CHAPTER 2

Linear Modelling: A Maximum Likelihood Approach

In the previous chapter, we introduced the idea of learning the parameters of a model by defining and minimising a loss function. By the end of this chapter, we will have derived exactly the same equation for the optimal parameter values from a different starting point. In particular, we will explicitly model the **noise** (the errors between the model and the observations) in the data by incorporating a **random variable**. We will demonstrate the considerable advantages of incorporating a noise term into our model. A large section of this chapter (Sections 2.2 to 2.5) is an introduction to random variables and **probability** which can be skipped by readers already familiar with these concepts.

2.1 ERRORS AS NOISE

In Figure 1.5 we saw the result of minimising the squared loss function to model the Olympic 100 m data with a linear model. The linear model appears to capture an interesting downward trend but is unable to explain each data point perfectly – there are errors between the model and the true values. These errors are highlighted in Figure 2.1.

When building our model, we assumed that there was a linear relationship between years and winning times. This model appeared to capture the general trend in the data whilst ignoring the, sometimes large, deviation between the model and the observed data. From a modelling perspective, ignoring these errors is hard to defend. If we know they are going to be present, we should make an effort to build them into our model.

In this chapter we will see the benefits of explicitly modelling these errors. In particular, it allows us to express the level of uncertainty in our estimate of the model parameters, $\mathbf{w}$ – if we change $\mathbf{w}$ a bit, do we still have a *good* model? This in turn allows us to express a degree of uncertainty in our predictions – 'we believe the winning time will be between a and b' rather than 'we believe the winning time will be exactly c'.

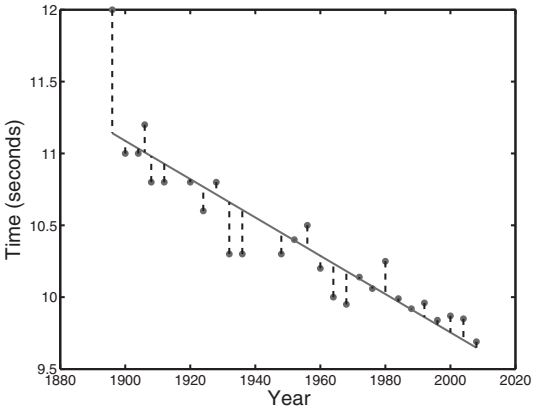


FIGURE 2.1 Linear fit to the Olympic men's 100 m data with errors highlighted.

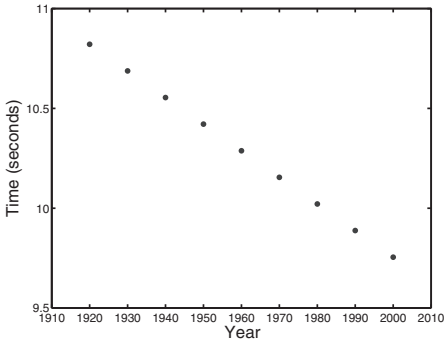


FIGURE 2.2 Dataset generated from the linear model.

2.1.1 Thinking generatively

The process that generated this particular dataset is very complex – we couldn't even begin to make a near-perfect model of one sprinter and the events surrounding his preparation and performance, let alone several of them *and* all of the other factors. However, it is still useful to think of our modelling problem as a **generative** one: can we build a model that could be used to create (or generate) a dataset that *looks* like ours? Although we are happy to accept that this isn't in fact how the data were generated, we shall see that this is a useful strategy.

How might we go about generating data from our current model? We have an equation, $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}$, that, if we plug in the values for $\mathbf{w}$ that we found in the previous chapter could be used to *generate* a winning time for any particular year. Figure 2.2 shows winning times generated in this way for a number of years between

1920 and 2000. It doesn't look much like the data in Figure 2.1. To make it more realistic, we need to add some errors. Examining Figure 2.1, we notice a couple of important features of the errors:

1. They are different at each year. Some are positive, some negative and they all have different magnitudes.
2. There does not seem to be any obvious relationship between the size (or direction) of the error and the year. The error does not appear to be a function of x, the Olympic year.

If we had a method for generating a random amount of time (in seconds) that could be either positive or negative and was, on average, roughly the same size as the errors in Figure 2.1, we could generate one such value for each data point we wished to generate, and add it to $\mathbf{w}^\mathsf{T}\mathbf{x}$. The tools that we will need to incorporate this variability into our model come from **statistics**. In the next section we will introduce random variables and some of the ways in which they can be manipulated. Readers already familiar with this can jump straight to Section 2.7.

2.2 RANDOM VARIABLES AND PROBABILITY

Any model we build will be a simplification of the real system that generated the data we observe. This will lead to a discrepancy between the model and reality which the tools presented in this section will help us to model and understand. As we must start with the basics, it may at first seem slightly disconnected from the particular problem of adding errors to our generated 100 m data and being able to express uncertainties in our predictions, but the connection will become clearer as we progress.

2.2.1 Random variables

The equation

$$y = 5x - 2$$

has two variables, x and y. If we were given a value for one (say $y = 8$) we could solve for the other ($x = 2$). *Random* variables are very different. They allow us to assign numerical values to **random events**. For example, I would like to model the outcome of a coin toss. As a starting point, I create a variable called X to which I will assign the value 1 if the coin lands heads and 0 if it lands tails. X is a random variable – the 'variable' part describes the fact that it can take a number of different values (in this case, 0 and 1) and the 'random' part is so called because we don't know what value X will take before the coin toss takes place – we couldn't express the outcome as a function of standard variables (e.g. $y = 5x - 2$). It is a common convention to use upper-case letters to describe random variables and lower-case ones for possible values that the random variable can take.

There are two types of random variable and they must be treated slightly differently. **Discrete** random variables are the easiest to conceptualise, as they are used for random events for which we can systematically list (or count) all possible outcomes. A discrete random variable could be used to, for example, describe a coin toss (possible outcomes are 0 and 1) or the rolling of a die (1, 2, 3, 4, 5 or 6). The collection of possible outcomes is known as the **sample space**.

It might seem that being able to systematically write all of the possible events in order should be true of almost anything. In fact, there are many possible events for which this is not the case. Taking our Olympic 100 m example and assuming that the winning time is going to be between 9 and 10 seconds, we could attempt to systematically write down all possibilities:

$$9,9.1,9.2,\ldots$$

at which point, we realise that we've missed some out (all the ones between 9 and 9.1 for example) so we start again:

$$9,9.01,9.02,\ldots,9,1,\ldots$$

But what about all the ones between 9 and 9.01? Starting a third time,

$$9,9.001,9.002,9.003,\ldots,9.01,\ldots \text{etc.}$$

The possible outcomes of this event cannot be systematically listed (after writing down any two values, someone could point out the missing ones in between). For events like this, we must use **continuous** random variables.

Table 2.1 gives examples of events or quantities that we might wish to model with random variables and whether or not they are discrete or continuous. We will now introduce some important concepts through discrete random variables before extending the ideas to the continuous case.

TABLE 2.1 Events we might want to model with random variables.

Process	Discrete or continuous
Toss of a coin	Discrete
Roll of a die	Discrete
Outcome of a 100 m race	Continuous
Failure of a node in a computer network	Discrete
Outcome of a court case	Discrete
Height of a human	Continuous
Mass of a pebble	Continuous
Score in a football match	Discrete
Errors in our 100 m linear regression model	See Exercise 2.1

2.2.2 Probability and distributions

Let Y be a random variable that represents the toss of a coin. If the coin lands heads, $Y = 1$ and if tails, $Y = 0$. To model this event (the coin toss), we need to be able to quantify how likely either outcome is. For discrete random variables, we do this by defining the probabilities of the different outcomes. One intuitive way of thinking about the probability of a particular outcome is to imagine that it represents the proportion of times this outcome would happen if the event were to be repeated many times. If a fair (i.e. not biased to land either way) coin were tossed 1000 times, we might expect to see heads roughly half of the time (and tails the rest of time). It would seem sensible to define the probability of seeing a head, which we will denote

$P(Y = 1)$, as a half, or 0.5. If the coin doesn't land as a head, it lands as a tail (there are only two options in our sample space) and so the proportion of tails must be one minus the proportion of heads. Therefore, $P(Y = 0) = 1 - P(Y = 1) = 0.5$.

Conceptualising probabilities as the proportion of times a particular outcome would occur if an event were repeated many times is not the only way they can be thought of. It is not always the most natural analogy, particularly for events that can only occur once. It will be sufficient for our needs, but the reader is encouraged to investigate this interesting area further.

From our short discussion on proportions, we can write down two important rules governing probabilities:

- Probabilities must be greater than or equal to 0 (a proportion cannot be negative) and less than or equal to 1.
- The sum of the probabilities of each possible individual outcome must be equal to 1.

$$\begin{aligned} \text{e.g. for a coin:} \quad & P(Y = 1) + P(Y = 0) = 1 \\ \text{For a die:} \quad & P(Y = 1) + P(Y = 2) + \cdots + P(Y = 6) = 1 \end{aligned}$$

The mathematical equivalents of these statements are:

$$0 \leq P(Y = y) \leq 1, \tag{2.1}$$

$$\sum_y P(Y = y) = 1 \tag{2.2}$$

where the lower case y is used, by convention, to represent values that the random variable Y can take. Note that we will often need to write summations over the values that a random variable can take – to keep notation concise, $\sum_y$ will be used to denote a sum over all of the possible values that can be taken by a random variable Y.

$P(Y = y)$ is a scalar value – the probability that the random variable Y has outcome y. This notation can sometimes become unwieldy and so we will sometimes use the following shorthand:

$$P(Y = y) = P(y).$$

The set of all of the possible outcomes (all of the ys) and their probabilities, $P(y)$, is known as a probability **distribution**. It tells us how the total probability (1) is distributed (or shared out) over all possible outcomes.

