|_{x=x_0}.$$

Note that the first order term vanished in our Taylor series since the first derivative vanishes at maxima. Now, the approximation above leads to

$$f(x) \approx f(x_0) \exp\left\{-\frac{1}{2}A(x - x_0)^2\right\}.$$

We recognise the functional form of this unnormalised density as that of a Gaussian distribution with mean x_0 and variance $1/A$. Therefore, the Laplace approximation of $p(x)$ is

$$p(x) \approx \frac{1}{\sqrt{2\pi/A}} \exp\left\{-\frac{1}{2}A(x - x_0)^2\right\}.$$

When applying the Laplace approximation in the Bayesian setting, we obtain the approximation

$$\pi(\theta \mid y) \approx \mathrm{N}(\theta_{\mathrm{MAP}}, 1/A),$$

where

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} \pi(\theta \mid y) = \arg\max_{\theta} \log \pi(\theta \mid y) = \arg\max_{\theta} \{\log \pi(\theta) + \log \pi(y \mid \theta)\} \tag{3.1}$$

is the *maximum a posteriori* estimate, and

$$A = -\left.\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log \pi(\theta \mid y)\right|_{\theta=\theta_{\mathrm{MAP}}} = -\left.\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \{\log \pi(\theta) + \log \pi(y \mid \theta)\}\right|_{\theta=\theta_{\mathrm{MAP}}}. \tag{3.2}$$

Exercise 27. Verify the relations (3.1) and (3.2). Why are these useful?

Exercise 28. After having gone through the above calculations, the Laplace approximation is best illustrated with an example. Suppose we want to approximate the complicated density $p(x) \propto 1/(1 + x^4) \times \sigma(2x)$, where $\sigma(a) = 1/(1 + \exp(-a))$ is the logistic sigmoid function. We can infact numerically integrate this function over the reals to find the normalisation constant $Z = 1.1107$, but let us see how the Laplace approximation compares. Find the Laplace approximation and plot its density alongside $p(x)$. Comment on the quality of the approximation.

Exercise 29. In many situations, θ is a p-dimensional vector, not just a scalar. Derive the Laplace approximation in the multidimensional setting, giving explicit expressions for the p-vector θ_{MAP} and the $p \times p$ matrix A.

4 Model selection and model averaging

So far, we have seen how Bayesian inference works under the specification of a single model. However, in many situations, we have several candidate models which we would like to compare or combine. For example, two scientists could impose two different prior distributions on θ, or disagree about the data generating model $\pi(\cdot \mid \theta)$. In other situations, we might have two good models which we would like to combine in order to describe the data better. This is similar to how one combines machine learning models via *ensembling* to get better accuracy.

We need to introduce a discrete *model space* $\mathcal{M}$, containing the candidate models $m \in \mathcal{M}$. The parameter space containing θ will now depend on m, so write $\theta \in \Theta_m$. Hence, the extended parameter space for all the models is

$$\bigcup_{m \in \mathcal{M}} (\Theta_m \times \{m\}) = \bigcup_{m \in \mathcal{M}} \bigcup_{\theta \in \Theta_m} \{(\theta, m)\}$$

Now, remember the first sentence we learned in the course, namely that as Bayesians, we treat every unknown quantity as a random variable. We will therefore impose a prior distribution $\pi(m)$ on the models. Usually, when $\mathcal{M}$ is finite, this is just the discrete uniform prior, but in some cases, we might want uneven priors.

If we condition on a specific model m, then we can carry out inference like before under this model, with the posterior

$$\pi(\theta \mid y, m) = \frac{\pi(y \mid \theta, m)\pi(\theta \mid m)}{\pi(y \mid m)}$$

under model m, and and the marginal likelihood

$$\pi(y \mid m) = \int_{\Theta_m} \pi(y \mid \theta, m)\pi(\theta \mid m) \, \mathrm{d}\theta$$

under model m.

In order to evaluate the different models in $\mathcal{M}$, we do the same analysis, but at the model level. The posterior probability for choosing model m is

$$\pi(m \mid y) = \frac{\pi(y \mid m)\pi(m)}{\pi(y)},$$

where

$$\pi(y) = \sum_{m \in \mathcal{M}} \pi(y \mid m)\pi(m)$$

is the marginal likelihood, averaged over the models.

Suppose we have two competing models m and m', with equal prior probabilities, which we would like to compare. We would do this by comparing their respective posterior probabilities $\pi(m \mid y)$ and $\pi(m' \mid y)$. The ratio of these two is

$$\frac{\pi(m \mid y)}{\pi(m' \mid y)} = \frac{\pi(y \mid m)\pi(m)}{\pi(y \mid m')\pi(m')} = \frac{\pi(y \mid m)}{\pi(y \mid m')},$$

since the prior probabilities of m and m' were assumed to be equal. Hence, we see that the ratio of the posterior model probabilities equals the ratio of the model specific marginal likelihoods, a ratio more commonly known as the *Bayes factor*. We will write

$$B_{m,m'} = \frac{\pi(y \mid m)}{\pi(y \mid m')}.$$

If the Bayes factor is greater than 1, it means we favour model m over m'. Conversely, if it is less than 1, then we favour m' over m.

Bayes factors are a useful tool, but unfortunately, they require the evaluation of marginal likelihoods, which we know is difficult. In the final section of the course, we shall look at some ways of estimating these marginal likelihoods for models where they cannot be evaluated analytically. In Section 5.1, we shall see model comparison in action in the context of regression.

Exercise 30. Read the first two sections of MacKay (1992), which you can easily find on Google Scholar or Oria. You do not have to understand all the details - but try to get a sense of the big picture.

(a) What is meant by the first and second levels of inference?

(b) Study Figure 1. Which boxes correspond to the first and second level of inference?

(c) What is the principle of *Occam's razor?* How does it relate to statistical modelling?

(d) Study Figure 2. How does this figure illustrate Occam's razor?

(e) Suppose that we want to perform model comparison, but instead of comparing marginal likelihoods, we use the maximum likelihood values $\pi(y \mid \hat{\theta}, m)$ for each model m. What goes wrong?

Exercise 31.

(a) (Based on lecture notes by Nicholls (2023)) Suppose we have a single parameter $\theta \in \mathbb{R}$ and a finite set of models $\mathcal{M}$. Show that under quadratic loss, the Bayes estimator is the *model averaged* posterior mean. In particular, show that this will have a lower Bayes risk than the posterior mean in any single model.

(b) So it seems like averaging multiple models always leads to an improvement. This can be observed in practice in machine learning, where the combination of multiple trained models (commonly known as *ensembling*) leads to more accurate predictions. Can you nevertheless think of any downsides to model averaging? Come up with scenarios where model selection is preferred, and others where model averaging is preferred.

4.1 The Bayesian information criterion (BIC)

We have seen that marginal likelihoods (and therefore also Bayes factors) are difficult to calculate due to the intractability of the integral. The Bayesian information criterion (BIC) offers an approximation of the marginal likelihood, based on the Laplace approximation, which can be a useful

heuristic when comparing models. For ease of notation, we omit the dependence on the model m in this section, so the marginal likelihood takes its usual form $\pi(y) = \int \pi(y \mid \theta)\pi(\theta)\,\mathrm{d}\theta$. The BIC is given by the expression

$$\mathrm{BIC} = -2\log\pi(y \mid \hat{\theta}) + p\log n, \tag{4.1}$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE) and $p = \dim(\theta)$ is the number of parameters in the model. The expression (4.1) is an approximation of $-2\log\pi(y)$, and so a smaller value of BIC is favoured. Note that we only need to compute the MLE to compute the BIC. However, it should be noted that it yields only an approximate answer, and so we should not base our decisions in model comparison on the BIC alone.

4.2 Derivation of the BIC

This section is non-examinable.

The remainder of this section is devoted to deriving the BIC approximation. We start with the marginal likelihood

$$\pi(y) = \int \pi(y \mid \theta)\pi(\theta)\,\mathrm{d}\theta = \int \exp\{\log\pi(y \mid \theta)\}\pi(\theta)\,\mathrm{d}\theta = \int \exp\{\ell(\theta)\}\pi(\theta)\,\mathrm{d}\theta, \tag{4.2}$$

where $\ell(\theta)$ is the log likelihood.

Now we employ the Laplace approximation. Expanding the log likelihood about the MLE $\hat{\theta}$, we get

$$\ell(\theta) = \ell(\hat{\theta}) - \frac{n}{2}(\theta - \hat{\theta})^\top A_{\hat{\theta}}(\theta - \hat{\theta}) + \cdots, \tag{4.3}$$

where

$$A_{\hat{\theta}} = -\left.\frac{1}{n}\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta^\top}\right|_{\theta=\hat{\theta}}.$$

This is a slightly different derivation as the one we used for the Laplace approximation in Section 3. The reason that we have included a factor of $1/n$ in the expression for $A_{\hat{\theta}}$ is that we want this quantity to be $O(1)$ with respect to n. By the variance of the asymptotic normality of the MLEs, we know that this holds for $A_{\hat{\theta}}$.

Now, assuming the prior is approximately linear around the MLE, we can expand $\pi(\theta)$ around $\hat{\theta}$ and ignore all but linear terms, which yields

$$\pi(\theta) = \pi(\hat{\theta}) + (\theta - \hat{\theta})^\top \left.\frac{\partial\pi(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}} + \cdots. \tag{4.4}$$

Substituting (4.3) and (4.4) into (4.2), we obtain

$$\begin{aligned}\pi(y) &= \int \exp\left\{\ell(\hat{\theta}) - \frac{n}{2}(\theta - \hat{\theta})^\top A_{\hat{\theta}}(\theta - \hat{\theta}) + \cdots\right\}\left\{\pi(\hat{\theta}) + (\theta - \hat{\theta})^\top \left.\frac{\partial\pi(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}\right\}\mathrm{d}\theta \\ &\approx \exp\{\ell(\hat{\theta})\}\pi(\hat{\theta}) \int \exp\left\{-\frac{n}{2}(\theta - \hat{\theta})^\top A_{\hat{\theta}}(\theta - \hat{\theta})\right\}\mathrm{d}\theta,\end{aligned} \tag{4.5}$$

where we have used that

$$\int (\theta - \hat{\theta}) \exp\left\{-\frac{n}{2}(\theta - \hat{\theta})^\top A_{\hat{\theta}}(\theta - \hat{\theta})\right\}\mathrm{d}\theta = 0.$$

Now, (4.5) is a Gaussian integral, which we can evaluate to

$$\int \exp\left\{-\frac{n}{2}(\theta-\hat{\theta})^\top A_{\hat{\theta}}(\theta-\hat{\theta})\right\}\,\mathrm{d}\theta = (2\pi)^{p/2}n^{-p/2}|A_{\hat{\theta}}|^{-1/2}.$$

Putting everything together, we get the approximation

$$\pi(y) \approx \exp\{\ell(\hat{\theta})\}\pi(\hat{\theta})(2\pi)^{p/2}n^{-p/2}|A_{\hat{\theta}}|^{-1/2}. \tag{4.6}$$

Taking the logarithm of (4.6) and multiplying by -2, we get

$$-2\log\pi(y) \approx -2\ell(\hat{\theta}) - 2\log\pi(\hat{\theta}) - p\log(2\pi) + p\log n + \log|A_{\hat{\theta}}|. \tag{4.7}$$

Finally, we ignore all terms less than or equal to $O(1)$ with respect to n to obtain the BIC (4.1).

5 Regression and classification

This section builds on Bishop (2006, Chapters 3 and 4).

In this section, we shall consider problems where our data[1] $y = (y_1, \ldots, y_n)^\top$ depend on a number of *inputs* $x = (x_1, \ldots, x_n)^\top$, where each input x_i is a d-dimensional vector. The main goal of such models is to predict the value of an outcome y' given a new input x'. This section is all about linear models, which provides a fundamental understanding of prediction models in the Bayesian setting. The derivations in this chapter will form the foundations needed for understanding more advanced models, such as neural networks or Gaussian processes. There are two main types of problems to study: regression and classification. In the regression setting, the dependent variables y are continuous, whereas in the classification setting, they are discrete. We shall attack the former setting fist.

5.1 Linear models for regression

At first glance, you may think that linear models for regression are too restrictive, implying a linear dependence between the inputs x_i and the outcomes y_i. However, we shall allow ourselves a lot more flexibility by first mapping the inputs x_i to a set of suitable *features*,

$$\phi(x_i) = (\phi_0(x_i), \phi_1(x_i), \ldots, \phi_{p-1}(x_i))^\top,$$

which are usually non-linear functions of the x_i. For reasons we shall see shortly, we let ϕ_0 be the constant function 1.