Often, we can use Equations 2.1 and 2.2 to define probabilities based on some fundamental assumptions. For example, in the coin example, we might assume that the two outcomes are equally likely: $P(Y = 1) = P(Y = 0) = r$. Plugging this into Equation 2.2 and remembering that r must lie between 0 and 1 (Equation 2.1), we can use some algebra to work out the value of r (See Exercise 2.2):

$$\begin{aligned} P(Y = 0) + P(Y = 1) &= 1 \\ 2r &= 1 \\ r &= \frac{1}{2}. \end{aligned}$$

2.2.3 Adding probabilities

Let Y be a random variable for modelling the outcome of rolling a fair die. If we encode our assumption that the die is fair by assuming that all outcomes are equally likely, we know enough from the previous section to compute the probabilities of each possible outcome – 1, 2, 3, 4, 5 or 6. The die is rolled and the result is a 4. If it is rolled again, what is the probability of the result being lower than 4? Maybe we are playing a betting game and want to know whether the odds on offer are acceptable. The outcomes that are lower than 4 are 1, 2 and 3, suggesting that we need to be able to calculate the probability that the die lands 1 *or* 2 *or* 3. If the die were to be rolled many times, we could compute the proportion of times that this was the case. The proportion of times the die lands 1 or 2 or 3 is equal to the proportion of times the die lands 1 *plus* the proportion of times the die lands 2 *plus* the proportion of times the die lands 3. This leads us to the following additive law of probability:

$$P(Y < 4) = P(Y = 1) + P(Y = 2) + P(Y = 3).$$

Exactly the same result applies if the outcomes in which I'm interested are not in order. For example, the probability that I roll a 1 or a 6 would be $P(Y = 1)+P(Y = 6)$. It is also not just restricted to individual outcomes. For example, the probability that I don't roll a 4 could be computed as

$$\begin{aligned} P(Y \neq 4) &= P(Y < 4) + P(Y > 4) \\ &= P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 5) + P(Y = 6). \end{aligned}$$

As an aside, it is worth remembering that there is generally more than one way to compute any probability. In this example, it would in fact be easier to make use of Equation 2.2 and compute

$$\begin{aligned} P(Y \neq 4) + P(Y = 4) &= 1 \\ P(Y \neq 4) &= 1 - P(Y = 4). \end{aligned}$$

2.2.4 Conditional probabilities

Often one event will affect the outcome of another. For example, I toss a coin and then tell you what the result was (you cannot see the coin). There are two events – the first is tossing the coin, the second is me communicating the outcome of the coin toss to you. Let's assume that these two events are represented by two random variables. X is 1 if the coin lands heads and 0 if tails. Y is 1 if I tell you heads, and zero if I tell you tails. Unless I'm behaving very strangely, the outcome of Y will depend on the outcome of X. We can use **conditional probabilities** to express the probability that Y takes a particular value given that X has taken a particular value. We express this as

$$P(Y = y|X = x), \tag{2.3}$$

which reads as the probability that Y has the outcome y given that X has the outcome x. As for unconditional probabilities, we will also make use of the following shorthand:

$$P(Y = y|X = x) = P(y|x).$$

In our example, if we assume that I always tell the truth, the probability that I say heads if the coin lands heads is 1 (it will always happen):

$$P(Y = 1|X = 1) = 1.$$

Similarly for tails:

$$P(Y = 0|X = 0) = 1.$$

Using Equation 2.2 and these probabilities, we can deduce $P(Y = 0|X = 1)$ and $P(Y = 1|X = 0)$:

$$\begin{aligned} P(Y = 0|X = 1) + P(Y = 1|X = 1) &= 1 \\ P(Y = 0|X = 1) = 1 - P(Y = 1|X = 1) &= 0. \\ P(Y = 1|X = 0) + P(Y = 0|X = 0) &= 1 \\ P(Y = 1|X = 0) = 1 - P(Y = 1|X = 1) &= 0. \end{aligned}$$

Things get a bit more interesting if I'm not so truthful. Let's assume that, if the coin lands tails, I always tell the truth but the proportion of times I tell the truth if it lands heads is 0.8. This implies that, if the coin lands heads, I'll say heads with probability 0.8 and tails with probability 0.2. The full list of conditional probabilities under this assumption is

$$\begin{aligned} P(Y = 1|X = 1) &= 0.8 \\ P(Y = 0|X = 1) &= 0.2 \\ P(Y = 1|X = 0) &= 0 \\ P(Y = 0|X = 0) &= 1. \end{aligned}$$

Just as in non-conditional probabilities, Equation 2.2 must be satisfied, i.e. $\sum_y P(Y = y|X = x) = 1$. We can check this for the values just computed:

$$\begin{aligned} \sum_y P(Y = y|X = 1) &= P(Y = 1|X = 1) + P(Y = 0|X = 1) = 0.8 + 0.2 = 1 \\ \sum_y P(Y = y|X = 0) &= P(Y = 1|X = 0) + P(Y = 0|X = 0) = 0 + 1 = 1 \end{aligned}$$

Armed with the conditional probabilities and assuming that $P(X = 1) = P(X = 0) = 0.5$ (i.e., our coin is fair), we might ask 'what is the probability that the coin lands heads *and* I say heads?' This is different from $P(Y = 1|X = 1)$; the conditional distribution assumes that $X = 1$ has already happened and the only uncertainty that remains is what will happen with Y whereas my question concerns both events. If neither has happened, what is the probability that they will both have a particular outcome? Other interesting quantities that we may want to evaluate are $P(Y = 1)$ and $P(Y = 0)$, the probability that I say heads or tails. To compute any of these, we need to understand probabilities and distributions of more than one variable.

2.2.5 Joint probabilities

Given two (or more) random variables, we may wish to know the probability that they each take a particular value. Continuing our previous coin tossing example, we might want to know the probability that the coin shows heads *and* I say heads or the probability that the coin shows heads *and* I say tails. These are joint probabilities and are denoted as

$$P(Y = y, X = x) \tag{2.4}$$

(or, in functional form, $p(y, x)$). How we deal with these joint distributions depends on whether or not the random variables are *dependent*. In our example, Y (what I say) depends on X (how the coin lands). This is the case even when I'm not always being truthful – how the coin lands determines how I *decide* what to say. If there is no dependence between the variables (e.g. if two random variables represent different coin tosses, the outcome of one is unlikely to affect the outcome of the other), the **joint probability** can be computed by multiplying the individual probabilities together:

$$P(Y = y, X = x) = P(Y = y) \times P(X = x).$$

The probability that Y takes value y *and* X takes value x is equal to the probability that Y takes value y multiplied by the probability that X takes value x. More generally (and here we switch to the functional form – $p(y_1, \ldots, y_J)$ rather than $P(Y_1 = y_1, \ldots Y_J = y_j)$ for convenience), for a family of J random variables $Y_1, \ldots, Y_J$,

$$P(y_1, y_2, \ldots, y_J) = P(y_1) \times p(y_2) \times \cdots \times P(y_J) = \prod_{j=1}^{J} P(y_j). \tag{2.5}$$

If the events are dependent, we cannot decompose the joint probability in this manner. However, if we can create conditional distributions, we can decompose the joint probability using the following definitions:

$$P(Y = y, X = x) = P(Y = y|X = x) \times P(X = x) \tag{2.6}$$

or as

$$P(Y = y, X = x) = P(X = x|Y = y) \times P(Y = y). \tag{2.7}$$

So, the probability that the coin lands heads *and* I say heads is

$$P(Y = 1, X = 1) = P(Y = 1|X = 1) \times P(X = 1) = 0.8 \times 0.5 = 0.4$$

or, in other words, if we repeated this many times, the proportion of times that the coin landed heads and I said heads is 0.4. The fact that I occasionally lie when the coin shows heads has reduced the probability that you will hear heads from 0.5 (if I were always honest) to 0.4.

There are four possible combinations of X and Y and hence four possible outcomes of the event. Equation 2.2 tells us that, if we sum the probabilities of all four of these events, we should get 1:

$$\sum_{x,y} P(X = x, Y = y) = 1. \tag{2.8}$$

(Note that $\sum_{x,y}$ corresponds to a summation over all possible combinations of x and y). We can test this by working them all out from Equation 2.6. We already

know $P(X = 1, Y = 1) = 0.4$. The others are

$$\begin{aligned} P(Y = 0, X = 1) &= P(Y = 0|X = 1)P(X = 1) = 0.2 \times 0.5 = 0.1 \\ P(Y = 1, X = 0) &= P(Y = 1|X = 0)P(X = 0) = 0 \times 0.5 = 0 \\ P(Y = 0, X = 0) &= P(Y = 0|X = 0)P(X = 0) = 1 \times 0.5 = 0.5. \end{aligned}$$

Adding these together gives $0.4 + 0.1 + 0 + 0.5 = 1$, as required.

Before we move on, we will quickly consider these three values. The first (0.1) gives the probability that I say tails and the coin lands heads. This has increased from the truthful case (it would be zero if I always told the truth) because I sometimes lie if the coin is heads. The second (0) is the probability that I say heads when the coin is actually tails. This is zero because I never lie if the coin is tails. The final value is the probability that I say tails and the coin lands tails. This is 0.5 – the coin lands tails half the time and if it does, I always tell the truth.

2.2.6 Marginalisation

If you recorded the proportion of times I said heads or tails, you would in effect be computing $P(Y = 1)$ and $P(Y = 0)$. These expressions do not involve X – they just refer to what I say. $P(Y = y)$ can be obtained by **marginalising** out X from the joint distribution $P(Y = y, X = x)$. This is done by summing the joint probabilities over all possible values of X:

$$P(Y = y) = \sum_x P(Y = y, X = x). \tag{2.9}$$

In our coin example, X can take one of two values, so this summation would become

$$P(Y = y) = P(Y = y, X = 0) + P(Y = y, X = 1).$$

In general, for joint probabilities of J random variables, to get $P(Y_j = y_j)$, the marginal distribution of one of them is given by

$$P(Y_j = y_j) = P(y_j) = \sum_{y_1,\ldots,y_{j-1},y_{j+1},\ldots,y_J} P(y_1, \ldots, y_J). \tag{2.10}$$

The summation in this expression looks a bit strange. It is summing over all combinations of the remaining $J - 1$ variables (y_j is missing). For example, if $J = 3$ and each variable can take only the values 0 or 1, to compute $P(Y_1 = y_1) = p(y_1)$ would require summation over four different combinations of y_2 and y_3:

y_2	y_3
0	0
0	1
1	0
1	1

If $J = 4$, this increases to 8:

y_2	y_3	y_4
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

In general, for binary variables, the number of combinations will be 2^{J-1}, which rapidly increases with J. If our random variables have more than two outcomes, it gets even worse (e.g. 6^{J-1} for a die). Marginalisation is important in some probabilistic areas of Machine Learning and can be very challenging, inspiring approximation methods such as those that we shall see in Chapter 4.

Returning to our coin example, $P(Y = 1)$ is

$$\begin{aligned} P(Y=1) &= \sum_x P(Y=1, X=x) \\ &= P(Y=1, X=0) + P(Y=1, X=1) \\ &= 0 + 0.4 = 0.4 \end{aligned}$$

and $P(Y = 0)$ is

$$\begin{aligned} P(Y=0) &= \sum_x P(Y=0, X=x) \\ &= P(Y=0, X=0) + P(Y=0, X=1) \\ &= 0.5 + 0.1 = 0.6. \end{aligned}$$

We could also have computed $P(Y = 0)$ by using the value for $P(Y = 1)$ and Equation 2.2. These probabilities tell us the proportion of times I say heads and tails. They are different from the proportion of times that the coin lands heads or tails ($P(X = 1) = P(X = 0) = 0.5$). This discrepancy is due to the uncertainty in my communication of the results – in the context of this chapter, I am effectively a source of noise or errors. A further example of conditional probabilities and marginalisation is provided in Comment 2.1.

Comment 2.1 – Conditional probabilities and marginalisation – an example: Let's assume that we have a fair coin and two dice (one of which is a little unusual). We will generate a coin toss (X) and a dice roll (Y) using the following procedure. Firstly, toss the coin. If it gives heads, roll die 1. If it gives tails, roll die 2. Die 1 and die 2 are different, with probabilities defined in the following table:

	1	2	3	4	5	6	
Die 1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$= P(y\|X = H)$
Die 2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{12}$	$= P(y\|X = T)$

So, the probability of rolling say a 3 is 1/6 with die 1 and 1/4 with die 2. As we roll die 1 if our coin showed heads and dice 2 if tails, we have the following conditional distributions:

$$P(y|X = H),\ P(y|X = T),$$

i.e. the distribution over Y depends on the outcome of X. The joint distribution is given as (Equation 2.6)

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

We can use this to compute the probability of rolling a 3 *and* a head:

$$P(Y = 3, x = H) = P(Y = 3|X = H)P(X = H) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}.$$

Alternatively, a 3 *and* a tail:

$$P(Y = 3, X = T) = P(Y = 3|X = H)P(X = T) = \frac{1}{4} \times \frac{1}{2} = \frac{1}{8}.$$

Perhaps more interestingly, we can compute the marginal distribution for Y. From our definition (Equation 2.9)

$$P(y) = \sum_x P(y, x) = \sum_x P(y|x)P(x).$$

Therefore, the probability of rolling a 3 is

$$\begin{aligned} P(Y = 3) &= \sum_x P(Y = 3|x)P(x) \\ &= P(Y = 3|X = H)P(X = H) + P(Y = 3|X = T)P(X = T) \\ &= \frac{1}{6} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = \frac{5}{24}. \end{aligned}$$

2.2.7 Aside – Bayes' rule

Although we won't need it in this chapter, it is worth introducing Bayes'[1] rule, as it will feature heavily from Chapter 3 onwards. The left hand sides of Equations 2.6

[1] Named after the Reverend Thomas Bayes, a British mathematician and Presbyterian minister, who first proposed this reversing of conditional probabilities.

and 2.7 are identical so we can also equate the right hand sides:

$$P(Y = y|X = x)P(X = x) = P(X = x|Y = y)P(Y = y).$$

Rearranging, we can get an expression for the probability of X conditioned on a particular value of Y ($P(X = x|Y = y)$) that depends on the probability of Y conditioned on a particular value of X ($P(Y = y|X = x)$), which is known as Bayes' rule:

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}. \tag{2.11}$$

In our example, this is the probability that the coin landed in a particular way given (or conditioned on) what I said. This is likely to be of interest to you if you want to make predictions about how the coin actually landed. Substituting our numerical values, we can work out $P(X = 1|Y = 1)$,

$$P(X = 1|Y = 1) = \frac{P(Y = 1|X = 1)P(X = 1)}{P(Y = 1)} = \frac{0.8 \times 0.5}{0.4} = 1$$

from which we can also deduce that $P(X = 0|Y = 1) = 0$ (Equation 2.2 again). Similarly, we can compute $P(X = 0|Y = 0)$,

$$P(X = 0|Y = 0) = \frac{P(Y = 0|X = 0)P(X = 0)}{P(Y = 0)} = \frac{1 \times 0.5}{0.6} = 0.83,$$

from which we can deduce that $P(X = 1|Y = 0) = 0.17$.

The first two values give the probabilities of the true coin toss if I say heads (i.e. $Y = 1$) and the second two the true probabilities if I say tails ($Y = 0$). $P(X = 1|Y = 1) = 1$ tells us that my saying heads must mean that heads was the true outcome of the coin toss. $P(X = 0|Y = 0) = 0.83$ tells us that, if tails is heard, it is more likely that the coin was tails (probability 0.83) than heads (probability 0.17). Reversing the conditioning in this way is very useful when building models and is something that we shall return to in Chapter 3 and beyond.

2.2.8 Expectations

When dealing with random variables, it is useful to summarise a distribution with a value or values that encapsulate its characteristics. An obvious example is the mean value – the average value that we expect the random variable to take. The mean is an example of an **expectation**. An expectation tells us what value we would expect some function $f(X)$ of a random variable X to take and is defined (for discrete random variables) as

$$\mathbf{E}_{P(x)}\{f(X)\} = \sum_x f(x)P(x). \tag{2.12}$$

For example, if we're interested in the expected value of X (the mean), $f(X) = X$, and the expression becomes

$$\mathbf{E}_{P(x)}\{X\} = \sum_x xP(x).$$

For a fair die ($P(x) = 1/6$), the expected value of X would be

$$\mathbf{E}_{P(x)}\{X\} = \sum_x x\frac{1}{6} = \frac{1}{6} + \frac{2}{6} + \ldots + \frac{6}{6} = \frac{21}{6} = 3.5.$$

Notice from this example that the expected value doesn't have to be one of the values that the random variable can take (we can never roll 3.5).

Expected values of other functions are computed in exactly the same manner. For example, the expected value of $f(X) = X^2$ is

$$\mathbf{E}_{P(x)}\left\{X^2\right\} = \sum_x x^2\frac{1}{6} = \frac{1}{6} + \frac{4}{6} + \ldots + \frac{36}{6} = \frac{91}{6}.$$

It is important to realise that the expected value of a function of X is not in general the function evaluated at the expected value of X. Mathematically, $\mathbf{E}_{P(x)}\{f(X)\}$ does not necessarily equal $f\left(\mathbf{E}_{P(x)}\{X\}\right)$. As an example, we've just computed $\mathbf{E}_{P(x)}\left\{X^2\right\} = 91/6$, which is not equal to $\left(\mathbf{E}_{P(x)}\{X\}\right)^2 = (21/6)^2$. One situation where the two are equal is when the function is just a constant multiplied by X. In this case, doing a little algebra allows us to show that the two are equivalent:

$$\begin{aligned} f(X) &= aX \\ \mathbf{E}_{P(x)}\{f(X)\} &= \sum_x axP(x) \\ &= a\sum_x xP(x) \\ &= a\mathbf{E}_{P(x)}\{X\} \\ &= f\left(\mathbf{E}_{P(x)}\{X\}\right). \end{aligned}$$

Another important case is when the function is simply a constant. In this case, the expectation disappears due to the fact that the distribution has to sum to 1 over all possible outcomes:

$$\begin{aligned} f(X) &= a \\ \mathbf{E}_{P(x)}\{f(X)\} &= \sum_x aP(x) \\ &= a\sum_x P(x) \\ &= a. \end{aligned}$$

A final special case that will prove useful is that the expectation of a sum of different functions is equal to a sum of the individual expectations:

$$\begin{aligned} \mathbf{E}_{P(x)}\{f(X) + g(X)\} &= \sum_x (f(x) + g(x))P(x) \\ &= \sum_x f(x)P(x) + \sum_x g(x)P(x) \\ &= \mathbf{E}_{P(x)}\{f(X)\} + \mathbf{E}_{P(x)}\{g(X)\}. \end{aligned}$$

The two most common expectations that we will come across are the mean ($\mathbf{E}_{P(x)}\{X\}$ as defined above) and the **variance**. Variance is a measure of how variable the random variable is and is defined as the expected squared deviation from the mean:

$$\mathsf{var}\{X\} = \mathbf{E}_{P(x)}\left\{(X - \mathbf{E}_{P(x)}\{X\})^2\right\}. \tag{2.13}$$

Multiplying out the bracket gives us the following convenient expression for the variance of a random variable:

$$\begin{aligned}\mathsf{var}\{X\} &= \mathbf{E}_{P(x)}\left\{(X - \mathbf{E}_{P(x)}\{X\})^2\right\} \\ &= \mathbf{E}_{P(x)}\left\{X^2 - 2X\mathbf{E}_{P(x)}\{X\} + \mathbf{E}_{P(x)}\{X\}^2\right\} \\ &= \mathbf{E}_{P(x)}\left\{X^2\right\} - 2\mathbf{E}_{P(x)}\{X\}\,\mathbf{E}_{P(x)}\{X\} + \mathbf{E}_{P(x)}\{X\}^2.\end{aligned}$$

To get from the second to third lines, we have used the fact that

$$\mathbf{E}_{P(x)}\left\{\mathbf{E}_{P(x)}\{f(X)\}\right\} = \mathbf{E}_{P(x)}\{f(X)\}.$$

The result of $\mathbf{E}_{P(x)}\{f(X)\}$ is a constant (all X terms are removed by the expectation). The outer expectation is the expected value of a constant, which we have already shown is equal to the constant. Collecting together the $\mathbf{E}_{P(x)}\{X\}^2$ terms gives

$$\mathsf{var}\{X\} = \mathbf{E}_{P(x)}\left\{X^2\right\} - \mathbf{E}_{P(x)}\{X\}^2. \tag{2.14}$$

Random variables with high variance would, on average, take values further away from their mean than random variables with low variance.

Comment 2.2 – Vector random variables: It will often be necessary to define probability distributions over vectors. This is nothing more than a shorthand way of defining large joint distributions. For example, the values that could be taken on by random variables $X_1, X_2, \ldots, X_N$ can be expressed as the vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^\mathsf{T}$. Using this shorthand:

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_N) = P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N).$$

Even though $\mathbf{x}$ is a vector, $p(\mathbf{x})$ is a scalar quantity, just as $P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$ is.

Expectations are computed for vector random variables (see Comment 2.2) in exactly the same way. For a random variable X that can take vector values $\mathbf{x}$, expectations are defined as

$$\mathbf{E}_{P(\mathbf{x})}\{f(\mathbf{x})\} = \sum_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x})$$

where the sum is over all possible values of the vector $\mathbf{x}$. Therefore, the mean vector is defined as

$$\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\} = \sum_{\mathbf{x}} \mathbf{x}P(\mathbf{x}).$$

When dealing with vectors, the concept of variance is generalised to a **covariance** matrix. This is defined as

$$\mathsf{cov}\{\mathbf{x}\} = \mathbf{E}_{P(\mathbf{x})}\left\{\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)^\mathsf{T}\right\} \tag{2.15}$$

If $\mathbf{x}$ is a vector of length D, then $\mathsf{cov}\{\mathbf{x}\}$ is a $D \times D$ matrix. The diagonal elements correspond to the variance of the individual elements of $\mathbf{x}$ whilst the off-diagonal elements tell us to what extent different elements of $\mathbf{x}$ co-vary, that is, how dependent they are on one another. A high positive value between, say, elements x_d and x_e, suggests that if x_d increases, so does x_e. A high negative value suggests that they are related but move in opposite directions (x_d increases whilst x_e decreases) and a value of (or close to) zero suggests that there is no relationship between them (they are independent). We give some examples of covariance matrices and the associated densities in Section 2.5.4. Just as for variance, the covariance expression can be manipulated into a more convenient form as follows:

$$\begin{aligned}\mathsf{cov}\{\mathbf{x}\} &= \mathbf{E}_{P(\mathbf{x})}\left\{\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)^{\mathsf{T}}\right\} \\ &= \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}} - 2\mathbf{x}\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^{\mathsf{T}} + \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\,\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^{\mathsf{T}}\right\}.\end{aligned}$$

Rearranging this expression results in

$$\mathsf{cov}\{\mathbf{x}\} = \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\,\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^{\mathsf{T}}. \tag{2.16}$$

2.3 POPULAR DISCRETE DISTRIBUTIONS

In all of our examples thus far, we have worked with random variables for which we can list the probabilities of each possible outcome. This is useful for explanative purposes but rapidly becomes infeasible as the number of possible outcomes increases. In reality, we will often work with well-known families of distributions. Each family is suitable for particular types of events and in general these distributions have parameters that can be tuned to change their characteristics. In this section we will describe some common discrete distributions that you are likely to come across in Machine Learning.

2.3.1 Bernoulli distribution

We have already come across the Bernoulli distribution several times without realising it. It is used for events like a coin toss that have two possible outcomes. For a random variable X that can take two values, 0 or 1 (a binary random variable), where the probability that it takes the value 1 is defined as q, the Bernoulli distribution is

$$P(X = x) = q^x(1-q)^{1-x}. \tag{2.17}$$

The Bernoulli distribution is also a special case of the **binomial** distribution (see below) when $N = 1$.