The models we shall study take the form

$$y_i = w^\top\phi(x_i) + \epsilon_i = w_0 + w_1\phi_1(x_i) + \cdots + w_{p-1}\phi_{p-1}(x_i) + \epsilon_i. \tag{5.1}$$

Here, the $w = (w_0, \ldots, w_{p-1})^\top$ is the vector of *coefficients* of the model, and the noise $\epsilon_i \sim N(0, \beta^{-1})$ are iid Gaussian variables with precision β. The *linearity* thus refers to the linearity in

[1] In previous sections, we have not cared much about whether to interpret y as a row or column vector. However, since the linear algebra is particularly important in this section, we will consistently treat y as a column vector.

the features $\phi(x_i)$, not the inputs x_i themselves. We now see why we declared $\phi_0 = 1$, namely so that our model includes a *constant term* w_0. Note that (5.1) can alternatively be written as

$$y_i \sim N(w^\top \phi(x_i), \beta^{-1}), \tag{5.2}$$

independently for $i = 1, \ldots, n$. Let us consider some examples of models of this kind.

Example 1.

(a) Let $\phi_j(x) = x^j$. Then the model becomes

$$y_i = w^\top \phi(x_i) + \epsilon_i = w_0 + w_1 x_i + w_2 x_i^2 + \cdots + w_{p-1} x_i^{p-1} + \epsilon_i.$$

That is, y_i is modelled as a degree $p-1$ polynomial with Gaussian noise.

(b) Let

$$\phi_j(x) = \exp\left\{-\frac{1}{2s^2}(x - \mu_j)^2\right\}.$$

These are usually called *Gaussian features*, due to their functional form. Here, s is a global hyperparameter, and the μ_j are the locations of the centres of the features. We do not have to normalise these features, since they are not probability distributions – they are simply functions mapping the inputs x_i to a richer feature space.

It is usually more convenient to write (5.1) in matrix form,

$$y = \Phi w + \epsilon. \tag{5.3}$$

Here, the $n \times p$ matrix

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{p-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_{p-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \cdots & \phi_{p-1}(x_n) \end{pmatrix} \tag{5.4}$$

is called the *design matrix*, and the vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a vector of iid Gaussians, all with precision β. We may also rewrite (5.2) in matrix form,

$$y \sim \mathrm{N}(\Phi w, \beta^{-1} I), \tag{5.5}$$

where I is the $n \times n$ identity matrix. Having set up the model, our goal is two-fold: First, we would like to estimate the parameters w and β of the model. Secondly, given a new input $x^\star$, we would like to predict the outcome $y^\star$.

5.1.1 The frequentist solution: Least squares and penalisation

Before we look at the Bayesian solution to the regression problem, let us first recap the frequentist solution, which you may have seen before. In the frequentist framework, the parameters w are

estimated via maximum likelihood estimation. We will treat the parameter β as fixed and known. From (5.2), the likelihood L_n takes the form

$$L_n(w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left\{-\frac{1}{2}\beta(y_i - w^\top \phi(x_i))^2\right\}, \tag{5.6}$$

and so the log likelihood is given by

$$\begin{aligned}
\ell_n(w) &= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\beta) - \frac{\beta}{2}\sum_{i=1}^{n}\{y_i - w^\top\phi(x_i)\}^2 \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\beta) - \frac{\beta}{2}\|y - \Phi w\|^2 \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\beta) - \frac{\beta}{2}(y - \Phi w)^\top(y - \Phi w),
\end{aligned} \tag{5.7}$$

where $\|\cdot\|$ is the ℓ^2 norm, also known as the Euclidean norm. To find the MLEs, we maximise (5.7) with respect to the parameters. Differentiating with respect to w gives

$$2\Phi^\top(y - \Phi w) = 0,$$

and so, solving for w, we obtain the MLE

$$\hat{w} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top y \tag{5.8}$$

That is, we recover the *normal equations*, i.e. the solution to the least squares regression problem. The matrix $\Phi^\dagger = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top$ is called the Moore-Penrose pseudo inverse. Note that if Φ is a nonsingular square matrix, then $\Phi^\dagger = \Phi^{-1}$. In this way, the pseudo inverse can be thought of as a generalisation of inverses to non-square matrices.

It should also be mentioned that we could treat β as its own parameter to be estimated. Doing so would yield the MLE

$$\frac{1}{\hat{\beta}} = \frac{1}{n}\|y - \Phi\hat{w}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{w}^\top\phi(x_i)\right\}^2. \tag{5.9}$$

Exercise 32. Download the `sinusoidal.csv` data set from course webpage, which you can find on the website. This was made by collecting $n = 20$ evenly spaced points $0 = x_1 < \cdots < x_n = 1$ on the unit interval, and then computing $y_i = f(x_i) + \varepsilon_i$, where

$$f(x) = \sin(2\pi x)$$

and $\varepsilon_i \sim \mathrm{N}(0, 0.2^2)$, independently, for $i = 1, \ldots, n$.

(a) Plot the data and the underlying function f, recreating Figure 1.

(b) Using the polynomial basis functions, use the normal equations to find $\hat{w}$ and plot the resulting curves along with the data, for $p = 4$ (i.e. cubic regression). Compute the mean squared error (MSE) and $\|w\|^2/p$.

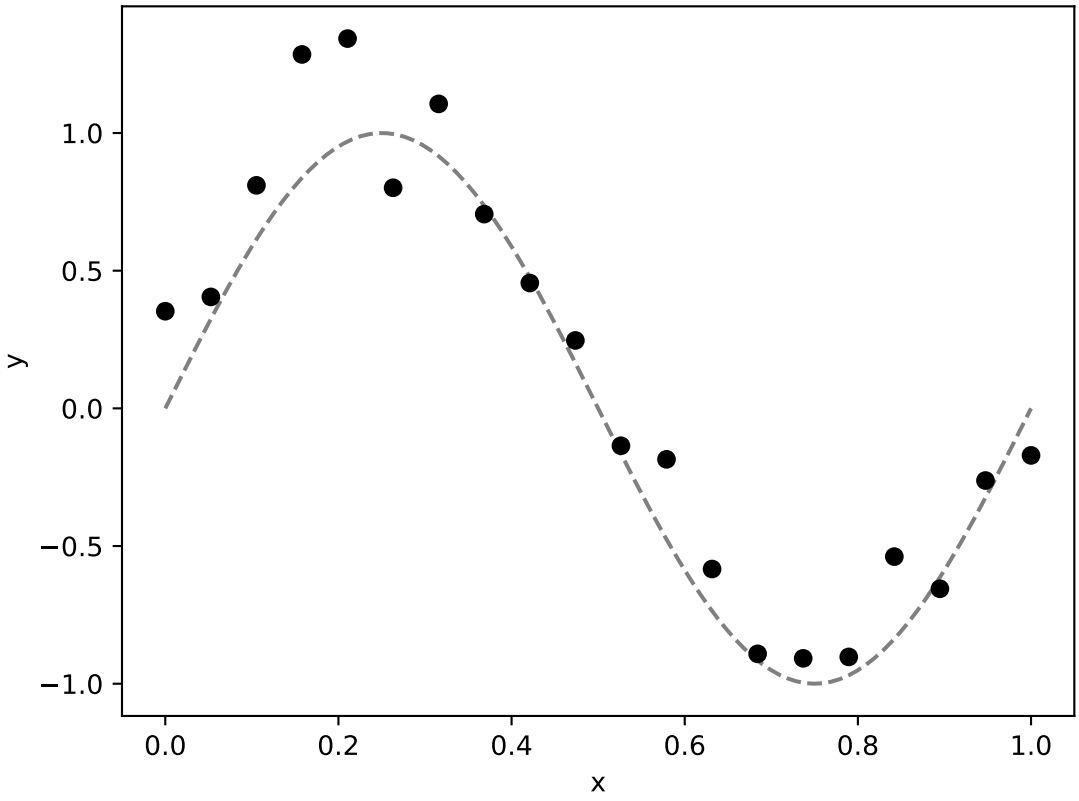

Figure 1: The `sinusoidal` data ($n = 20$), along with the underlying function.

(c) Repeat the above analysis for $p = 2, 3$ and $p = n = 20$. What happens when $p = n$? What are the values of the MSE and $\|w\|^2/p$, and how do they relate to the linear algebra involved?

In the previous exercise we saw that a more complicated model resulted in overfitting. In order to penalise model complexity, least squares regression often includes a *penalty term*, so we minimise instead

$$\min_w \left\{ \|y - \Phi w\|^2 + \lambda \|w\|^2 \right\}. \tag{5.10}$$

We see that large values of $\|w\|^2$ are discouraged. By maximising (5.10) with respect to w, we can show that

$$\hat{w} = (\lambda I + \Phi^\top \Phi)^{-1} \Phi^\top y, \tag{5.11}$$

where I is the $p \times p$ identity matrix. In the next section, we shall see that, unlike the frequentist approach, Bayesian regression penalises model complexity by construction. Indeed, we shall recover the penalised least squares minimiser (5.10) automatically.

It should be pointed out that other penalties can be used. Recall that for $1 \leq q < \infty$, the ℓ^q norm (denoted $\|\cdot\|_q$) of a vector $w = (w_0, \ldots, w_{p-1})^\top$ is given by

$$\|w\|_q = \left(\sum_{j=0}^{p-1} |w_j|^q \right)^{\frac{1}{q}}.$$

If we use the ℓ^q norm and raise to the power q rather than the ℓ^2 norm squared in (5.10), we obtain the optimisation problem

$$\min_w \left\{ \|y - \Phi w\|^2 + \lambda \|w\|_q^q \right\} = \min_{w,\beta} \left\{ \|y - \Phi w\|^2 + \lambda \sum_{j=0}^{p-1} |w_j|^q \right\}. \tag{5.12}$$

The choice of $q = 1$ leads to the famous *lasso*, which naturally to yields *sparse* minimisers (i.e. lots of zero entries in $\hat{w}$). This is important in the high-dimensional setting, where we want to exclude features with little predictive value.

5.1.2 Bayesian linear regression

Having seen the well-known frequentist solution to the regression problem (including regularisation), we are ready to look at the Bayesian approach to regression. As always, the key difference between these approaches is that in the Bayesian setting, we treat unknown quantities as random variables. Thus, we will impose a prior distribution on the vector of coefficients w. From (5.5), the likelihood is a Gaussian, and therefore the conjugate prior is also a Gaussian. For the sake of further simplicity, we shall impose a zero mean isotropic Gaussian,

$$w \sim \mathrm{N}(0, \alpha^{-1} I), \tag{5.13}$$

meaning that $\mathbb{E}[w] = 0$ and that the prior covariance is a constant multiple of the identity matrix. Rather than assigning priors to α and β, we shall treat them as hyperparameters of the model for

now. In Section 5.1.4, we will use empirical Bayes (see Section 2.2) to find optimal values for these parameters.

Now, having chosen a conjugate, Gaussian prior, we know that the posterior[2] for w will also be Gaussian,

$$w \mid y \sim \mathrm{N}(m_n, S_n).$$

It remains to find expressions for the posterior mean m_n and covariance matrix S_n. Fortunately, we do not have to do any extra work, as the derivations from Section 2.1.1 will do all the heavy lifting for us. The prior (5.13) and the likelihood (5.5) form pricesely a linear Gaussian model (see (2.2)), and we can therefore simply read off the posterior mean and covariance matrix using (2.3) and (2.4). Doing so yields

$$m_n = \beta S_n \Phi^\top y \tag{5.14}$$

$$S_n^{-1} = \alpha I + \beta \Phi^\top \Phi. \tag{5.15}$$

Exercise 33. Verify relations (5.14) and (5.15).