2.3.2 Binomial distribution

The binomial distribution extends the Bernoulli distribution to define the probability of observing a certain number of heads in a total of N tosses. More generally, we might think of events that have two outcomes (success or failure). If we have N such

events, the binomial random variable Y can take values from 0 (no successes) to N (N successes). The probability of observing a particular number of successes is given by

$$P(Y = y) = P(y) = \left(\begin{array}{c} N \\ y \end{array} \right) q^y (1-q)^{N-y}. \tag{2.18}$$

The second part of this expression looks very similar to the Bernoulli expression we have already seen. In fact, if we define the N binary outcomes as $x_1, \ldots, x_N$, the second part of the binomial expression is the product of the N binomial probabilities:

$$\begin{aligned} \prod_{n=1}^{N} q^{x_n} (1-q)^{1-x_n} &= q^{\sum_n x_n} (1-q)^{N-\sum_n x_n} \\ &= q^y (1-q)^{N-y}, \end{aligned}$$

where $y = \sum_n x_n$: the number of successes (a success corresponds to $x_n = 1$). The first part of the binomial expression is required because there is potentially more than one set of $x_1, x_2, \ldots, x_N$ that corresponds to, say, $y = 3$. $q^y(1-q)^{N-y}$ gives us the probability of just one of these sets. Summing over all possible sets is equivalent to multiplying by the number of such sets, given by the combinations function, $\left(\begin{array}{c} N \\ y \end{array} \right)$ (read as N choose y – see Comment 2.3 for details). Figure 2.3 shows an example of the distribution function when $N = 50$ and $q = 0.7$.

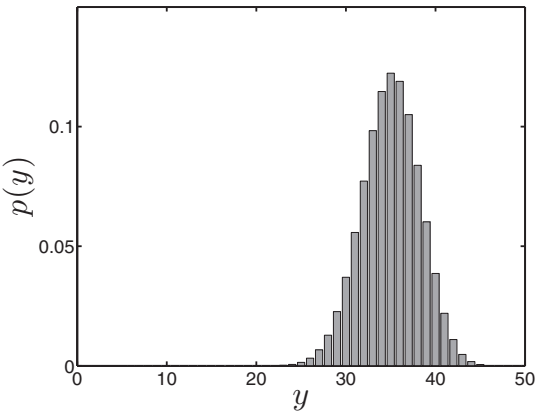


FIGURE 2.3 An example of the probability distribution function for a binomial random variable when $N = 50$ and $q = 0.7$ (see Equation 2.18).

2.3.3 Multinomial distribution

Our previous two examples have been distributions over scalar random variables – we will now look at a distribution that assigns probabilities to vectors of discrete values. The fundamental ideas are exactly the same – the distribution assigns a

probability to every possible vector and the sum of these probabilities must equal 1. As a motivation for vector random variables, imagine you were building a machine that would produce random documents of N words and you wanted to define a distribution over these documents. This isn't as foolish as it might sound – Machine Learning techniques are often used to analyse text data by defining distributions over documents in just this manner. One way of representing a document would be with a vector of word counts. Assuming J possible words in our vocabulary, the vector would be of length J and the jth element would hold the number of times the jth word appears in the document. The **multinomial** distribution allows us to define a distribution over such vectors. Let Y be a random variable that represents a document. An instance of this random variable is a vector of word counts $\mathbf{y} = [y_1, \ldots . y_J]^\mathsf{T}$; the multinomial distribution defines the probability of $\mathbf{y}$ as

$$P(Y = \mathbf{y}) = P(\mathbf{y}) = \frac{N!}{\prod_j y_j!} \prod_j q_j^{y_j} \tag{2.19}$$

where q_j are the parameters of the multinomial distribution and represent the probabilities of the individual words ($\sum_j q_j = 1$).

Comment 2.3 – Combinations: N choose y, written as

$$\begin{pmatrix} N \\ y \end{pmatrix}$$

is mathematical shorthand for the number of ways in which y distinct objects can be chosen from a set of N objects. For example, $\begin{pmatrix} 4 \\ 1 \end{pmatrix}$ would be 4 – there are 4 ways I can choose one object from four objects – object 1 on its own, object 2 on its own, object 3 on its own or object 4 on its own. $\begin{pmatrix} 4 \\ 2 \end{pmatrix}$ is 6 – the possible choices are 1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4 or 3 and 4. In general,

$$\begin{pmatrix} N \\ y \end{pmatrix} = \frac{N!}{y!(N-y)!}$$

where $N!$ (read N factorial) is

$$\prod_{i=1}^{N} i = N \times (N-1) \times (N-2) \times \ldots \times 1.$$

2.4 CONTINUOUS RANDOM VARIABLES – DENSITY FUNCTIONS

We saw at the start of this section that we are unable to systematically write down all possible outcomes of a continuous random variable. Unfortunately, this precludes us from assigning probabilities to particular values. To overcome this, we work with the probabilities of the outcome falling within some range or interval. For example, given a continuous random variable X that can take on any value between minus infinity and infinity, it makes sense to try and work out

$$P(x_1 \leq X \leq x_2)$$

but not

$$P(X = x).$$

When working with continuous random variables, we need a continuous analogue to the probability distribution (recall that this, for a discrete random variable, was the set of outcomes (x) and the probabilities of each outcome, expressed as a function of x, $p(x)$). This is provided by a **probability density function** (pdf), also denoted $p(x)$. To compute the probability that X lies in a particular range, we compute the definite integral (see Comment 2.4) of $p(x)$ with respect to x over this range:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)dx.$$

If our random variable may only take values in the range $x_1 \leq X \leq x_2$, it stands to reason that the probability that it lies in this range must be 1. This leads us to the continuous equivalent of Equation 2.2:

$$\int_{x_1}^{x_2} p(x)dx = 1 \quad \text{where } x_1 \leq X \leq x_2. \tag{2.20}$$

Equation 2.1 also has a continuous equivalent,

$$p(x) \geq 0, \tag{2.21}$$

that tells us that a pdf can never be negative. Note that there is no upper bound on the value of the pdf – it is not a probability and so can (and often will) be higher than 1 for a particular value of x.

Comment 2.4 – Definite Integrals: When differentiating a function including a constant term, the term disappears, e.g.

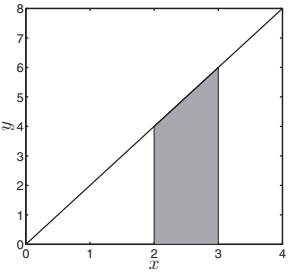
$$\frac{d}{dx}(x^2 + 3) = 2x.$$

Hence, when we are integrating a function, we have to admit the possibility that there might be a constant term

$$\int 2x \; dx = x^2 + C.$$

This is called an indefinite integral, as we don't know the value of C.

Often we will be interested in using integration to compute the area under a curve. For example, here we are interested in computing the area under the curve $y = 2x$ between $x = 2$ and $x = 3$, as shown in the plot on the right. This is calculated as



$$\int_2^3 2x \; dx = [x^2 + C]_2^3$$

where the $[\cdot]_a^b$ means take the value of the object inside the brackets when $x = a$ away from the value when $x = b$. In this case, this suggests

$$(3^2 + C) - (2^2 + C) = 9 - 4 + C - C = 5.$$

This is a definite integral – the constants cancel out and the answer is exact.

Joint and conditional continuous densities Just as with the discrete case, we can define joint probability density functions over several continuous random variables. For example, $p(x, y)$ is the joint density of two random variables X and Y, and $p(\mathbf{w})$, is the density of a vector, $\mathbf{w}$ which could be thought of as the joint density of $p(w_0, w_1, \dots)$ – random variables representing each element in the vector. Although we cannot compute $P(X = x, Y = y)$, we can compute

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{x=x_1}^{x_2} \int_{y=y_1}^{y_2} p(x, y) \; dx \; dy.$$

The same applies for conditional distributions, although the conditioning is done on an exact value (as this event is assumed to have happened). For example, we would compute

$$P(x_1 \leq X \leq x_2 | Y = y) = \int_{x=x_1}^{x_2} p(x|Y = y) \; dx.$$

Often we will use the shorthand $p(x|y)$ to describe the density function of X given that $Y = y$.

Marginalisation You may have already guessed that to marginalise over a continuous random variable, we replace the summation from the discrete case with an integral. For example, the pdf $p(y)$ can be computed from $p(y, x)$ as follows:

$$p(y) = \int_{x=x_1}^{x_2} p(y, x)\ dx$$

where $x_1 \leq X \leq x_2$ describes the sample space of X.

Expectations Expectations with respect to continuous random variables are performed by integrating over the range of values that the random variable can take:

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} = \int f(x)p(x)dx. \tag{2.22}$$

All of the expressions derived in Section 2.2.8 are identical in the continuous case.

In many practical scenarios, we will not be able to perform this integral – we may not know the exact form of $p(x)$ or it might simply be impossible to integrate. However, if we can generate samples from $p(x)$, it can be approximated by

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s) \tag{2.23}$$

where x_s is one of the S samples from $p(x)$. This is an example of a **Monte Carlo** approximation to an integral which we will see a lot more of in subsequent chapters.

2.5 POPULAR CONTINUOUS DENSITY FUNCTIONS

Just as for the discrete case, there are several common families of continuous density functions that we will often come across. In this section, we will describe three of them.

2.5.1 The uniform density function

The simplest continuous density function is the uniform density function. The uniform density function, $p(y) = \mathcal{U}(a, b)$, is constant between a and b and zero elsewhere:

$$p(y) = \begin{cases} r & \text{for } a \leq y \leq b \\ 0 & \text{otherwise.} \end{cases} \tag{2.24}$$

An example where $a = 3$ and $b = 8$ can be seen in Figure 2.4. We can compute the value of r for any values of a and b by remembering that the integral of the pdf over the sample space must be equal to 1 by definition. In this case,

$$\begin{aligned} P(a \leq Y \leq b) = 1 &= \int_{y=a}^{b} p(y)\ dy = \int_{y=a}^{b} r\ dy \\ &= [yr]_a^b = rb - ra = r(b-a) \\ r &= \frac{1}{b-a}. \end{aligned}$$

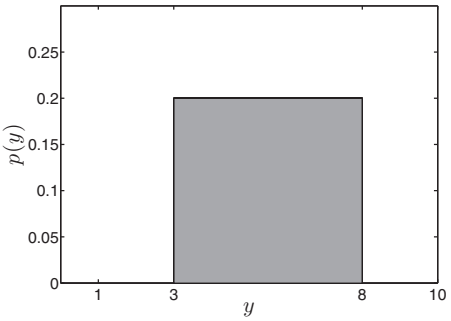


FIGURE 2.4 An example of the uniform pdf.

This is quite intuitive – it is the total probability available – 1 – divided by the length of the interval in which the variables must lie $(b-a)$. We can also easily define multidimensional uniform random variables. For example, if $\mathbf{y} = [y_1, y_2]^\mathsf{T}$,

$$p(\mathbf{y}) = \begin{cases} r & \text{for } a \leq y_1 \leq b \text{ and } c \leq y_2 \leq d \\ 0 & \text{otherwise} \end{cases}$$

and we can compute r in just the same way:

$$\begin{aligned} P(a \leq y_1 \leq b, c \leq y_2 \leq d) = 1 &= \int_{y_1=a}^{b} \int_{y_2=c}^{d} r \; dy_1 \; dy_2 \\ &= \int_{y_1=a}^{b} [ry_2]_c^d \; dy_1 = \int_{y_1=a}^{b} r(d-c) dy_1 \\ &= [r(d-c)y_1]_a^b = r(d-c)(b-a) \\ r &= \frac{1}{(d-c)(b-a)}. \end{aligned}$$

Again, this is intuitive – it is the total probability – 1 – divided by the area of the interval in which the variables must lie $(d-c)(b-a)$.