It is worth looking at what the MAP estimator looks like in this setting. Recall from Section 3 that the MAP is the value of w which maximises the (log) posterior, which in this case means that

$$\begin{aligned}
w_{\mathrm{MAP}} &= \arg\max_w \log \pi(w \mid y) \\
&= \arg\max_w \left\{\log \pi(y \mid w) + \log \pi(w)\right\} \\
&= \arg\max_w \left\{-\frac{\beta}{2}(y - \Phi w)^\top (y - \Phi w) - \frac{\alpha}{2} w^\top w\right\} \\
&= \arg\max_w \left\{-\frac{\beta}{2}\|y - \Phi w\|^2 - \frac{\alpha}{2}\|w\|^2\right\} \\
&= \arg\min_w \left\{\|y - \Phi w\|^2 + \frac{\alpha}{\beta}\|w\|^2\right\},
\end{aligned}$$

meaning that we have recovered regularised least squares regression (5.10) with $\lambda = \alpha/\beta$. This illustrates quite nicely how the Bayesian framework penalises model complexity by design. We do not have to arbitratily impose regularisers - the machinery does the job for us.

We just saw that the Gaussian prior naturally recovered regularised least squares regression with the Euclidean penalty term $\lambda\|w\|^2$. At this point, we may ask whether it is possible to recover the more generalised regularisation (5.12) by choosing a sufficiently clever prior for w. It turns out that this is indeed possible, and the required prior for w is given by

$$\begin{aligned}
\pi(w) &= \left\{\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q} \frac{1}{\Gamma(1/q)}\right\}^p \exp\left\{-\frac{\alpha}{2}\|w\|_q^q\right\} \\
&= \left\{\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q} \frac{1}{\Gamma(1/q)}\right\}^p \exp\left\{-\frac{\alpha}{2}\sum_{j=0}^{p-1} |w_j|^q\right\}.
\end{aligned}$$

[2]You may wonder why we do not condition on α, β or x in the notation for the posterior. Some authors do, but it is useful to only condition on *random* variables, which neither α, β or x are. This makes it easier for us to separate between what is random and what is not, and also makes the notation less cluttered.

It is not difficult to show that if we plug in $q = 2$, then we recover the isotropic Gaussian prior (5.13).

Usually, our main interest is not in analysing the posterior distribution $\pi(w \mid y)$, but rather the predictive distribution $\pi(y' \mid y)$ of a new output y' from a new input x', given the observed outputs y. This distribution can be written as

$$\begin{aligned}\pi(y' \mid y) &= \int \pi(y', w \mid y)\,\mathrm{d}w \\ &= \int \pi(y' \mid w)\pi(w \mid y)\,\mathrm{d}w,\end{aligned}$$

where both $\pi(y' \mid w)$ and $\pi(w \mid y)$ are Gaussian. By (2.5), we know that $\pi(y' \mid y)$ will also be Gaussian, and we can write its mean and variance in explicit form, namely

$$y' \mid y \sim \mathrm{N}(m_n^\top \phi(x'), \sigma_n^2(x')), \tag{5.16}$$

where

$$\sigma_n^2(x') = \frac{1}{\beta} + \phi(x')^\top S_n \phi(x').$$

Exercise 34. Return to the analysis of the `sinusoidal` data set, using polynomial basis functions for $p = 4$.

(a) Set $\alpha = 0.0001$ and $\beta = 10$, and go through the Bayesian analysis. Along with the data, plot the predictive mean, with $\pm$ one standard deviation on either side.

(b) Now modify the data set by removing all points x_i satisfying $x_i < 0.4$ (and remove the corresponding y_i). Create the same plot as you did in part (a), mutatis mutandis. What happens?

5.1.3 Model comparison

We will now discuss how to compare different models in the regression setting. This will allow us to evaluate the performance of different choices of basis functions, along with different choices of p. As we say in Section 4, this is done by evaluating the marginal likelihoods $\pi(y)$ for the competing models. In our setting, we have a Gaussian likelihood and a Gaussian prior, and so it would be possible to use (2.5) to marginalise directly. However, it turns out that in this case, we get the answer in a more insightful form if we complete the square in the integrand explicitly, and so that is what we shall do here. Now, the marginal likelihood $\pi(y)$ takes the form

$$\begin{aligned}\pi(y) &= \int \pi(y \mid w)\pi(w)\,\mathrm{d}w \\ &= \left(\frac{\beta}{2\pi}\right)^{n/2} \left(\frac{\alpha}{2\pi}\right)^{p/2} \int \exp\left\{-E(w)\right\}\,\mathrm{d}w,\end{aligned} \tag{5.17}$$

where

$$E(w) = \frac{\beta}{2}\|y - \Phi w\|^2 + \frac{\alpha}{2}\|w\|^2. \tag{5.18}$$

Completing the square over w like before yields

$$E(w) = E(m_n) + \frac{1}{2}(w - m_n)^\top A(w - m_n), \tag{5.19}$$

where

$$A = S_n^{-1} = \alpha I + \beta \Phi^\top \Phi.$$

Having completed the square, we can now compute the integral (5.17) to obtain

$$\begin{aligned}\int \exp\{-E(w)\}\,\mathrm{d}w &= \exp\{-E(m_n)\} \int \exp\{-\frac{1}{2}(w - m_n)^\top A(w - m_n)\}\,\mathrm{d}w \\ &= \exp\{-E(m_n)\}(2\pi)^{p/2}|A|^{-1/2}.\end{aligned}$$

Hence, the log marginal likelihood can be written as

$$\log \pi(y) = \frac{p}{2}\log\alpha + \frac{n}{2}\log\beta - E(m_n) - \frac{1}{2}\log|A| - \frac{n}{2}\log(2\pi). \tag{5.20}$$

Exercise 35. Continue the analysis of the `sinusoidal` data set, again with $\alpha = 0.0001$ and $\beta = 10$. Use (5.20) to compare the polynomial regression models for different values of p. Which model is the best? What do the results tell you about the compromise between the goodness of fit versus model complexity?

Exercise 36. In this exercise, we shall derive (5.20) directly using (2.5).

(a) Show that marginally,

$$y \sim \mathrm{N}(0, B),$$

where $B = \beta^{-1} I + \alpha^{-1}\Phi\Phi^\top$, and that therefore,

$$\log \pi(y) = -\frac{1}{2} y^\top B^{-1} y - \frac{1}{2}\log|B| - \frac{n}{2}\log(2\pi).$$

(b) Next, we need the binomial inverse theorem, which states that if X, Y, U, V are matrices of sizes $n \times n, p \times p, n \times p, p \times n$, respectively, then

$$(X + UYV)^{-1} = X^{-1} - X^{-1}UY(Y + YVX^{-1}UY)^{-1}YVX^{-1}.$$

Use this to show that

$$y^\top B^{-1} y = \beta y^\top y - \beta y^\top \Phi m_n.$$

(c) Now show that

$$\frac{\beta}{2} m_n^\top \Phi^\top \Phi m_n + \frac{\alpha}{2} m_n^\top m_n = \frac{\beta}{2} y^\top \Phi m_n,$$

and conclude thus that in fact,

$$\frac{1}{2} y^\top B^{-1} y = E(m_n).$$

(d) Next, we look at the determinant of B. Again, using the binomial inverse theorem, show that

$$B^{-1} = \beta\left\{I - \beta\Phi A^{-1}\Phi^\top\right\}.$$

(e) Recall that λ_j denote the eigenvalues of the matrix $\beta\Phi^\top\Phi$, with corresponding eigenvectors u_j. By showing that Φu_j is an eigenvector of $\beta\Phi A^{-1}\Phi^\top$, deduce that

$$|I - \beta\Phi A^{-1}\Phi^\top| = \prod_{j=0}^{p-1}\left(1 - \frac{\lambda_j}{\alpha+\lambda_j}\right).$$

Conclude that

$$|B| = \frac{1}{\alpha^p\beta^n}|A|.$$

(f) Finally, put everything together to recover (5.20).

5.1.4 Empirical Bayes

There is one final outstanding question in our regression story, namely how to decide the values of the hyperparameters α and β. The fully Bayesian solution is, as you might expect, to impose a prior distribution $\pi(\alpha, \beta)$, but this complicates the previous mathematical analysis, making the posterior $\pi(\alpha, \beta, w \mid y)$ and the predictive $\pi(y' \mid y)$ mathematically intractable. These issues can be overcome with more sophisticated machinery, like Markov chain Monte Carlo (MCMC) techniques, which we shall later on. However, in this section, we shall employ the technique of empirical Bayes, which we saw in Section 2.2, to choose α and β so that the (log) marginal likelihood is maximised. As we will discover, the analysis is quite fruitful and yields a nice geometrical interpretation of the effective number of parameters in the regression model.

For empirical Bayes, we need to optimise the log marginal likelihood (5.20) with respect to α and β. Let us consider the optimisation with respect to α first. To do so, we need to introduce the eigenvalues and eigenvectors of the matrix $\beta\Phi^\top\Phi$. Write

$$\left(\beta\Phi^\top\Phi\right)u_j = \lambda_j u_j, \tag{5.21}$$

for[3] $j = 0, \ldots, p-1$. Having introduced λ_j and u_j, we note that the eigenvectors of $A = \alpha I + \beta\Phi^\top\Phi$ are also the u_j, but with eigenvalues $\alpha + \lambda_j$. Now, the determinant of a matrix equals the product of its eigenvalues, and therefore

$$|A| = \prod_j(\alpha + \lambda_j).$$

Hence, when we differentiate (5.20) with respect to α, we have

$$\frac{\partial}{\partial\alpha}\log|A| = \frac{\partial}{\partial\alpha}\sum_j \log(\alpha+\lambda_j) = \sum_j \frac{1}{\alpha+\lambda_j}. \tag{5.22}$$

We also need to differentiate $E(m_n)$ with respect to α. It turns out that

$$\frac{\partial}{\partial\alpha}E(m_n) = \frac{1}{2}m_n^\top m_n. \tag{5.23}$$

[3] we index from $j = 0$ to be consistent with the indexing of the design matrix Φ in (5.4).

We now set the derivative of (5.20) with respect to α equal to zero, which yields

$$\frac{p}{2\alpha} - \frac{1}{2} m_n^\top m_n - \frac{1}{2} \sum_j \frac{1}{\alpha + \lambda_j} = 0. \tag{5.24}$$

Multiyplying through by 2α and rearranging, we get

$$\alpha = \frac{\gamma}{m_n^\top m_n}, \tag{5.25}$$

where

$$\gamma = \sum_j \frac{\lambda_j}{\alpha + \lambda_j}. \tag{5.26}$$

The quantity γ has a particularly nice geometric interpretation to which we shall return shortly. Note that (5.25) is only an implicit equation. Indeed, the parameter γ depends on α and the eigenvalues λ_j depend on the other hyperparameter β. However, as we shall see shortly, we can derive a similar implicit equation for β, which will allow us to make an initial guess for α and β, and then sequentially update their values using the implicit equation until we have optimised the marginal likelihood.

Exercise 37. Verify relation (5.23).

We now differentiate (5.20) with respect to β to obtain the implicit equation for β. First, we need to find the derivative of $\log|A|$ with respect to β. By differentiating (5.21), one can show that

$$\frac{\mathrm{d}\lambda_j}{\mathrm{d}\beta} = \frac{\lambda_j}{\beta}, \tag{5.27}$$

and so

$$\frac{\partial}{\partial \beta} \log|A| = \frac{\partial}{\partial \beta} \sum_j \log(\alpha + \lambda_j) = \sum_j \frac{\lambda_j/\beta}{\alpha + \lambda_j} = \frac{\gamma}{\beta}. \tag{5.28}$$

Using this to differentiate (5.20), we see that any stationary point must satisfy

$$\frac{n}{2\beta} - \frac{1}{2}\|y - \Phi m_n\|^2 - \frac{\gamma}{2\beta} = 0. \tag{5.29}$$

Multiyplying by 2β and rearranging, we obtain

$$\frac{1}{\beta} = \frac{1}{n-\gamma}\|y - \Phi m_n\|^2 = \frac{1}{n-\gamma} \sum_i \left\{ y_i - m_n^\top \phi(x_i) \right\}^2. \tag{5.30}$$

Again, this is only an implicit equation for β (as the eigenvalues λ_j depend on β), but we can use (5.25) and (5.30) in tandem to sequentially update the values of α and β until convergence to maximise the marginal likelihood.

Exercise 38. Verify relation (5.27).

Exercise 39. Return to the analysis of the `sinusoidal` data set (with the cubic regression model). Starting with $\alpha = 0$ and $\beta = 1$, use the updates (5.25) and (5.30) sequentially until convergence to optimise the log marginal likelihood with respect to α and β. What happens if you start with $\alpha = 1$ and $\beta = 1$?

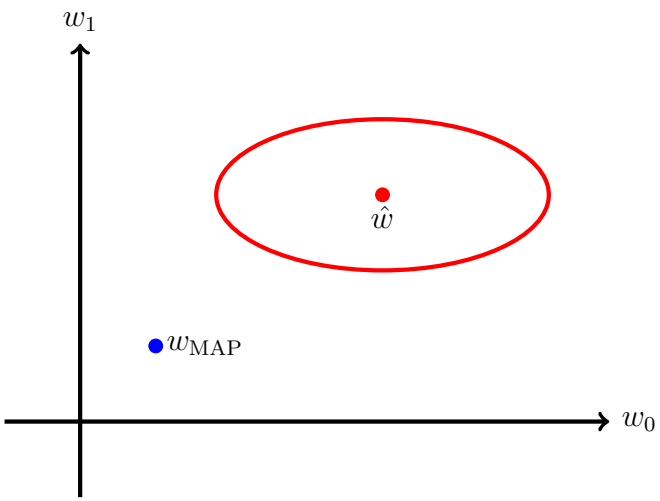

Figure 2: Contours of the likelihood function, drawn on axes aligned with the eigenvectors of $\beta\Phi^\top\Phi$.

The parameter γ has a particularly nice geometric interpretation which we will now explain. The key in this interpretation is to study the matrix $\beta\Phi^\top\Phi$. First of all, note that this matrix is positive definite, so that it has positive eigenvalues. Also, we can see that $\beta\Phi^\top\Phi$ is actually the Hessian of the log likelihood function with respect to w. Indeed,

$$\begin{aligned}\nabla\nabla^\top \log \pi(y \mid w) &= -\frac{\beta}{2}\nabla\nabla^\top (y-\Phi w)^\top (y-\Phi w) \\ &= -\beta\nabla(y-\Phi w)^\top\Phi \\ &= \beta\Phi^\top\Phi,\end{aligned}$$

and so the eigenvalues λ_j describe the *curvature* of the contours of the log likelihood function along the axes aligned with the eigenvectors u_j. See Figure 2 for an example in two dimensions. Here the axes for w_0 and w_1 have been aligned with the eigenvectors u_0 and u_1. In red, we see a contour of the log likelihood function, which is an ellipse centred at the MLE and aligned with the eigenvectors. Now, since the curvature is smaller in the direction of u_0 than u_1, we know that λ_0 is *smaller* than λ_1. In other words, the axis with the larger eigenvalue contributes more to the curvature of the log likelihood surface.

As all the eigenvalues λ_j are positive, the quantities $\lambda_j/(\alpha+\lambda_j)$ are bounded between 0 and 1, and so γ, defined in (5.26) is bounded between 0 and p. Now, if $\lambda_j \ll \alpha$, then $\lambda_j/(\alpha+\lambda_j) \approx 0$, and this term will not contribute much to the sum (5.26). Conversely, if $\lambda_j \gg \alpha$, then $\lambda_j/(\alpha+\lambda_j) \approx 1$, and so we get a substantial contribution. Viewed in this light, γ measures the effective number of parameters in the model.

The discussion above now yields some useful insight into the estimation procedure for β, defined by (5.30). Comparing with (5.9), we see a great similarity. Both express the variance $1/\beta$ as an empirical average of the squared distance between the data and the predictions. However, in (5.9), we divide by n in the average, whereas in (5.30), we divide by $n-\gamma$. This is quite similar to the problem of estimating the mean and variance of a distribution from data. Indeed, suppose that

$z_1, \ldots, z_n$ are iid samples from a distribution with mean μ and variance σ^2. We would estimate the mean as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Having done so, we could estimate the variance by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x - \hat{\mu})^2,$$

however, this estimator is biased. In order to obtain an unbiased estimator, we use the estimator $\hat{\sigma}^2 n/(n-1)$. Similarly, in frequentist regression problems, the corresponding unbiased estimator for the variance divides by n minus the number of parameters, rather than n. This is precisely what happens in (5.30).

Exercise 40. Return once more to the `sinusoidal` data set. Using the optimal values you found in Exercise 39, along with the eigenvalues λ_j of $\beta \Phi^\top \Phi$, calculate the effective number of parameters.

5.2 Linear models for classification

We now move on to the second kind of input-dependent models which we shall study, namely classification models. In the regression setting in the previous section, the outputs y_i were continuous. In classification, they are discrete. More specifically, we shall only consider the case of *binary classification*, so each $y_i \in \{0, 1\}$. It is not too difficult to extend the results we will derive here to the multiclass setting.

Like in the previous section, we shall consider linear models, where the inputs x_i have been mapped to some fixed features $\phi(x_i) = (\phi_0(x_i), \ldots, \phi_{p-1}(x_i))^\top$, where, as before, $\phi_0(x) = 1$. We still let $w = (w_0, \ldots, w_{p-1})^\top$ denote the vector of coefficients. However, there is no Gaussian noise this time, since we assume all the data are correctly labelled. Instead, we want to map the inner product $\eta_i = w^\top \phi(x_i)$ to the interval $[0, 1]$, so that this will represent the *probability* that $y_i = 1$, given input x_i. We do this by choosing a suitable smooth and increasing function $f : \mathbb{R} \to [0, 1]$, and model the probability that $y_i = 1$ by

$$\mathbb{P}(y_i = 1 \mid w) = f(\eta_i) = f(w^\top \phi(x_i)), \tag{5.31}$$

and $\mathbb{P}(y_i = 0 \mid w) = 1 - \mathbb{P}(y_i = 1 \mid w)$.

A common choice of function f is the logistic sigmoid function σ, defined by

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{5.32}$$

This function has the nice property that

$$\frac{\mathrm{d}}{\mathrm{d}a} \sigma(a) = \sigma(a)[1 - \sigma(a)]. \tag{5.33}$$

Classification with σ is commonly referred to as *logistic regression*, but we emphasise that it is not regression in the sense of the previous section, since the outcomes y_i are discrete, not continuous. It is merely a convention of terminology.

Another common choice is the probit function Φ, defined by

$$\Phi(a) = \int_{-\infty}^{a} (2\pi)^{-1/2} \exp\{-t^2/2\} \, \mathrm{d}t, \tag{5.34}$$

i.e., it is the standard normal cdf. Classification with Φ is commonly referred to as *probit regression*.

From (5.31), we see that the outcomes y_i as Bernoulli random variables, with parameters $f(\eta_i)$. The likelihood of the outcomes $y = (y_1, \ldots, y_n)^\top$ given w is thus

$$\pi(y \mid w) = \prod_{i=1}^{n} f(\eta_i)^{y_i} [1 - f(\eta_i)]^{1-y_i}, \tag{5.35}$$

and the log likelihood is

$$\log \pi(y \mid w) = \sum_{i=1}^{n} \{y_i \log f(\eta_i) + (1 - y_i) \log[1 - f(\eta_i)]\}. \tag{5.36}$$

5.2.1 Bayesian classification

To achieve a Bayesian solution to classification, we impose a prior distribution on w. Like in the regression setting, we shall impose a Gaussian prior, as this simplifies the upcoming mathematical analysis. Letting m_0 and S_0 denote the prior mean and covariance matrix, we let

$$w \sim \mathrm{N}(m_0, S_0). \tag{5.37}$$

Unlike the the previous section, we do not insist on using a zero mean isotropic prior, as there is not much to gain from this assumption in the classification problem.

Combining (5.36) and in Bayes theorem, we see that the log posterior takes the form

$$\log \pi(w \mid y) = -\frac{1}{2}(w - m_0)^\top S_0^{-1}(w - m_0) + \sum_{i=1}^{n} \{y_i \log f(\eta_i) + (1 - y_i) \log[1 - f(\eta_i)]\} + \text{constant}. \tag{5.38}$$

Due to the presence of the Bernoulli likelihood, we are not able to evaluate the posterior distribution in closed form. We will therefore employ the Laplace approximation, introduced in Section 3. That is, we use a Gaussian approximation $\pi(w \mid y) \approx q(w)$, with mean w_{MAP} and precision matrix

$$S_n^{-1} = -\nabla\nabla^\top \log \pi(w \mid y). \tag{5.39}$$

Exercise 41.

(a) Using (5.33), show that for logistic regression, the approximate posterior covariance matrix S_n takes the form

$$S_n^{-1} = S_0^{-1} + \sum_{i=1}^{n} \sigma(\eta_i)[1 - \sigma(\eta_i)]\phi(x_i)\phi(x_i)^\top.$$

(b) Show also that for probit regression, we have

$$S_n^{-1} = S_0^{-1} + \sum_{i=1}^{n} \left\{ y_i \frac{\eta_i \mathrm{N}(\eta_i; 0, 1)\Phi(\eta_i) + \mathrm{N}^2(\eta_i; 0, 1)}{\Phi^2(\eta_i)} + (1 - y_i) \frac{\mathrm{N}^2(\eta_i; 0, 1) - \eta_i \mathrm{N}(\eta_i; 0, 1)\left[1 - \Phi(\eta_i)\right]}{\left[1 - \Phi(\eta_i)\right]^2} \right\} \phi(x_i)\phi(x_i)^\top.$$

We have used the notation $\mathrm{N}(x; 0, 1)$ rather than the standard $\phi(x)$ here to avoid confusion with the features $\phi(x_i)$.

(c) The expression for the probit model is not so nice, so we tend to work with its *expectation* instead (with respect to y). Show that

$$\mathbb{E}_y[S_n^{-1}] = S_0^{-1} + \sum_{i=1}^{n} \frac{\mathrm{N}^2(\eta_i; 0, 1)}{\Phi(\eta_i)\left[1 - \Phi(\eta_i)\right]} \phi(x_i)\phi(x_i)^\top.$$

What have we in fact calculated here?

(d) Assume that S_0 is a positive definite matrix. Verify that the matrices in (a) and (c) are also positive definite for any w. Why is this useful when computing w_{MAP}?

(e) Write down the updating equation for the Newton-Raphson formula for computing w_{MAP} for the probit model, where you use $-\mathbb{E}_y[S_n^{-1}]$ as a proxy for the Hessian matrix of the log posterior.

Having derived our Gaussian approximation $q(w)$ of the posterior, we move on the the predictive distribution $\pi(y' \mid y)$ of a new outcome y' from a new input x', given the data y. With the Laplace approximation, we have

$$\begin{aligned} \mathbb{P}(y' = 1 \mid y) &= \int \mathbb{P}(y' = 1 \mid w)\pi(w \mid y)\,\mathrm{d}w \\ &\approx \int \mathbb{P}(y' = 1 \mid w)q(w)\,\mathrm{d}w \\ &= \int f(w^\top \phi(x'))q(w)\,\mathrm{d}w. \end{aligned} \tag{5.40}$$

A priori, this integral is quite difficult to evaluate, as we have to integrate over all the coefficients w. However, we can exploit that $f(w^\top \phi(x'))$ only depends on w through the inner product $\eta' = w^\top \phi(x')$, and so simplify the above integral significantly by working with the distribution of η' instead. Now, η' is a linear combination of Gaussian random variables, so we know it must be Gaussian. From the Laplace approximation we derive that $\pi(\eta' \mid y) = \mathrm{N}(\eta'; \mu, \sigma^2)$, where

$$\begin{aligned} \mu &= \mathbb{E}[\eta' \mid y] = w_{\mathrm{MAP}}^\top \phi(x'), \\ \sigma^2 &= \mathrm{Var}(\eta' \mid y) = \sigma^2 = \phi(x')^\top S_n \phi(x'). \end{aligned}$$

Hence, we can write (5.40) as

$$\begin{aligned}\mathbb{P}(y' = 1 \mid y) &= \int f(\eta')\pi(\eta' \mid y)\,\mathrm{d}\eta' \\ &= \int f(\eta')\mathrm{N}(\eta'; \mu, \sigma^2)\,\mathrm{d}\eta'.\end{aligned}$$

In general, we are not able to simplify this integral further. However, in the case of probit regression, we can actually evaluate this integral analytically. Doing so, we obtain

$$\mathbb{P}(y' = 1 \mid y) = \int \Phi(\eta')\mathrm{N}(\eta'; \mu, \sigma^2)\,\mathrm{d}\eta' = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \tag{5.41}$$

as our final expression of the predictive probability that $y' = 1$.

Exercise 42. Verify relation (5.41) by differentiating both sides with respect to μ. It is useful to use the substitution $\eta' = \mu + \sigma z$ in the integral.