As an aside, Equation 2.23 shows how we can approximate expectations by taking samples (realisations of the random variable) from the appropriate distribution. We will demonstrate this approach by computing the expected value of y^2 analytically and via sampling. The analytical result is given by

$$\begin{aligned} \mathbf{E}_{p(y)}\left\{y^2\right\} &= \int_{y=a}^{b} y^2 p(y) \; dy = \int_{y=a}^{b} \frac{y^2}{b-a} \; dy \\ &= \left[\frac{y^3}{3(b-a)}\right]_a^b = \frac{b^3 - a^3}{3(b-a)}. \end{aligned}$$

Substituting $a = 0, b = 1$ gives

$$\mathbf{E}_{p(y)}\left\{y^2\right\} = \frac{1}{3}.$$

To compute the sample based approximation, we need to be able to draw samples

from $\mathcal{U}(0,1)$. In MATLAB, the command `rand` generates samples from this distribution. If we generate S samples, y_s, we can approximate the expectation as

$$\mathbf{E}_{p(y)}\left\{y^2\right\} \approx \frac{1}{S}\sum_{s=1}^{S} y_s^2. \tag{2.25}$$

Figure 2.5 shows how this approximation improves as we increase the number of samples from 1 to 10^4. The true value, $\frac{1}{3}$, is shown as the dashed line (MATLAB script: `approx_expected_value.m`). After only 100 samples, the approximation is reasonably good. Approximating expectations with samples will be used extensively in later chapters (see Exercise 2.4).

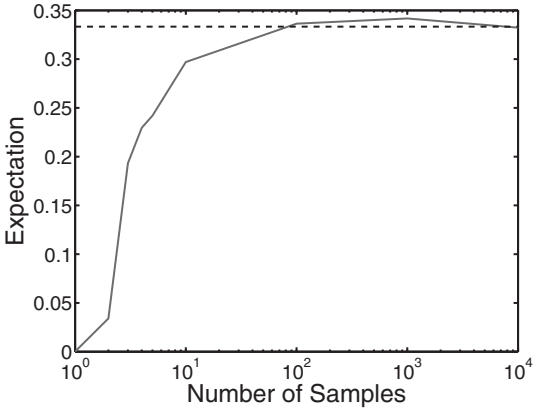


FIGURE 2.5 Effect of increasing the number of samples on the approximation to the expectation given in Equation 2.25 where $p(y) = \mathcal{U}(0,1)$. The dashed line is the true value of 1/3. Note the log scale on the x-axis.

2.5.2 The beta density function

The beta density function can be used for continuous random variables that are restricted to between 0 and 1. The beta density function is defined as

$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}, \tag{2.26}$$

where α and β are parameters that control the shape of the density function, both of which must be positive. $\Gamma(z)$ is known as the gamma function and we will omit a discussion here except to say that it can be computed in MATLAB using the inbuilt function `gamma`. Figure 2.6 shows the beta pdfs corresponding to three different sets of parameters. We will use the beta density function considerably in Chapter 3 and so will leave more discussion until then.

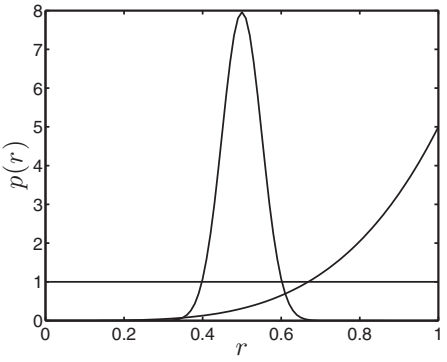


FIGURE 2.6 Examples of beta pdfs with three different pairs of parameters.

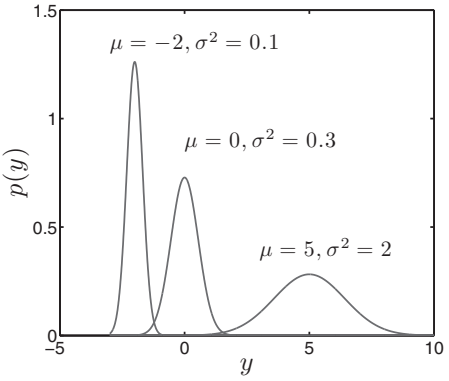


FIGURE 2.7 Three Gaussian pdfs with different means and variances.

2.5.3 The Gaussian density function

Gaussian random variables are used in many continuous applications. One reason is the ease with which the Gaussian pdf can be manipulated in certain, useful situations. The Gaussian distribution is defined over a sample space that includes all real numbers (i.e. all numbers between $-\infty$ and ∞) and has a pdf for a random variable Y defined as

$$p(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \tag{2.27}$$

and is characterised by two variables – the mean (μ) and variance (σ^2). Figure 2.7 shows three Gaussian pdfs with different μ, σ^2 values. The highest value of the pdf is obtained when $y = \mu$ and the density is symmetric about this point. The width of the density is controlled by σ^2 – the higher the value, the wider the density. If we used the leftmost Gaussian in Figure 2.7 to generate instances of a random variable, we

would only expect values from a small range around −2. For the rightmost Gaussian, we would anticipate values from quite a large range around 5. A common shorthand for the Gaussian pdf is $\mathcal{N}(\mu, \sigma^2)$. Therefore, if Y has a Gaussian pdf, we could write

$$p(y|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2),$$

which reads as 'the density function for the random variable Y is normal (Gaussian and normal are used interchangeably) with mean μ and variance σ^2'.

2.5.4 Multivariate Gaussian

The Gaussian distribution can also be generalised to define a density function over continuous vectors. This multivariate Gaussian density for a vector $\mathbf{x} = [x_1, \ldots, x_D]^\mathsf{T}$ is something we will use a great deal in subsequent chapters. The density function is defined as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{2.28}$$

where the mean $\boldsymbol{\mu}$ is now a vector (of the same size as $\mathbf{x}$), the dth element of which tells us the mean value of x_d, and the variance has become a $D \times D$ covariance matrix. A graphical example is perhaps the best way of getting a feel for this density and the effects of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The first example is shown in the top line of Figure 2.8. In this example, the parameters are

$$\boldsymbol{\mu} = [2,\ 1]^\mathsf{T},\ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This is a special case of the multivariate Gaussian where the two variables (say x_1 and x_2) are independent. To show this, we note that $\boldsymbol{\Sigma} = \mathbf{I}$. So,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{I}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{I}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Now, $\mathbf{I}^{-1} = \mathbf{I}$ (see Comment 1.10), allowing us to manipulate this expression to obtain a product over univariate Gaussian pdfs. Starting with the expression above (having swapped the $\mathbf{I}^{-1}$ for $\mathbf{I}$), we can convert the matrix product inside the exponential into a sum over the D different elements (see Exercise 2.5):

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}|\mathbf{I}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{I}(\mathbf{x}-\boldsymbol{\mu})\right\} \\ &= \frac{1}{(2\pi)^{D/2}|\mathbf{I}|^{1/2}} \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}(x_d-\mu_d)^2\right\}. \end{aligned}$$

Comment 2.5 – Matrix determinant: The determinant of a square matrix, denoted $|\mathbf{A}|$ for matrix $\mathbf{A}$, is a useful quantity, especially when dealing with multivariate Gaussians. For large matrices, it is too cumbersome to calculate by hand but it can be done for small matrices. For example, for a 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \; |\mathbf{A}| = ad - bc,$$

but for anything bigger than this it is safest to resort to a computer unless the matrix has a special structure. One special matrix that we will see a lot of is a square matrix that only has diagonal elements (all off-diagonal elements are zero). In this case, the determinant is simply the product of these elements. For example,

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{DD} \end{bmatrix}, \; |\mathbf{A}| = \prod_{d=1}^{D} a_{dd}.$$

It is not easy to gain an intuition into what the determinant represents. Its role in the normalisation constant of the multivariate Gaussian leads us to think of it as related to the volume of the Gaussian unnormalised Gaussian (remember that the normalised volume must be equal to 1) and it may be useful to think of it in this way.

The exponential of a sum is a product of exponentials, allowing us to rewrite the expression as follows:

$$p(\mathbf{x}) \quad = \quad \frac{1}{(2\pi)^{D/2}|\mathbf{I}|^{1/2}} \prod_{d=1}^{D} \exp\left\{-\frac{1}{2}(x_d - \mu_d)^2\right\}.$$

$|\mathbf{I}|$ is the determinant of $\mathbf{I}$, which, from the discussion of diagonal matrices in Comment 2.5, is equal to 1. The other constant term, $(2\pi)^{D/2}$, could be written as $\prod_{d=1}^{D}(2\pi)^{1/2}$ and so our expression can be rewritten as

$$p(\mathbf{x}) = \prod_{d=1}^{D} \frac{1}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2}(x_d - \mu_d)^2\right\}.$$

Each term in the product is a univariate Gaussian (with mean μ_d and variance 1) and therefore, by the definition of independence, the elements of $\mathbf{x}$ are independent. This result doesn't just hold for $\Sigma = \mathbf{I}$, it holds for any covariance matrix that has non-zero elements only in the diagonal positions. These diagonal elements will be the variances of the individual, univariate Gaussians (see Exercises 2.5 and 2.6 for further exercises and practice at this kind of Gaussian manipulation).

The second row in Figure 2.8 gives another example, with parameters:

$$\boldsymbol{\mu} = [2, \; 1]^{\mathsf{T}}, \; \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

In this example, we could not write the pdf as a product of univariate Gaussians,

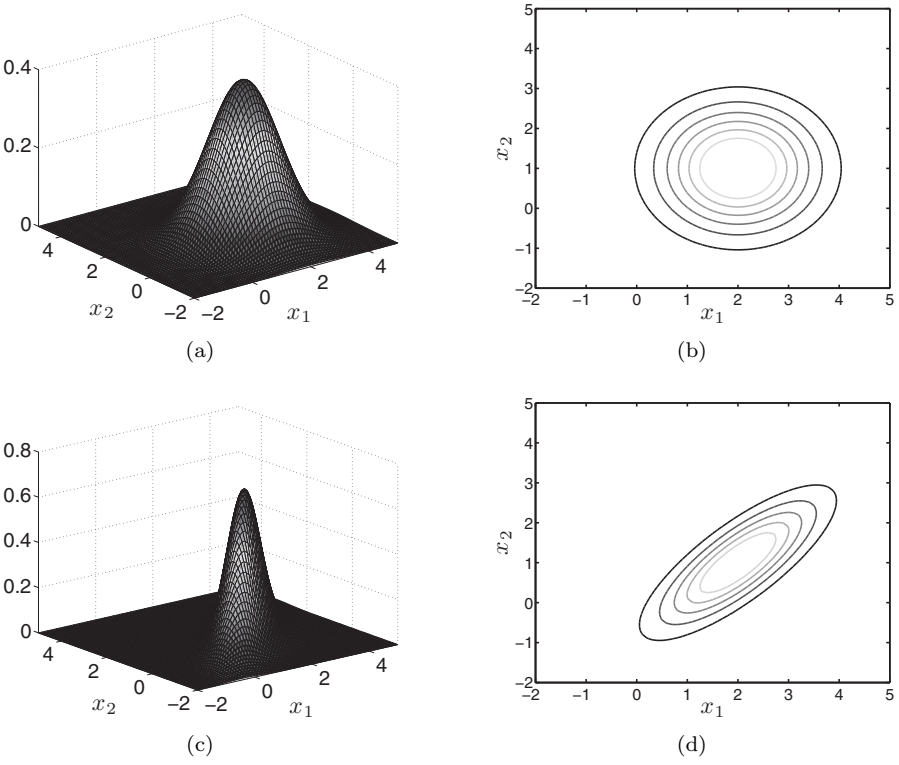


FIGURE 2.8 Example surface (left) and contour (right) plots for two different two-dimensional Gaussian pdfs.

suggesting that the elements of $\mathbf{x}$ are not independent. We can also see the dependence between them in the contour plot (bottom right of Figure 2.8). If x_1 and x_2 are independent, $p(x_2|x_1)$ should not vary with different values of x_1. Imagine that $x_1 = 3$. It looks, from Figure 2.8, that when $x_1 = 3$, values for x_2 are grouped around 2. If $x_1 = 1$, the values are grouped around 0. Clearly we expect different values of x_2 in both cases and, intuitively, x_1 and x_2 are dependent (MATLAB script: `gauss_surf.m`). Experiment with the values in the covariance matrix to see the effect this has on the surface and contour plots.