Exercise 43. We shall now apply the theory from this Section on a real data set, namely the famous `Iris` data set, which comprises $N = 150$ observations, each of which has a four dimensional input x and a target $y \in \{0, 1, 2\}$. Since we want to only work with binary classification, we only use the observations with $y = 0$ or $y = 1$. In total, there are $n = 100$ such observations.

(a) Import the Iris data set. In `Python`, you can find it in `sklearn.data sets`. Make sure to delete the rows of X and y where $y = 2$. Also, add a column of ones to X so that we include the constant feature. Thus, $p = 5$.

We want to project the inputs X onto two dimensions so that we can plot the data and see what is going on. Find the first two principal components. If you are not familiar with the principal component analysis (PCA), use the following recipe: First find the *centred empirical covariance matrix* S, given by

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top.$$

Note that S is symmetric and positive semidefinite. Therefore, by the spectral theorem, its eigenvalues are real and we can write

$$S = U^\top \Lambda U,$$

where Λ is a diagonal matrix of real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$, and U is an orthogonal matrix whose columns are the corresponding eigenvectors. The dimension reduction map we want to use is given by the $2 \times p$ matrix M, whose rows are u_1 and u_2, corresponding to the two largest eigenvalues λ_1 and λ_2. To make sure our eigenvalues all have the same sign, let us declare that the last entry of u_1 is positive and that the last entry of u_2 is negative. Using the map M, recreate the scatter plot in Figure 3.

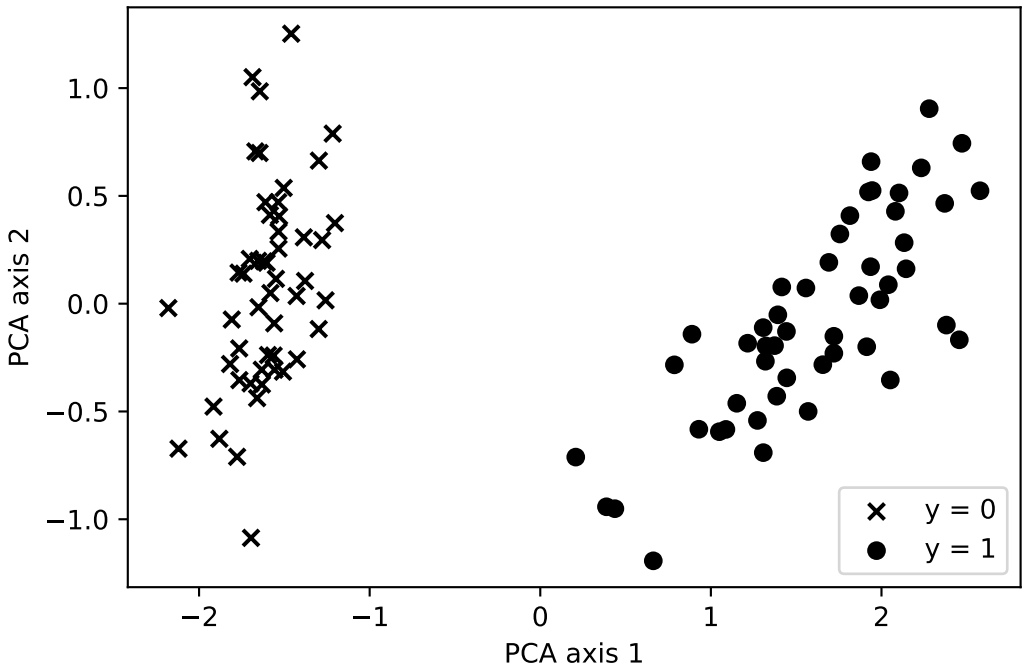

Figure 3: The two first principal components of the Iris data sets, restricted to the $n = 100$ observations with $y \in \{0, 1\}$.

(b) Let us use probit regression to train a binary classifier for our "binary Iris" data set. Use $m_0 = 0$ and $S_0 = \alpha^{-1} I$ in the prior, with $\alpha = 0.01$. Implement the Newton-Raphson algorithm you wrote down in Exercise 41 (e) to calculate w_{MAP}. You will have to work on a log scale here, so that things do not blow up numerically.

(c) The equation for the decision boundary is given by

$$\mathbb{P}(y' = 1 \mid y) = \frac{1}{2}.$$

From (5.41), show that this equation is indeed what we expect it to be, namely

$$w_{\text{MAP}}^{\top} \phi(x') = 0.$$

Use this, along with map M you found, to include the decision boundary in your scatter plot.

(d) Compute the posterior predictive mean μ and variance σ^2 at all points in your plot, and use these to plot the predictive density in your plot. In `Python`, you may find the `pcolormesh` function helpful.

6 Exchangeability and De Finetti's Theorem

It is time for a theoretical interlude. In the present section and the next, we shall make a detour into two aspects of theoretical foundations of Bayesian inference (without proofs), starting with de Finetti's theorem. The upshot of this theorem is the following assertion. Suppose we have *exchangeable* data $y = (y_1, \ldots, y_n)$, which means, roughly speaking, that shuffling the order of the observations makes no difference. Then there must exist some parameter θ and an accompanying prior $\pi(\theta)$ such that when we condition on θ, the data are iid.

Hence, de Finetti's theorem provides justification for the Bayesian perspective. It provides sufficient conditions for the existence of a prior and a data model. To understand the theorem fully, we first need to properly define the notion of exchangeability.

6.1 Exchangeability

We begin with the definition of finite exchangeable sequences. A *permutation* is a bijective map $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Definition 6.1. We say that a finite sequence $Y = (Y_1, \ldots, Y_n)$ of random variables is *exchangeable* if, for any permutation σ, we have

$$(Y_1, \ldots, Y_n) \sim (Y_{\sigma(1)}, \ldots, Y_{\sigma(n)}),$$

where $\sim$ means 'distributed equally' in this setting.

That is, the joint distribution of the Y_i does not change after we shuffle the coordinates. If the joint density of the Y_i is $f(y_1, \ldots, y_n)$, then $Y_1, \ldots, Y_n$ is exchangeable if and only if

$$f(y_1, \ldots, y_n) = f(y_{\sigma(1)}, \ldots, y_{\sigma(n)})$$

for all permutations σ.

We now extend the above definition of infinite sequences.

Definition 6.2. Let $Y_1, Y_2, \ldots$ be a sequence of random variables. We say that the sequence is *exchangeable* if for any number n, the finite sequence $Y_1, \ldots, Y_n$ is exchangeable.

Exercise 44. Show that any iid sequence is also exchangeable.

Exercise 45. We shall now look at Pólya's urn, which is a famous example of a sequence which is exchangeable but not iid. Suppose we have α red and β blue marbles in an urn. We repeatedly select a marble from the urn at random, look at it, and then return the marble, along with another marble of the same colour. Let $Y_i = 1$ if the ith marble drawn is red, and $Y_i = 0$ if it is blue.

(a) Show that Y_1 and Y_2 are not independent. Conclude that the sequence of Y_i cannot be iid.

(b) Write down the probability that we first draw k red marbles, followed by $n - k$ blue marbles.

(c) Show that the probability you calculated in (b) remains the same if you permute the draws. That is, the first n draws forms an exchangeable sequence. Conclude that the entire sequence of draws must be exchangeable.

From the previous exercise, we see that exhangeable sequences are not always iid. De Finetti's theorem builds a bridge between exchangeable and iid sequences, via the existence of a hyperparameter and a prior distribution. We state the theorem here only for binary random variables (like we saw in the Pólya urn example), but note that the theorem holds in much more general settings as well.

Theorem 6.1. *Let $Y_1, Y_2, \ldots$ be an infinite exchangeable sequence of binary random variables. Then there exists a random variable $\theta \in [0,1]$ such that conditined on θ, the sequence $Y_1, Y_2, \ldots$ are iid Bernoulli random variables with parameter θ. Furthermore, $\theta = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n Y_i$.*

Let us expand a bit more on de Finetti's theorem to make sure we understand it fully. Let $\pi(y)$ denote the pmf of the first n variables $y = (y_1, \ldots, y_n)$ in the sequence. Then the theorem states that there exists a cdf $F(\theta)$ on $[0,1]$ such that

$$\pi(y) = \int_0^1 \left\{ \prod_{i=1}^n \pi(y_i \mid \theta) \right\} \mathrm{d}F(\theta), \tag{6.1}$$

where each $\pi(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$. You might not be too familiar with the idea of integrating with respect to a distribution, but if θ has a density $\pi(\theta)$, then $\mathrm{d}F(\theta) = \pi(\theta)\,\mathrm{d}\theta$, so that (6.1) simplifies to

$$\pi(y) = \int_0^1 \pi(\theta)\pi(y \mid \theta)\,\mathrm{d}\theta,$$

where $\pi(y \mid \theta) = \prod_{i=1}^n \pi(y_i \mid \theta)$. Hence, de Finetti's theorem tells us that for any exchangeable sequence, there is an underlying Bayesian model such that $\pi(y)$ is the marginal likelihood of that model. This basically tells us that the Bayesian perspective naturally arises whenever we are dealing with exchangeable data. It is important to note, however, that the theorem does not tell us what the prior $\pi(\theta)$ is. It merely asserts its existence.

Exercise 46. According to de Finetti's theorem, there should be an underlying parameter θ in the Pólya urn model such that (6.1) holds. Show that in fact $\theta \sim \mathrm{Beta}(\alpha, \beta)$ is the correct prior. That is, show that the probability that you found in Exercise 45 can be written in the form (6.1), with $\mathrm{d}F(\theta) = \mathrm{Beta}(\theta; \alpha, \beta)\,\mathrm{d}\theta$.

An important point to make is that de Finetti's theorem only holds for *infinite* exchangeable sequences, as the next exercise[4] illustrates.

Exercise 47. Let Y_1 and Y_2 be random variables satisfying

$$\mathbb{P}(Y_1 = 0, Y_2 = 1) = \mathbb{P}(Y_1 = 1, Y_2 = 0) = \frac{1}{2},$$
$$\mathbb{P}(Y_1 = Y_2 = 0) = \mathbb{P}(Y_1 = Y_2 = 1) = 0.$$

Verify that (Y_1, Y_2) is an exchangeable sequence but that de Finetti's theorem does not apply.

[4] originally taken from Diaconis and Freedman (1980).

7 The Bernstein-von Mises theorem

This section builds on van der Vaart (1998, Chapter 10).

We have now come to the second part of our theoretical interlude, in which we will look at (but not prove) the celebrated Bernstein–von Mises theorem, which bridges the gap between the frequentist and Bayesian perspectives as the sample size n grows to infinity. Before introducing the theorem itself, let us look at the frequentist story, which you may have seen before. A key property of MLEs is their *asymptotic normality*, which, in the simplest case, can be summarised as follows. Let $y = (y_1, \ldots, y_n)$ be a vector of iid data from a smooth density $\pi(y \mid \theta)$ depending on a parameter θ (no priors here, we are in the frequentist world). Letting $\hat{\theta}_n$ denote the MLE, we have that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} Z \sim \mathrm{N}(0, \mathcal{I}(\theta)^{-1}), \tag{7.1}$$

where $\xrightarrow{d}$ denotes convergence in distribution.

The above basically tells us that we can approximate

$$\hat{\theta}_n \approx \mathrm{N}\left(\theta, \frac{1}{n}\mathcal{I}(\theta)^{-1}\right),$$

which provide a universal recipe for constructing confidence intervals for the parameter θ of interest.

As Bayesians, we impose a prior distribution π for θ and ask whether the distribution of the variable $\sqrt{n}(\theta - \hat{\theta}_n) \mid y$ is approximately normal with the same mean and variance. The key to be able to answer this question will be to study the *distance* between the distribution $\mathrm{N}(\hat{\theta}_n, \mathcal{I}^{-1}(\theta_0)/n)$ and the posterior $\pi(\theta \mid y)$. Hence, we need some notion of distance between two probability distributions, which we will now define.

Definition 7.1. Let P and Q be two probability measures on a measurable space $(\Omega, \mathscr{F})$. The *total variation distance* $\delta(P, Q)$ is given by

$$\delta(P, Q) = \sup_{A \in \mathscr{F}} |P(A) - Q(A)|.$$

That is, the maximal distance between the probabilities that P and Q can assign to the same event. Note that δ is symmetric, so $\delta(P, Q) = \delta(Q, P)$.

Exercise 48. It turns out that if P and Q emit densities p and q, then the total variation distance between P and Q can be written as

$$\delta(P, Q) = \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)| \, \mathrm{d}x. \tag{7.2}$$

The point of this exercise is to verify relation (7.2).

(a) Let

$$B = \{p \geq q\} = \{x \in \mathbb{R} : p(x) \geq q(x)\}.$$

By splitting the integral in (7.2) into two integrals over B and $\mathbb{R} \setminus B$, verify the "$\geq$" part.

(b) We will now prove the other inequality, which is a bit more difficult. We break it down into several easier steps. Show first that

$$\int_B \{p(x) - q(x)\}\,\mathrm{d}x = \int_{\mathbb{R}\setminus B} \{q(x) - p(x)\}\,\mathrm{d}x.$$

(c) Next, show that for any (measurable) set A, we have

$$\int_A \{p(x) - q(x)\}\,\mathrm{d}x \leq \int_B \{p(x) - q(x)\}\,\mathrm{d}x,$$

and use this, along with the result in (b), to show that

$$\int_A \{p(x) - q(x)\}\,\mathrm{d}x \leq \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)|\,\mathrm{d}x.$$

(d) Argue from symmetry that

$$\int_A \{q(x) - p(x)\}\,\mathrm{d}x \leq \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)|\,\mathrm{d}x.$$

(e) Put (c) and (d) together to conclude that indeed

$$\delta(P, Q) = \sup_{A\in\mathscr{F}} \left| \int_A \{p(x) - q(x)\}\,\mathrm{d}x \right| \leq \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)|\,\mathrm{d}x.$$

Now, having introduced the total variation distance between two probability measures, we are ready to state the Bernstein–von Mises theorem. Let P_{θ_0} denote the distribution of y given that the true parameter θ_0 is used. The Bernstein-von Mises theorem studies the distance

$$\delta\left(\pi(\cdot \mid y), \mathrm{N}(\hat{\theta}_n, \frac{1}{n}\mathcal{I}_{\theta_0}^{-1})\right),$$

which we note is itself a random variable, since it is a function of the data y.

Theorem 7.1 (Bernstein-von Mises)**.** *Under the above setup and mild regularity conditions, we have that*

$$\delta\left(\pi(\cdot \mid y), \mathrm{N}(\hat{\theta}_n, \frac{1}{n}\mathcal{I}_{\theta_0}^{-1})\right) \xrightarrow{P_{\theta_0}} 0. \tag{7.3}$$

That is, as the number n of data points grows, the distance between the posterior distribution and the distribution $N(\hat{\theta}_n, \mathcal{I}_{\theta_0}^{-1}/n)$ converges to zero in probability under the data model with $\theta = \theta_0$.

Exercise 49.

(a) In what way does the Bernstein von-Mises theorem justify the Laplace approximation?

(b) It looks like the frequentist and Bayesian statisticians will agree on most questions when the data are sufficiently numerous. Can you nevertheless think of questions where they will always differ in their conclusions, regardless of the magnitude of the sample size? *Hint: See Section 4.*

8 Markov chain Monte Carlo (MCMC)

We will now study the topic of Markov chain Monte Carlo (MCMC), which has been a crucial development for Bayesian methods over the last 30+ years. Let us recall the definition of the posterior density,

$$\pi(\theta \mid y) = \frac{\pi(\theta)\pi(y \mid \theta)}{\pi(y)} = \frac{\pi(\theta)\pi(y \mid \theta)}{\int \pi(\theta')\pi(y \mid \theta') \, \mathrm{d}\theta'}.$$

In most applications, both the prior $\pi(\theta)$ and the likelihood $\pi(y \mid \theta)$ are well-behaved and can be evaluated directly. However, unless we use a conjugate prior, the marginal likelihood $\pi(y)$ will be difficult to evaluate, or even to approximate accurately. This is the fundamental challenge of Bayesian statistics, and makes posterior inference difficult.

The main idea behind MCMC methods is to instead try to generate a sample $\theta_1, \ldots, \theta_S \sim \pi(\cdot \mid y)$ *approximately* from the posterior, bypassing evaluations of its density altogether. This is done by constructing a clever Markov chain, whose long term behaviour mimics that of the posterior distribution. Since the Markov chain is not an exact iid posterior sample, MCMC is an example of an approximate inference technique. However, it works very well in practice, and is relatively straight-forward to implement once we have understood the fundamental ideas. We start with the theory of Markov chains in a general state space.

8.1 General state space Markov chains

Definition 8.1. Let $X_1, X_2, \ldots$ be a sequence of random variables taking values in a set Ω, called the state space. We say that the sequence is a *Markov chain* if, for all n and all (measurable) sets A,

$$\mathbb{P}(X_{n+1} \in A \mid X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} \in A \mid X_n = x_n). \tag{8.1}$$

That is, when generating X_{n+1}, knowing the previous step in the chain is equivalent to knowing the entire previous history. We have in some sense "forgotten" all the history of the chain apart from the previous step.

Definition 8.2. We say that a Markov chain $\{X_n\}$ is *homogeneous* if the transition probabilities $\mathbb{P}(X_{n+1} \in A \mid X_n = x_n)$ are independent of n. In this case, we can specify the chain by the initial distribution P_1 of X_1, and the *kernel* k of the chain, defined by

$$k(x, A) = \mathbb{P}(X_{n+1} \in A \mid X_n = x).$$

Note that k does not depend on n by assumption. In many cases, the kernel emits a density, and we write

$$k(x, A) = \int_A k(x, y) \, \mathrm{d}y,$$

so it will be clear from context whether $k(\cdot, \cdot)$ refers to the transition distribution or to its density.

Henceforth, we will only study homogeneous Markov chains.

Exercise 50. Explain why $\int_\Omega k(x, y) \, \mathrm{d}y = 1$.

Example 2.

(a) Let $P_1 = \mathrm{N}(0, 1)$ and let $k(x, y) = \mathrm{N}(y; x, \sigma^2)$ be the transition density, for some fixed $\sigma > 0$. Then $\{X_n\}$ is a homogeneous Markov chain with a Gaussian kernel.

(b) Let $P_1 = \mathrm{N}(0, 1)$ and let $0 < \alpha < 1$. Consider the kernel

$$k(x, A) = \alpha P_x(A) + (1 - \alpha)\delta_x(A),$$

where P_x is a fixed probability distribution (allowed to depend on x) and δ_x is the degenerate distribution at x. We can think of this kernel as follows. At each step in the chain, we generate a *proposal* $x' \sim P_x$, which is accepted with probability α. Otherwise, it is rejected, and the chain stays at the previous value x.

Note that strictly speaking, this kernel does not emit a density due to the presence of the degenerate distribution. However, we can abuse notation slightly using the Dirac delta function, so that if P_x has density p_x, we can write

$$k(x, y) = \alpha p_x(y) + (1 - \alpha)\delta_x(y).$$

Exercise 51.

(a) Show that

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{A_1 \times \cdots \times A_n} p_1(x_1) \prod_{i=1}^{n-1} k(x_i, x_{i+1}) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n,$$

where p_1 is the density of P_1.

(b) Show further that the m-step transition probabilities can be written as

$$k^m(x_n, A) = \mathbb{P}(X_{n+m} \in A \mid X_n = x_n) = \int_A k^m(x_n, x_{n+m}) \, \mathrm{d}x_{n+m},$$

where

$$k^m(x_n, x_{n+m}) = \int_{\Omega^{m-1}} \prod_{i=0}^{m-1} k(x_{n+i}, x_{n+i+1}) \, \mathrm{d}x_{n+1} \cdots \mathrm{d}x_{n+m-1}$$

is the m-step kernel.

8.2 Key properties

Recall that our goal is to construct a Markov chain whose long term behaviour mimics that of the posterior distribution. It turns out that when it comes to study long term behaviour, some chains are better than others. In this section, we will go through some key properties a chain must satisfy for us to be able to study its asymptotic properties. We start with irreducibility and aperiodicity.

Definition 8.3. We say that a Markov chain with kernel k is *μ-irreducible* if for any point $x \in \Omega$ and any (μ-measurable) set $A \subseteq \Omega$ such that $\mu(A) > 0$, we have

$$k^m(x, A) > 0 \quad \text{for some } m \geq 1.$$

That is, for any subset A to which μ assigns a nonzero probability, there is a positive probability that we can reach A from x in a finite number of steps. Note in particular that the chain is automatically irreducible if $k(x, A) > 0$ for all x and A.

Exercise 52. Verify that the chains from Example 2 are irreducible.

Definition 8.4. We say that a Markov chain $\{X_n\}$ is *periodic* with period $d \geq 2$ if there exists a (measurable) partition[5] $\mathcal{S}_1, \ldots, \mathcal{S}_d$ of the state space Ω such that

$$\mathbb{P}(X_{n+m} \in \mathcal{S}_j \mid X_n \in \mathcal{S}_i) = \begin{cases} 1 & \text{if } j - i = m \mod d \\ 0 & \text{otherwise.} \end{cases}$$

The chain is *aperiodic* if it is not periodic.

That is, a Markov chain with period d moves through $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_d$ and back to $\mathcal{S}_1$ in loops with probability one.

Exercise 53.

(a) Verify that the chains from example 2 are aperiodic.

(b) Show that any chain with $k(x, \{x\}) > 0$ for all $x \in \Omega$ is aperiodic.

(c) Construct an irreducible chain which is 2-periodic.

Let us now turn to the question of examining the long term behaviour of a Markov chain. Suppose we run the chain for a very long time. For some chains, it will look like the states are drawn from some distribution, say π. For example, the empirical mean of the states in the chain, $(1/n)\sum_{i=1}^n X_i$, will converge to the mean $\mathbb{E}_\pi[X]$, etc. In order to make this notion precise, we need to introduce the idea of a stationary distribution.

Definition 8.5. We say that π is a *stationary distribution* (or *invariant distribution*) for the Markov chain $\{X_n\}$ if

$$\int_\Omega \pi(x)k(x, y)\,\mathrm{d}x = \pi(y) \tag{8.2}$$

for all $y \in \Omega$.

This definition tells us that if π is stationary for $\{X_n\}$ and $X_n \sim \pi$, then $X_{n+1} \sim \pi$. Indeed, letting π_{n+1} denote the density of X_{n+1}, we have

$$\pi_{n+1}(x_{n+1}) = \int_\Omega \pi(x_n)\pi_{n+1}(x_{n+1} \mid x_n)\,\mathrm{d}x_n = \int_\Omega \pi(x_n)k(x_n, x_{n+1})\,\mathrm{d}x_n = \pi(x_{n+1}),$$

[5]Recall that a partition of a set Ω is a collection of non-empty, disjoint subsets whose union is Ω.

so $\pi_{n+1} = \pi$, as required.

In practice, it can be very difficult to verify condition (8.2). Fortunately, there exists an only slightly stronger condition, called detailed balance, which is much easier to verify and always implies stationarity.

Definition 8.6. Let π be a probability distribution and let k be a kernel. We say that k and π satisfy *detailed balance* if

$$\pi(x)k(x,y) = \pi(y)k(y,x)$$

for all $x, y \in \Omega$.

Exercise 54. Show that if k and π satisfy detailed balance, then π is stationary for k.

For irreducible Markov chains, π-stationarity is a sufficient for guaranteeing a property called *ergodicity*, which basically asserts that the law of large numbers holds for the chain.

Theorem 8.1. *Suppose a Markov chain $\{X_n\}$ is π-irreducible and that its kernel k has stationary distribution π. Then almost surely, for any integrable function $f : \Omega \to \mathbb{R}$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \int_\Omega f(x)\pi(x)\,\mathrm{d}x = \mathbb{E}_\pi[(f(X)],$$

for π-almost all[6] starting values x.

If we have aperiodicity also then we can state an even stronger convergence result.

Theorem 8.2. *Suppose an aperiodic Markov chain $\{X_n\}$ is π-irreducible and that its kernel k has stationary distribution π. Then*

$$\lim_{n\to\infty} \int_\Omega |k^n(x,y) - \pi(y)|\,\mathrm{d}y = 0$$

for π-almost all starting values x.

Note that by Exercise 48, this means that

$$\lim_{n\to\infty} \delta(k^n(x,\cdot),\pi) = 0$$

for π-almost all starting values x. That is, the total variation distance between the chain and the distribution π converges to zero as $n \to \infty$.

The study of which conditions guarantee what modes of convergence for Markov chains is a rich field of study. For example, we might ask what condition is necessary to remove the "for π-almost all x" condition. It turns out that the answer is a condition known as *Harris recurrence*, but this is beyond the scope of the course.