A nice feature of the multivariate Gaussian is that the conditional density function $p(x_2|x_1)$ is another Gaussian for which we can easily obtain the mean and variance.

2.6 SUMMARY

This completes our brief introduction to random variables and probability. Although we have only skimmed the surface of an enormous subject, the material presented

in the previous few sections is sufficient for us to extend our model to explicitly measure the discrepancy between predictions and measurements. In the remainder of this chapter, we will add a random variable to our model that will model the error between the linear model and our data. Assuming that the random variable follows a Gaussian density, we will end up with exactly the equation for $\widehat{\mathbf{w}}$ (the optimum parameter value) as in Chapter 1. However, the inclusion of the noise term allows us to obtain degrees of confidence in both our parameter values and predictions.

2.7 THINKING GENERATIVELY...CONTINUED

We now have a sufficient grounding in random variables to be able to handle the errors in our linear model (as shown in Figure 2.1). In Section 2.1.1 we began thinking about how we could generate data that looks like the data that we have observed. In particular, we considered generating the nth winning time from a function of the form $\mathbf{w}^{\mathsf{T}}\mathbf{x}_n$ and then adding a random quantity that we shall call ϵ_n – a random variable.

Our model now takes the following form:

$$t_n = \mathbf{w}^{\mathsf{T}}\mathbf{x}_n + \epsilon_n. \tag{2.29}$$

To complete the definition of this model, we need to decide on a distribution for ϵ_n. Firstly, it should be clear that the difference between the model and the actual winning times is a continuous quantity. Therefore, ϵ_n is a continuous random variable. We also do not just have one random variable, but one for each observed Olympic year. It seems reasonable to assume that these values are independent:

$$p(\epsilon_1, \ldots, \epsilon_N) = \prod_{n=1}^{N} p(\epsilon_n).$$

The final assumption is the form of $p(\epsilon_n)$. We will assume that this is a Gaussian (or normal) distribution with zero mean and variance σ^2. We will not make much effort to justify this assumption here except to say that this allows ϵ_n to be both positive and negative (allows data to lie both above and below the line $\mathbf{w}^{\mathsf{T}}\mathbf{x}$) and has interesting modelling properties that link it to the squared loss that we used in Chapter 1. As for the choice of loss functions discussed in Section 1.1.3, in a real modelling situation one should be much more careful to properly justify this choice.

Using a normal density for ϵ, i.e. $p(\epsilon) = \mathcal{N}(\mu, \sigma^2)$ (see Section 2.5.3), with a mean (μ) of zero and a variance of $\sigma^2 = 0.05$ (don't worry about the particular value here for now), we obtain a much more realistic looking dataset, shown in Figure 2.9 (MATLAB script: `genolymp.m`).

Our model now consists of two components:

1. A **deterministic** component ($\mathbf{w}^{\mathsf{T}}\mathbf{x}_n$), sometimes referred to as a *trend* or *drift*.
2. A random component (ϵ_n), sometimes referred to as *noise*.

We have already pointed out that we are not restricted to noise from a Gaussian distribution. We are also not restricted to *additive* noise. For some applications, a multiplicative term might be more appropriate (in which case, $t = f(\mathbf{x}; \mathbf{w})\epsilon$). For example, degradation of image pixels is often modelled with multiplicative noise.

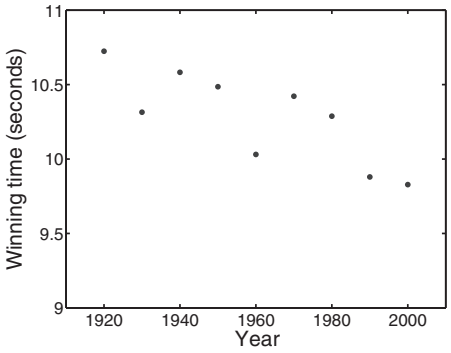


FIGURE 2.9 Dataset generated from a linear model with Gaussian errors.

However, as we shall see in the following sections, choosing additive Gaussian noise allows us to obtain exact expressions for the optimal parameter value $\widehat{\mathbf{w}}$.

2.8 LIKELIHOOD

Our model is of the following form:

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n, \ \epsilon_n \sim \mathcal{N}(0, \sigma^2).$$

As in Chapter 1, we need to find the optimal value of $\mathbf{w}$, $\widehat{\mathbf{w}}$. We also have an additional parameter σ^2 that needs to be set. In Chapter 1 we found the value of $\mathbf{w}$ that minimised the loss. The loss measured the difference between the observed values of t and those predicted by the model. The effect of adding a random variable to the model is that the output of the model, t, is now itself a random variable. In other words, there is no single value of t_n for a particular $\mathbf{x}_n$. As such, we cannot use the loss as a means of optimising $\mathbf{w}$ and σ^2.

Adding a constant $(\mathbf{w}^\mathsf{T}\mathbf{x}_n)$ to a Gaussian random variable is equivalent to another Gaussian random variable with the mean shifted by the same constant:

$$\begin{aligned} y &= a + z \\ p(z) &= \mathcal{N}(m, s) \\ p(y) &= \mathcal{N}(m + a, s) \end{aligned}$$

Therefore, the random variable t_n has the density function

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2).$$

Note the conditioning on the left hand side – the density of t_n depends on particular values of $\mathbf{x}_n$ and $\mathbf{w}$ (they determine the mean) and σ^2 (the variance).

To see how we can use this to find optimal values of $\mathbf{w}$ and σ^2, consider one of the years from our dataset – 1980. Based on the model (w_0, w_1) found in the previous chapter and assuming again that $\sigma^2 = 0.05$, we can plot $p(t_n|x_n = 1980, \mathbf{w}, \sigma^2)$ as a function of t_n, shown Figure 2.10. The solid line shows

$$p(t_n|\mathbf{x}_n = [1, \ 1980]^\mathsf{T}, \mathbf{w} = [36.416, -0.0133]^\mathsf{T}, \sigma^2 = 0.05),$$

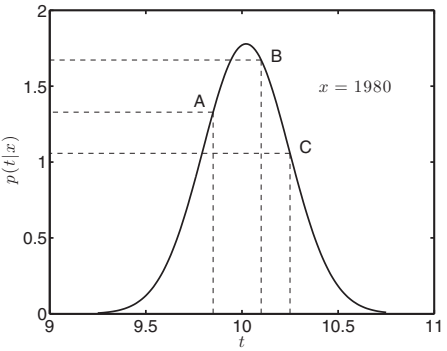


FIGURE 2.10 Likelihood function for the year 1980.

which is a Gaussian density with mean $\mu = 36.416 - 0.0133 \times 1980 = 10.02$ and variance $\sigma^2 = 0.05$. Recall that, for a continuous random variable, t, $p(t)$ cannot be interpreted as a probability. The height of the curve at a particular value of t can be interpreted as how *likely* it is that we would observe that particular t for $x = 1980$. The most *likely* winning time in 1980 would be 10.02 seconds (for a Gaussian, the most likely (highest) point corresponds to the mean). Also shown on the plot, are three example times – A, B and C. Of these, B is the most likely and C the least likely.

The actual winning time in the 1980 Olympics is C (10.25 seconds). The density $p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2)$ evaluated at $t_n = 10.25$ is an important quantity, known as the **likelihood** of the nth data point. We cannot change $t_n = 10.25$ (this is our data) but we can change $\mathbf{w}$ and σ^2 to try and move the density so as to make it as high as possible at $t = 10.25$. The idea of finding parameters that maximise the likelihood in this way is a key concept in Machine Learning.

2.8.1 Dataset likelihood

In general, we are not interested in the likelihood of a single data point but that of all of the data. If we have N data points, we are interested in the joint conditional density:

$$p(t_1, \ldots, t_N|\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma^2).$$

This is a joint density over all of the responses in our dataset (see Section 2.2.5). We will write this compactly (using vector notation and $\mathbf{X}$ as defined in Chapter 1) as $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$. Evaluating this density at the observed data points gives a single likelihood value for the whole dataset, which we can optimise by varying $\mathbf{w}$ and σ^2.

The assumption that the noise at each data point is independent ($p(\epsilon_1, \ldots, \epsilon_N) = \prod_n p(\epsilon_n)$) enables us to factorise this density into something more manageable. In particular, this joint conditional density can be factorised into N separate terms, one for each data object:

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2). \tag{2.30}$$

Note that we haven't gone as far as saying that the t_n values are themselves completely independent. This is not the case – the t_n values are, on average, decreasing over time, suggesting a clear statistical dependence between them. If they were completely independent, it would not be worthwhile actually trying to model the data at all. In fact, they are **conditionally independent** – given a value for $\mathbf{w}$ (the deterministic part of the model), the t_n are independent; without them they are not. If this sounds a bit strange, think of it in the following way: Imagine that we had values for all of the Olympic years and winning times except one of the ones in the middle – say 1960. For simplicity, we shall use $\mathbf{X}, \mathbf{t}$ to denote all Olympic years and winning times excluding 1960. If we want to use $\mathbf{X}$ and $\mathbf{t}$ to learn something about t_{1960}, we are interested in the conditional distribution

$$p(t_{1960}|\mathbf{x}_{1960}, \mathbf{X}, \mathbf{t}).$$

From the definition of conditional distributions, this is given by

$$p(t_{1960}|\mathbf{x}_{1960}, \mathbf{X}, \mathbf{t}) = \frac{p(t_{1960}, \mathbf{t}|\mathbf{x}_{1960}, \mathbf{X})}{p(\mathbf{t}|\mathbf{X})}.$$

Assuming that the elements of $\mathbf{t}$ are independent results in t_{1960} only depending on $\mathbf{x}_{1960}$:

$$p(t_{1960}|\mathbf{x}_{1960}, \mathbf{X}, \mathbf{t}) = \frac{p(t_{1960}|\mathbf{x}_{1960}) \prod_n p(t_n|\mathbf{x}_n)}{\prod_n p(t_n|\mathbf{x}_n)} = p(t_{1960}|\mathbf{x}_{1960}).$$

However, for our model to be any use, t_{1960} must, in some sense, be dependent on the other data. This dependence is encapsulated in the parameter $\mathbf{w}$. The deterministic part of our model captures this dependence. If we know $\mathbf{w}$, all that remains is the errors between the observed data and $\mathbf{w}^\mathsf{T}\mathbf{x}_n$. These errors are assumed to be independent. Hence, conditioned on $\mathbf{w}$, the observations are independent. Without a model (and therefore a $\mathbf{w}$), the observations are not independent.

We will now show how we can find the values of $\mathbf{w}$ and σ^2 that maximise the likelihood.