The main takeaway for us is the following: If we have a general state space Markov chain $\{X_n\}$ which is (i) π-irreducible, (ii) aperiodic, and (iii) has π as a stationary distribution, then we can assert that the long term behaviour of the chain will mimic that of iid samples from π. In the next section, we will look at chains which are constructed such that the stationary distribution π will be the posterior distribution, so that if we run these chains long enough, they will converge to the posterior.

[6] That is, for all values x except from a subset $\mathcal{S}$ with $\pi(\mathcal{S}) = 0$.

8.3 The Metropolis-Hastings algorithm

Having gone through our introduction to general state space Markov chains, we are now ready to see how they can be applied in Bayesian analysis. The first phenomenon we shall look at is the Metropolis-Hastings algorithm, whose roots date back to Metropolis et al. (1953), but whose current form was first published by Hastings (1970).

Let us first through through the Metropolis-Hastings algorithm in its general form before looking at posterior inference in Bayesian statistics specifically. We now turn back to the lower case notation convention, as this makes everything less cluttered. Suppose we have a *target density* $\pi(x)$ that we wish to sample from which we can evaluate pointwise. We want to construct a Markov chain $\{x_n\}$ which converges to π. The first element of the chain is generated by some distribution $x_1 \sim P_1$. Suppose now that the current state in the chain is $x_n = x$. How do we generate x_{n+1}? The first step is to draw a sample x' from a *proposal distribution* with density $q(x' \mid x)$, which is allowed to depend on the current state x. This value is then either *accepted* with probability

$$\alpha(x' \mid x) = \min\left\{1, \frac{\pi(x')q(x \mid x')}{\pi(x)q(x' \mid x)}\right\},$$

in which case we set $x_{n+1} = x'$. Otherwise, it is *rejected*, and the chain stays where it was, so $x_{n+1} = x_n = x$. The minimum operator ensures that $\alpha(x' \mid x) \leq 1$, so it is a valid probability.

The Metropolis-Hastings algorithm is given in pseudo-code in Algorithm 1.

Exercise 55. Verify that step 6 in Algorithm 1 is correct even though it is possible that $\alpha(x' \mid x) > 1$ here.

To study the properties of the Metropolis-Hastings algorithm, we need to identify its kernel. The probability of accepting the proposed value, given that the current state of the chain is x, is given by

$$\alpha(x) = \int \alpha(x' \mid x)q(x' \mid x)\,\mathrm{d}x'. \tag{8.3}$$

In this case, we accept the proposal. Otherwise, with probability $1 - \alpha(x)$, we stay where we are. Hence the kernel takes the form

$$k(x, y) = \alpha(y \mid x)q(y \mid x) + [1 - \alpha(x)]\delta_x(y). \tag{8.4}$$

Exercise 56.

(a) Show that if the proposal $q(y \mid x)$ is a Gaussian distribution centred at x, $N(y; x, \sigma^2)$, then the Metropolis-Hastings kernel is irreducible.

(b) Show that the Metropolis-Hastings kernel is always aperiodic.

We now want to show that the target distribution π is stationary for the Metropolis-Hastings kernel. We do this by verifying the detailed balance equation,

$$\pi(x)k(x, y) = \pi(y)k(y, x).$$

Algorithm 1 Metropolis-Hastings

Require: Initial distribution P_1
Require: Proposal distribution q
Require: Target density π
Require: Sample size N

$x_1 \sim P_1$
$\mathcal{S} \leftarrow \{x_1\}$ ▷ Initialise sample
for $n = 1, \ldots, N-1$ **do**
    $x \leftarrow x_n$
    $x' \sim q(x' \mid x)$ ▷ Generate proposal
    Calculate acceptance probability

$$\alpha(x' \mid x) \leftarrow \frac{\pi(x')q(x \mid x')}{\pi(x)q(x' \mid x)}$$

    $u \sim \text{Uniform}[0, 1]$
    if $u \leq \alpha(x' \mid x)$ **then**
        $x_{n+1} \leftarrow x'$ ▷ Accept proposal
    else
        $x_{n+1} \leftarrow x$ ▷ Reject proposal
    end if
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{x_{n+1}\}$ ▷ Update sample
end for

It suffices to prove this when $x \neq y$. Indeed, if $x = y$ then detailed balance is trivially satisfied. When $x \neq y$, the kernel (8.4) simplifies to the first term $\alpha(y \mid x)q(y \mid x)$, and so we have

$$\begin{aligned}
\pi(x)k(x,y) &= \pi(x)\alpha(y \mid x)q(y \mid x) \\
&= \pi(x)q(y \mid x)\min\left\{1, \frac{\pi(y)q(x \mid y)}{\pi(x)q(y \mid x)}\right\} \\
&= \min\left\{\pi(x)q(y \mid x), \pi(y)q(x \mid y)\right\} \\
&= \pi(y)q(x \mid y)\min\left\{\frac{\pi(x)q(y \mid x)}{\pi(y)q(x \mid y)}, 1\right\} \\
&= \pi(y)q(x \mid y)\alpha(x \mid y) \\
&= \pi(y)k(y,x).
\end{aligned}$$

Hence, we see that detailed balance is verified, and so the target distribution π is stationary for the Metropolis-Hastings algorithm. Hence, if we choose a proposal which yields irreducibility of the chain, we know that the long term behaviour of the chain will resemble that of the target distribution π.

In Bayesian statistics, we want to apply the Metropolis-Hastings algorithm to posterior inference. That is, we want to create a chain $\theta_1, \theta_2, \ldots$ using the Metropolis-Hastings algorithm with some proposal $q(\theta' \mid \theta)$ to target the posterior distribution $\pi(\theta \mid y)$. At this point, you might be worried that this will require pointwise evaluations of the posterior in computing the acceptance probability $\alpha(\theta' \mid \theta)$. This, in turn, will require evaluations of the marginal likelihood $\pi(y)$, which is the thing we are trying to avoid in the first place. However, there is a crucial simplification taking place in the acceptance probability. By Bayes' theorem,

$$\begin{aligned}
\alpha(\theta' \mid \theta) &= \min\left\{1, \frac{\pi(\theta' \mid y)q(\theta \mid \theta')}{\pi(\theta \mid y)q(\theta' \mid \theta)}\right\} \\
&= \min\left\{1, \frac{[\pi(y \mid \theta')\pi(\theta')/\pi(y)]q(\theta \mid \theta')}{[\pi(y \mid \theta)\pi(\theta)/\pi(y)]q(\theta' \mid \theta)}\right\} \\
&= \min\left\{1, \frac{\pi(y \mid \theta')\pi(\theta')q(\theta \mid \theta')}{\pi(y \mid \theta)\pi(\theta)q(\theta' \mid \theta)}\right\},
\end{aligned}$$

so we see that the marginal likelihoods *cancel* and so we only require evaluations of the prior, the likelihood and the proposal, all of which are available to us.

Appreciate the generality and simplicity of the above construction. There are virtually no restrictions on the choice of prior and likelihood here. This is in stark contrast to conjugate priors, which we require to be of the same functional form as the likelihood.

Exercise 57. Go back to exercise 3. Letting y be the data given in part (d), use the Metropolis-Hastings algorithm to generate an approximate posterior sample using the following priors:

- $\theta \sim \text{Gamma}(0.1, 0.1)$,
- $\theta \sim \text{Uniform}[0.5, 50]$,
- $\theta \sim \text{LogNormal}(0, 1)$, which means that $\log \theta \sim \text{N}(0, 1)$.

Run the algorithm for $S = 10,000$ iterations with a simple Gaussian proposal. You will have to think about what happens whenever we propose a negative value $\theta' < 0$ here.

Exercise 58. Return to the `sinusoidal.csv` dataset, with the cubic regression model. Instead of employing empirical Bayes, impose now a simple prior on both hyperparameters α and β, say $\alpha, \beta \sim \text{Exp}(1)$, independently. Simulate the posterior distributions of α and β, and use these simulations to simulate the posterior distribution of the coefficients w and the predictive distribution of a new outcome y'. Compare your results with those you obtained when using empirical Bayes.

8.4 Gibbs sampling

We have already seen the Metropolis-Hastings algorithm, which is one of the most fundamental algorithms of MCMC. We shall now move on to another MCMC technique, known as Gibbs sampling, which is particularly useful when the target distribution of interest is multivariate. For example, suppose we have two random variables x and y. There are many situations where it is difficult to sample from the joint distribution $\pi(x, y)$, but it is straightforward to sample from the conditional distributions $\pi(x \mid y)$ and $\pi(y \mid x)$. In the context of Bayesian inference, this would typically manifest itself as a multidimensional parameter $\theta = (\theta_1, \ldots, \theta_p)$ with a complicated joint posterior $\pi(\theta \mid y)$, but where the conditional distributions

$$\pi(\theta_j \mid \theta_{-j}, y)$$

are easy to sample from. Here, θ_{-j} is the vector we obtain by deleting component j from θ. That is,

$$\theta_{-j} = (\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p).$$

Exercise 59. Suppose we have n counts $y_1, \ldots, y_n$ which we assume are independent and follow a Poisson distribution. However, we suspect there is a *changepoint* τ in the data, so that the Poisson parameter before and after the point τ are different. That is, we assume

$$\begin{aligned} y_i \mid \{\tau, \lambda_l\} &\sim \text{Poisson}(\lambda_l) \quad \text{independently for } i = 1, \ldots, \tau, \\ y_i \mid \{\tau, \lambda_r\} &\sim \text{Poisson}(\lambda_r) \quad \text{independently for } i = \tau + 1, \ldots, n. \end{aligned}$$

The parameters of the model are therefore the changepoint τ and the two Poisson parameters λ_l and λ_r. As priors, we let

$$\begin{aligned} \tau &\sim \text{Uniform}\{1, 2, \ldots, n\}, \\ \lambda_1 &\sim \text{Gamma}(\alpha, \beta), \\ \lambda_2 &\sim \text{Gamma}(\alpha, \beta), \end{aligned}$$

independently of each other. The aim of the problem is to infer the location of the changepoint τ and the Poisson parameters λ_l, λ_r on either side.

(a) Write down the log posterior distribution for $\theta = (\lambda_l, \lambda_r, \tau)$ given the data y, up to a constant.

(b) The joint posterior you found in the previous exercise is difficult to sample from directly. Find the three conditional distributions of λ_l and λ_r given everything else,

$$\pi(\lambda_l \mid \lambda_r, \tau, y), \quad \pi(\lambda_r \mid \lambda_l, \tau, y), \quad \pi(\tau \mid \lambda_l, \lambda_r, y).$$

We are now ready to explain the Gibbs sampling algorithm. To keep the notation uncluttered, let us stick to the case where $p = 2$, so we have two random variables x and y whose joint distribution $\pi_{X,Y}(x, y)$ is our target. We assume that sampling directly from $\pi_{X,Y}$ is difficult, but that we can sample from the conditional distributions $\pi_{X|Y}$ and $\pi_{Y|X}$. Gibbs sampling basically says that if we continuously sample from these conditional distributions repeatedly, then we eventually converge to the joint distribution $\pi_{X,Y}$. The algorithm (for $p = 2$) is given in Algorithm 2

Algorithm 2 Gibbs (two-dimensional)

Require: Initial sample (x_1, y_1).
Require: Conditional distributions $\pi_{X|Y}, \pi_{Y|X}$.
Require: Sample size N
1: $\mathcal{S} \leftarrow \{(x_1, y_1)\}$ ▷ Initialise sample
2: **for** $n = 1, \dots, N-1$ **do**
3: $\quad x_{n+1} \sim \pi_{X|Y}(\cdot \mid y_n)$
4: $\quad y_{n+1} \sim \pi_{Y|X}(\cdot \mid x_{n+1})$
5: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{(x_{n+1}, y_{n+1})\}$ ▷ Update sample
6: **end for**

Note that when we sample y_{n+1} from the conditional distribution in step 4, we condition on $X = x_{n+1}$, not x_n. Thus, Gibbs sampling sequentially updates all the variables, conditioning on the latest information.

Exercise 60. Generalise Algorithm 2 to the three-dimensional case.

Note that there is no accept/reject step in Gibbs sampling. We always accept the proposed sample. From Algorithm 2, we see that Gibbs sampling creates a Markov chain with the kernel

$$k((x, y), (x', y')) = \pi_{X|Y}(x' \mid y)\pi_{Y|X}(y' \mid x'). \tag{8.5}$$

Proposition 8.1. *The joint distribution $\pi_{X,Y}$ is stationary for the Gibbs sampler.*

Proof. We have

$$\begin{aligned}
\int \pi_{X,Y}(x, y)k((x, y), (x', y'))\,\mathrm{d}x\,\mathrm{d}y &= \int \pi_{X,Y}(x, y)\pi_{X|Y}(x' \mid y)\pi_{Y|X}(y' \mid x')\,\mathrm{d}x\mathrm{d}y \\
&= \pi_{Y|X}(y' \mid x') \int \pi_Y(y)\pi_{X|Y}(x' \mid y)\,\mathrm{d}y \\
&= \pi_{Y|X}(y' \mid x') \int \pi_{X,Y}(x', y)\,\mathrm{d}y \\
&= \pi_{Y|X}(y' \mid x')\pi_X(x') \\
&= \pi_{X,Y}(x', y'),
\end{aligned}$$

as required. □

Hence, provided we have irreducibility and aperiodicity, the Gibbs sampler will converge to the joint distribution $\pi_{X,Y}$. Note, however, that in general, the Gibbs kernel does not satisfy detailed balance in general. On the other hand, the Gibbs kernel consists of two parts, namely the x-update and the y-update[7]. Separately, these updates have the kernels

$$\begin{aligned} k_1((x,y),(x',y')) &= \pi_{X|Y}(x' \mid y)\delta_y(y'), \\ k_2((x,y),(x',y')) &= \pi_{Y|X}(y' \mid x)\delta_x(x'). \end{aligned} \tag{8.6}$$

Exercise 61. Show that the joint distribution $\pi_{X,Y}$ satisfies detailed balance for both kernels k_1 and k_2.

A natural question to ask at this point is whether there is any direct link between Gibbs sampling and the Metropolis-Hastings algorithm. In order to answer this question, we need to consider what happens when we combine two Markov chain kernels k_1 and k_2 with the same stationary distribution π.

Proposition 8.2. *Let $k_1(x,y)$ and $k_2(x,y)$ be two Markov chain kernels with the same stationary distribution $\pi(x)$. Then π is also stationary for the combination of k_1 and k_2, defined by*

$$k_1 k_2(x,y) = \int k_1(x,x')k_2(x',y)\,\mathrm{d}x'.$$

Proof. We have

$$\begin{aligned} \int \pi(x)k_1k_2(x,y)\,\mathrm{d}x &= \int\int \pi(x)k_1(x,x')k_2(x',y)\,\mathrm{d}x'\,\mathrm{d}x \\ &= \int \underbrace{\int \pi(x)k_1(x,x')\,\mathrm{d}x}_{\pi(x')} k_2(x',y)\,\mathrm{d}x' \\ &= \int \pi(x')k_2(x',y)\,\mathrm{d}x' \\ &= \pi(y), \end{aligned}$$

as required. □

The above proof also generalises to the case where $p \geq 3$. We thus see that if we have multiple Markov chain kernels with the same stationary distribution π, then we can combine them to obtain another chain whose stationary distribution is still π.

It turns out that we can view the separate Gibbs kernels (8.6) as Metropolis-Hastings kernels, which means that the full Gibbs kernel (8.5) corresponds to a combination of Metropolis-Hastings kernels. Indeed, the separate kernels take the form of a proposal. But if they are Metropolis-Hastings kernels, why is there no accept/reject step in the kernels? To see why, consider the

[7] In the general p-dimensional case, there are p updates

acceptance probability for the separate Gibbs kernels, e.g. k_1. We have

$$\begin{aligned}
\alpha((x', y') \mid (x, y)) &= \min\left\{1, \frac{\pi_{X,Y}(x', y')\pi_{X|Y}(x \mid y')\delta_{y'}(y)}{\pi_{X,Y}(x, y)\pi_{X|Y}(x' \mid y)\delta_y(y')}\right\} \\
&= \min\left\{1, \frac{\pi_{X,Y}(x', y')\pi_{X,Y}(x, y)\delta_{y'}(y)/\pi(y)}{\pi_{X,Y}(x, y)\pi_{X,Y}(x', y')\delta_y(y')/\pi(y)}\right\} \\
&= \min\{1, 1\} \\
&= 1,
\end{aligned}$$

so the acceptance probability is always 1 and we therefore always accepts the proposed sample. This concludes our discussion about the connection between the Gibbs sampler and the Metropolis-Hastings algorithm.

Exercise 62.

(a) Let $k_1(x, y)$ and $k_2(x, y)$ be two Markov chain kernels with the same stationary distribution $\pi(x)$. For any fixed number $\gamma \in (0, 1)$, show that π is also stationary for the mixture kernel

$$k(x, y) = \gamma k_1(x, y) + (1 - \gamma)k_2(x, y).$$

(b) Generalise the above argument to p kernels. That is, suppose π is a stationary distribution for γ separate kernels $k_1, \ldots, k_p$, and let $\gamma_1, \ldots, \gamma_p$ be numbers satisfying $\gamma_j \geq 0$ for all $j = 1, \ldots, p$, and $\sum_{j=1}^{p} \gamma_j = 1$. Show that π is also stationary for the kernel

$$k(x, y) = \sum_{j=1}^{p} \gamma_j k_j(x, y).$$

(c) Consider the random scan Gibbs sampler, in which we choose uniformly at random which component to update at each iteration. For example, in the two-dimensional case, at each iteration, we either sample from $\pi_{X|Y}$ or $\pi_{Y|X}$, each with probability $1/2$. Conclude from parts (a) and (b) that this sampler has the joint distribution $\pi_{X,Y}$ as its stationary distribution.

Exercise 63. Return to Exercise 59. Apply this model to the `mining.csv` dataset, which lists the number of coal mining disasters in the UK between 1851 and 1962. Set $\alpha = \beta = 1$.

(a) Use Gibbs sampling to get a posterior sample of the parameters $\tau, \lambda_1, \lambda_2$, using $S = 1000$ iterations. What is the posterior mode of τ? What does this value represent?

(b) Consider also the model where there is no changepoint, so there is only a single parameter λ across the entire dataset. Find the posterior distribution of λ given the data (using e.g. exercise 3) and plot its density in a histogram. Compare your results to those you obtained in part (a).

(c) What ingredient is missing to compare the models in parts (a) and (b), i.e. to address whether there is a changepoint present in the data?

8.5 Convergence diagnostics

In the previous section, we saw how to set up and run MCMC algorithms for posterior inference. In this section, we shall see how we can assess the quality of their output via *convergence diagnostics*. This is important, because MCMC is not an exact algorithm for posterior sampling, it yields only an approximation. We therefore need to develop tools to assess the quality of that approximation.

Let us think about the two things that could go wrong in an MCMC algorithm, as a consequence of it being an approximate sampling technique. Firstly, the chain may not have reached equilibrium. Note that even though we have a theoretical guarantee that the chain will converge to its stationary distribution asymptotically, we have no guarantee that this will happen in finite time for a finite chain. Secondly, the correlation between terms in the chain might be high, and will affect the variance of our estimates. Indeed, suppose we want to estimate some function $f(X)$ from an MCMC output $X_1, X_2, \ldots, X_S$. Then we would use the empirical estimate

$$\bar{f}_S = \frac{1}{S}\sum_{s=1}^{S} f(X_s), \tag{8.7}$$

and so we need to ensure that S is large enough to make the variance of $\bar{f}_S$ small.

A bad practice, which many students and even some researchers are guilty of, is to simply state for how long the algorithm was run. If someone has run their algorithm for $S = 100,000$ iterations, we still have no information about the convergence and autocorrelation properties of the chain.

Let us first look at how we mitigate the problem of convergence to equilibrium. A standard solution for this is to delete a small proportion of the chain, usually called *burn-in*. To see how long burn-in should be, we usually draw *trace plots* of the chain (that is, a plot of the values taken by the chain over time), which allows us to see where mixing starts to improve.

Exercise 64. Return to your analysis from exercise 57 with the uniform prior. Make trace plots of the samples, and see how they change as you vary the variance of the Gaussian proposal distribution. What do you observe?

The next question is how to assess the autocorrelation of the chain. The simplest way to do this is to plot the autocorrelation as a function of lag. More specifically, suppose we have a Markov chain $X_1, \ldots, X_S$, starting at equilibrium. That is, through the rest of this section, we assume that $X_1 \sim \pi$, the stationary distribution. Suppose we want to estimate a function $f(X)$. For $r > s$, define the *correlation of f at lag r* to be

$$\rho_r = \frac{\mathrm{cov}(f(X_s), f(X_{s+r}))}{\mathrm{Var}(f(X_s))}. \tag{8.8}$$

Note that ρ_r only depends on r since the chain is assumed to start at its stationary distribution. In order to estimate ρ_r, we estimate the numerator and denominator of (8.8) separately. For the numerator $\gamma_r = \mathrm{cov}(f(X_s), f(X_{s+r}))$, use

$$\hat{\gamma}_r = \frac{1}{S}\sum_{i=1}^{S-r}(f(X_i) - \bar{f}_S)(f(X_{i+r}) - \bar{f}_S), \tag{8.9}$$

where $\bar{f}_S$ is defined by (8.7). There are reasons for why we divide by S rather than $S - r$ in (8.9), but we will not go into those here. Plugging $r = 0$ into (8.9) gives $\hat{\gamma}_0$, an estimate of $\sigma^2 = \text{Var}(f(X_s))$. Thus, our estimate for ρ_r is $\hat{\gamma}_s/\hat{\gamma}_0$. Note that σ^2 does not depend on s since we assume that the chain is stationary to begin with.

Exercise 65. Continue the analysis from exercise 57. This time, plot the autocorrelation of the samples as a function of the lag. Try different values for the variance of the proposal. What do you observe?

The derivation of ρ_r leads another summary of the quality of the MCMC output, called the *effective sample size* (ESS). In short, the ESS tells us how many iid samples from the target distribution we would have needed to achieve the same variance for $\bar{f}_S$ as when we use samples from the Markov chain. That is,

$$\text{Var}(\bar{f}_S) = \frac{\text{Var}(f(X_s))}{\text{ESS}} = \frac{\sigma^2}{\text{ESS}}. \tag{8.10}$$

We therefore need an estimate of $\text{Var}(\bar{f}_S)$ to estimate ESS. We have

$$\begin{aligned}
\text{Var}(\bar{f}_S) &= \frac{1}{S^2} \sum_{i=1}^{S} \sum_{j=1}^{S} \text{cov}(f(X_i), f(X_j)) \\
&= \frac{\sigma^2}{S^2} \sum_{i=1}^{S} \sum_{j=1}^{S} \rho_{|i-j|} \\
&= \frac{\sigma^2}{S} \left[1 + 2 \sum_{r=1}^{S-1} \left(1 - \frac{r}{S} \right) \rho_r \right],
\end{aligned} \tag{8.11}$$

and thus we obtain an estimate for ESS,

$$\widehat{\text{ESS}} = S \left[1 + 2 \sum_{r=1}^{S-1} \left(1 - \frac{r}{S} \right) \hat{\rho}_r \right]^{-1}. \tag{8.12}$$

It should be noted that there are more sophisticated ways of calculating the effective sample size, so you may observe that standard packages in `Python` or `R` gives different answers than (8.12). However, the above derivation, although reasonable simple, outlines the meaning of the effective sample size and how to estimate it from data.

Exercise 66. Go once more back to your output from exercise 57. Calculate the effective sample size for different values of the variance of the proposal.

To summarise, we have the following tools for mitigating convergence and mixing issues when running MCMC:

1. **Multiple runs**: Run the algorithm multiple times with different inputs, and check that we obtain the same results each time, to our desired accuracy.

2. **Trace-plots**: Plot the output of the algorithm. Check that we have reached the stationary distribution and that the chain is mixing well.

3. **Burn-in**: Delete a small proportion of the samples from the beginning of the chain which are not from the stationary distribution.

4. **Autocorrelation plots**: Plot the autocorrelation function against lag and verify that we have a quick drop-off.

5. **Effective sample size**: Make sure this is sufficiently large.

9 References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

P. Diaconis and D. Freedman. Finite exchangeable sequences. *The Annals of Probability*, 8:745–764, 1980.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

Geoff Nicholls. SC7 Bayes Methods lecture notes. *Oxford University*, 2023.

Christian P. Robert. *The Bayesian Choice*. Springer, 2001.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.