2.8.2 Maximum likelihood

Equation 2.30 gives us a single value that tells us how likely our dataset is, given the current model (by model, we mean choice of $\mathbf{w}$ and σ^2). As our dataset is fixed, varying the model will result in different likelihood values. A sensible choice of model would be that which maximised the likelihood. In other words, we will select the model parameters that will make our observations most likely.

For analytical reasons, we will maximise the **natural logarithm** of the likelihood (we will follow the Machine Learning convention of using $\log(y)$ to denote the natural logarithm of y, often denoted elsewhere as $\ln(y)$). We can do this because the estimated arguments $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$ that maximise the log-likelihood will also maximise the likelihood.

Substituting the expression for the Gaussian density function (Equation 2.27) and separating the various terms gives us an expression that will be easier to deal

with:

$$\begin{aligned}\log L &= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2\right\}\right)\\ &= \sum_{n=1}^{N}\left(-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(t_n - f(\mathbf{x}_n,\mathbf{w}))^2\right)\\ &= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2.\end{aligned}$$

Substituting our particular deterministic component $f(\mathbf{x}_n;\mathbf{w}) = \mathbf{w}^\mathsf{T}\mathbf{x}_n$ gives us the log-likelihood expression that we will work with:

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2. \tag{2.31}$$

As for the least squares solution derived in Chapter 1, we can find the optimal parameters by taking derivatives, equating them to zero and solving for turning points, in a manner similar to that in Section 1.1.4. For $\mathbf{w}$ (noting that $\mathbf{w}^\mathsf{T}\mathbf{x}_n = \mathbf{x}_n^\mathsf{T}\mathbf{w}$),

$$\begin{aligned}\frac{\partial \log L}{\partial \mathbf{w}} &= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n(t_n - \mathbf{x}_n^\mathsf{T}\mathbf{w})\\ &= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n t_n - \mathbf{x}_n\mathbf{x}_n^\mathsf{T}\mathbf{w} = \mathbf{0}.\end{aligned}$$

Note that $\frac{\partial \log L}{\partial \mathbf{w}}$ is a vector and so we equate it to $\mathbf{0}$, a vector of zeros of the same size. Recall the shorthand matrix/vector forms we used in Chapter 1:

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\mathsf{T}\\ \mathbf{x}_2^\mathsf{T}\\ \vdots\\ \mathbf{x}_N^\mathsf{T}\end{bmatrix} = \begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_N\end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix}t_1\\ t_2\\ \vdots\\ t_N\end{bmatrix}.$$

In this notation, $\sum_{n=1}^{N}\mathbf{x}_n t_n$ can be written as $\mathbf{X}^\mathsf{T}\mathbf{t}$ and similarly $\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\mathbf{w}$ as $\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}$ (see Exercise 1.5). This allows us to write the derivative in the more convenient vector/matrix form:

$$\frac{\partial \log L}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) = \mathbf{0}. \tag{2.32}$$

Solving this expression for $\mathbf{w}$ will lead to an expression for the optimal value:

$$\begin{aligned}\frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) &= 0\\ \mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} &= 0\\ \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} &= \mathbf{X}^\mathsf{T}\mathbf{t}\\ \mathbf{w} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}.\end{aligned}$$

This is the **maximum likelihood** solution for $\mathbf{w}$:

$$\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}. \tag{2.33}$$

Remarkably, this solution is *exactly* that which we have already derived for the least squares case in Chapter 1 (Equation 1.16). Minimising the squared loss is equivalent to the maximum likelihood solution if the noise is assumed to be Gaussian. Also, the noise variance, σ^2, does not appear in this expression at all – it scales the likelihood but doesn't affect the value of $\widehat{\mathbf{w}}$ corresponding to its maximum.

To obtain an expression for σ^2 (assuming $\mathbf{w} = \widehat{\mathbf{w}}$), we can follow the same procedure. Taking partial derivatives and equating to zero results in

$$\frac{\partial \log L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(t_n - \mathbf{x}^\mathsf{T}\widehat{\mathbf{w}})^2 = 0. \tag{2.34}$$

Rearranging gives $\widehat{\sigma^2}$, the maximum likelihood estimate for σ^2:

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^\mathsf{T}\widehat{\mathbf{w}})^2. \tag{2.35}$$

This expression makes perfect sense – the variance is simply the average squared error. We would prefer this in matrix notation so, using the fact that $\sum_{n=1}^{N}(t_n - \mathbf{x}^\mathsf{T}\widehat{\mathbf{w}})^2$ is equivalent to $(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$,

$$\begin{aligned}\sigma^2 &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}}) \\ &= \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - 2\mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}} + \widehat{\mathbf{w}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}}).\end{aligned} \tag{2.36}$$

This can be further simplified by substituting $\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$ (note that $\widehat{\mathbf{w}}^\mathsf{T} = \mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ because $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ is **symmetric** and is therefore equal to its own transpose):

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - 2\mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} + \mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}) \\ &= \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - 2\mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} + \mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}) \\ &= \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t})\end{aligned}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}}). \tag{2.37}$$

Using the Olympic 100 m data, our optimal parameter values (for a first-order (linear) polynomial) are

$$\widehat{\mathbf{w}} = [36.4165, \ -0.0133]^\mathsf{T}, \ \widehat{\sigma^2} = 0.0503.$$

$\widehat{\mathbf{w}}$ is the same as the least squares solution provided in the previous chapter (they are both computed using the same expression). $\widehat{\sigma^2}$ tells us the variance of the Gaussian noise that we have assumed is used to corrupt our data. Later in this chapter we will see that modelling the noise in this way provides several benefits over loss minimisation. Before we do, we shall first look at some of the characteristics of the solution.

2.8.3 Characteristics of the maximum likelihood solution

In Chapter 1, we used the second derivatives of the loss function to ensure that we had found a minimum. We will now do a similar thing with the second derivatives of the likelihood to ensure that we have found a maximum. Our derivatives are now with respect to a vector, and to examine the second derivatives, we construct the **Hessian matrix** (see Comment 2.6). Each entry in this matrix is the second derivative with respect to a pair of elements of $\mathbf{w}$. To be sure that we have found a maximum, we must show that the Hessian matrix is *negative definite* (see Comment 2.7).

Comment 2.6 – Hessian matrix: A Hessian matrix is square matrix containing all of the second-order partial derivatives of a function. For example, the Hessian matrix for a function $f(\mathbf{x};\mathbf{w})$ with parameters $\mathbf{w} = [w_1, \ldots, w_K]^T$ would be

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_K} \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_K \partial w_1} & \frac{\partial^2 f}{\partial w_K \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_K^2} \end{bmatrix}.$$

We can use the Hessian to tell us something about turning points in $f(\mathbf{x};\mathbf{w})$. For example, if the Hessian is *negative definite* (see Comment 2.7) at some turning point $\widehat{\mathbf{w}}$, then we know that that turning point corresponds to a maximum.

The Hessian matrix of second-order partial derivatives can be computed by differentiating Equation 2.32 with respect to $\mathbf{w}^\mathsf{T}$:

$$\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}. \tag{2.38}$$

If we substitute $\mathbf{x}_n = [1,\ x_n]^\mathsf{T}$, the diagonal elements of this matrix are equivalent (they differ by multiplication by a constant) to the second derivatives obtained in Equation 1.9 (see Exercise 2.7).

Comment 2.7 – Negative definite matrices: A real-valued matrix $\mathbf{H}$ is negative definite if

$$\mathbf{x}^\mathsf{T}\mathbf{H}\mathbf{x} < 0$$

for all vectors of real values $\mathbf{x}$.

To be sure this is a maximum, we need to determine whether or not this matrix is negative definite. We can do this by showing that

$$-\frac{1}{\sigma^2}\mathbf{z}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{z} < 0$$

for any vector $\mathbf{z}$ or equivalently (because σ^2 must be positive) that

$$\mathbf{z}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{z} > 0$$

for any vector $\mathbf{z}$. At this stage, it is probably worth showing how this can be done. We will assume that each $\mathbf{x}_n$ is two dimensional so that we can explicitly multiply

out the various terms. To be more general, we will define $\mathbf{X}$ slightly differently from before as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_N^\mathsf{T} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix}.$$

Thus, $\mathbf{X}^\mathsf{T}\mathbf{X}$ is

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} \sum_{i=1}^N x_{i1}^2 & \sum_{i=1}^N x_{i1}x_{i2} \\ \sum_{i=1}^N x_{i2}x_{i1} & \sum_{i=1}^N x_{i2}^2 \end{bmatrix}.$$

Pre- and postmultiplying by an arbitrary real vector $\mathbf{z} = [z_1, z_2]^\mathsf{T}$,

$$\begin{aligned} \mathbf{z}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{z} &= \left[z_1 \sum_{i=1}^N x_{i1}^2 + z_2 \sum_{i=1}^N x_{i2}x_{i1},\ z_1 \sum_{i=1}^N x_{i1}x_{i2} + z_2 \sum_{i=1}^N x_{i2}^2 \right] \mathbf{z} \\ &= z_1^2 \sum_{i=1}^N x_{i1}^2 + 2z_1 z_2 \sum_{i=1}^N x_{i1}x_{i2} + z_2^2 \sum_{i=1}^N x_{i2}^2. \end{aligned}$$

Because the first and last terms must be positive, proving that this expression is greater than zero is equivalent to proving that their combined value is larger than the middle term:

$$z_1^2 \sum_{i=1}^N x_{i1}^2 + z_2^2 \sum_{i=1}^N x_{i2}^2 > 2z_1 z_2 \sum_{i=1}^N x_{i1}x_{i2}.$$

Defining $y_{i1} = z_1 x_{i1}$ and $y_{i2} = z_2 x_{i2}$ and substituting into our expression gives

$$\sum_{i=1}^N (y_{i1}^2 + y_{i2}^2) > 2 \sum_{i=1}^N y_{i1}y_{i2}.$$

Now, considering some arbitrary i,

$$\begin{aligned} y_{i1}^2 + y_{i2}^2 &> 2y_{i1}y_{i2} \\ y_{i1}^2 - 2y_{i1}y_{i2} + y_{i2}^2 &> 0 \\ (y_{i1} - y_{i2})^2 &> 0 \end{aligned}$$

which will only not be the case if $y_{i1} = y_{i2}$ and therefore $x_{i1} = x_{i2}$ – something unlikely to happen in practice. So, if $y_{i1}^2 + y_{i2}^2 > 2y_{i1}y_{i2}$ holds for any i, the summation of any number of these terms must also satisfy the inequality. Hence, $\mathbf{z}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{z}$ is always positive, our Hessian is negative definite and the solution corresponds to a maximum of the likelihood.

To ensure that our expression for $\widehat{\sigma^2}$ corresponds to a maximum of the likelihood, we differentiate Equation 2.34 again with respect to σ:

$$\frac{\partial^2 \log L}{\partial \sigma^2} = \frac{N}{\sigma^2} - \frac{3}{\sigma^4}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}}).$$

We can simplify this by substituting the value for $\widehat{\sigma^2}$ given in Equation 2.36, resulting in

$$\begin{aligned} \frac{\partial^2 \log L}{\partial \sigma^2} &= \frac{N}{\widehat{\sigma^2}} - \frac{3}{(\widehat{\sigma^2})^2} N\widehat{\sigma^2} \\ &= -\frac{2N}{\widehat{\sigma^2}}, \end{aligned}$$

which is always negative and hence $\widehat{\sigma^2}$ corresponds to a maximum.

2.8.4 Maximum likelihood favours complex models

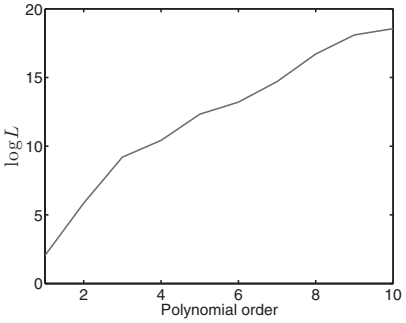
Plugging the expression for $\widehat{\sigma^2}$ (Equation 2.35) into the log-likelihood expression (Equation 2.31) gives us the value of the log-likelihood at the maximum:

$$\begin{aligned}\log L &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}}N\widehat{\sigma^2} \\ &= -\frac{N}{2}(1+\log 2\pi) - \frac{N}{2}\log\widehat{\sigma^2}.\end{aligned}$$

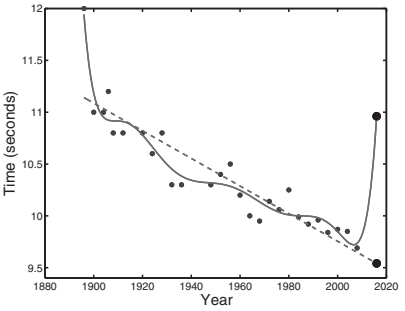
This tells us that the maximum value of L will keep increasing as we decrease $\widehat{\sigma^2}$. Recall that σ^2 is the variance of the noise incorporated into the model to capture effects that the deterministic part of our model (i.e. $f(\mathbf{x};\mathbf{w})$) cannot. One way to decrease σ^2 is to modify $f(\mathbf{x};\mathbf{w})$ so that it can capture more of the variability in the data – i.e., make it more flexible. For example, revisiting the Olympic men's 100 m data, we can investigate the increase in likelihood as model flexibility (or complexity) increases by fitting increasingly higher-order polynomial functions. Figure 2.11(a) shows that $\log L$ increases as polynomials of increasing order are fitted to the Olympic men's 100 m data (MATLAB script: `olymplike.m`). If we were to use $\log L$ to help choose which particular model to use, it would always point us to models of increasing complexity. This might seem like a sensible strategy – as $\widehat{\sigma^2}$ decreases, the deterministic part of our model must be capturing more of the variability in our data. However, consider the task of predicting the winning time for a year that we have not yet observed (e.g. 2016). Figure 2.11(b) shows first (dashed line) and eighth (solid line) order polynomial fits as well as their predictions for 2016 (shown as large dark circles). The more complex model makes a prediction of a winning time of close to 11 seconds (it would be one of the slowest ever) whereas the simpler model makes a much more realistic prediction. To the human eye, it looks like the simpler model has captured the important relationship in the data (the general downward trend) whilst the more complex model has not. This is a nice example of the trade-off between generalisation and over-fitting that we saw in Section 1.5. The simpler model is better able to generalise than the more complex one. The more complex model is over-fitting – we have given the model too much freedom and it is attempting to make sense out of what is essentially noise. In Section 1.6 we showed how regularisation could be used to penalise overcomplex parameter values. The same can be done with probabilistic models through the use of **prior distributions** on the parameter values. This will be introduced in the next chapter.

2.9 THE BIAS-VARIANCE TRADE-OFF

The trade-off between generalisation and over-fitting discussed in Section 1.5 is also sometimes described as the bias-variance trade-off. Imagine that we had access to the distribution from which the data were sampled, $p(\mathbf{x}, t)$. Using this distribution, we could, in theory, compute the expected value of the squared error between estimated parameter values and the true values. We would like this value, $\bar{\mathcal{M}}$, to be as low as



(a) Increase in log-likelihood as the polynomial order increases.



(b) First- and eighth- order polynomial functions fitted to the Olympic men's 100 m data. Large dark circles correspond to predictions for the 2016 Olympics.

FIGURE 2.11 Model complexity example with Olympic men's 100 m data.

possible. It can be decomposed into two terms called the bias $\mathcal{B}$ and the variance $\mathcal{V}$:

$$\bar{\mathcal{M}} = \mathcal{B}^2 + \mathcal{V}.$$

The bias describes the systematic mismatch between our model and the process that generated the data. A model that is too simple will have a high bias (underfitting). We can therefore decrease the bias and its contribution to the $\bar{\mathcal{M}}$ by making the model more complex. Unfortunately, more complex models have higher variance, thus increasing the $\mathcal{V}$ component of $\bar{\mathcal{M}}$. Finding the correct balance between generalisation and over/underfitting can thus also be thought of as finding the correct balance between bias and variance.

We omit further details here, but more details can be found in the suggested reading at the end of this chapter.

2.9.1 Summary

In the previous sections we have introduced a number of new concepts. Firstly, we made a case for explicitly modelling the noise (or errors) in our dataset. Making the assumption that these errors could be adequately modelled by a Gaussian random variable, we showed that we could compute a quantity called the *likelihood* that describes how likely our data is as a function of our model parameters. This is a reasonable quantity to maximise when choosing our parameters, and maximising the likelihood and minimising the squared loss give identical expressions for the optimal parameter values when we assume that the noise is Gaussian. In the remainder of the chapter we will look at two important benefits of explicitly modelling the noise: the ability to quantify the uncertainty in our parameters and the ability to express uncertainties in our predictions.

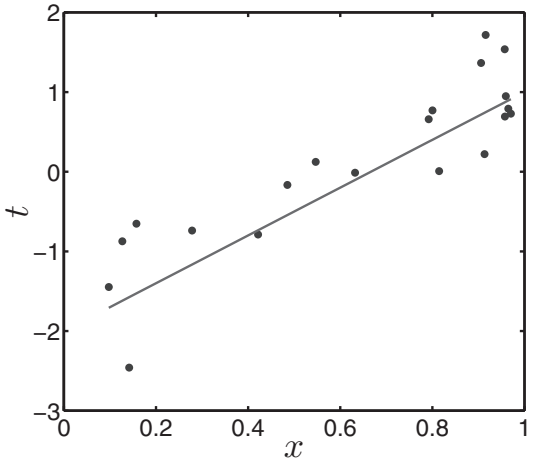


FIGURE 2.12 Data generated from the model given in Equation 2.39 and the true function.

2.10 EFFECT OF NOISE ON PARAMETER ESTIMATES

In this section we shall derive expressions for how much confidence we should place in our parameter estimates – how much could we change the straight line and still have a *good* model. If there is a lot of noise (σ^2 is high), it is likely that we could tolerate reasonably large changes in $\widehat{\mathbf{w}}$. If there is very little noise, the quality of the fit will deteriorate rapidly. Before we derive these expressions, it is useful to explore the variability in $\widehat{\mathbf{w}}$ by generating synthetic data. In particular, we shall generate lots of datasets with the same true $\mathbf{w}$ and σ^2 and see how our maximum likelihood estimate $\widehat{\mathbf{w}}$ varies. Consider the following model:

$$t_n = w_0 + w_1 x_n + \epsilon_n, \ \epsilon_n \sim \mathcal{N}(0, \sigma^2). \tag{2.39}$$

Assuming that the true parameter values are $w_0 = -2, w_1 = 3$ and the noise variance is $\sigma^2 = 0.5^2$, we can generate as many sets of responses $(t_1, \ldots, t_N)$ as we like for a particular set of attributes $(x_1, \ldots, x_N)$ and compute $\widehat{\mathbf{w}}$ for each set. An example of one such dataset and the *true* function can be seen in Figure 2.12, where the set of attributes consists of 20 values drawn from a uniform distribution between 0 and 1, i.e. $p(x) = \mathcal{U}(0, 1)$. Figure 2.13 shows the results of generating 10,000 datasets and fitting $\widehat{\mathbf{w}}$ in each case. The left panel shows a histogram where the height of each bar represents the number of datasets that resulted in parameter values within a particular range, and the right panel shows the same information as a contour plot. We can see a wide variability around the true values in both $\widehat{w_0}$ and $\widehat{w_1}$. It is hard, from these values, to get a feel for how much variability this implies in the model, so examples of $\widehat{\mathbf{w}}$ from ten datasets as well as the true function are plotted in Figure 2.14.

If we assume our real data to have been generated by such a process, it is useful to be able to quantify how variable our resulting estimates are. Unfortunately, we don't have access to many datasets from which we can compare values of $\widehat{\mathbf{w}}$. In the

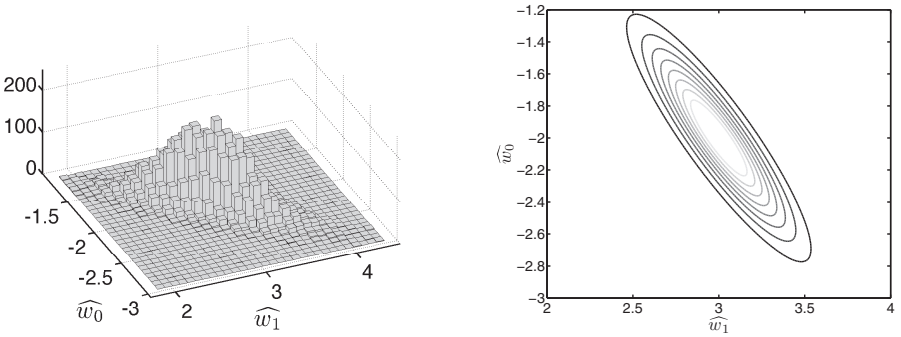


FIGURE 2.13 Variability in $\widehat{\mathbf{w}}$ for 10,000 datasets generated from the model described in Equation 2.39.

next section we will show how we can quantify this uncertainty using just the data that are available.

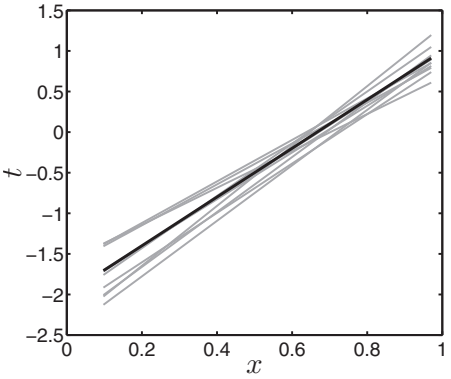


FIGURE 2.14 Functions inferred from ten datasets generated from the model given in Equation 2.39 as well as the true function (wider, darker line).

2.10.1 Uncertainty in estimates

We showed in the last section that the value we obtain for $\widehat{\mathbf{w}}$ is strongly influenced by the particular noise values in the data. In light of this, it would be useful to know how much uncertainty there was in $\widehat{\mathbf{w}}$. In other words, is this $\widehat{\mathbf{w}}$ unique in explaining the data well or are there many that could do almost as well?

To progress, we must be very clear about what $\mathbf{w}$ and $\widehat{\mathbf{w}}$ mean. We have hy-

pothesised a model which was responsible for the data. This model is

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

where $\mathbf{w}$ represents the *true* value of the parameters and ϵ_n is a random variable that we have defined to be normally distributed. This assumption means that the *generating* distribution (or likelihood), $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$, is a product of normal densities:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n\mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2).$$

In Section 2.5.4, we showed how a product of univariate Gaussian densities could be written as a multivariate Gaussian density with a diagonal covariance. It will be neater to work with a single multivariate Gaussian than a product over univariate ones. In this case, the multivariate Gaussian is

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}).$$

Satisfy yourself that the mean and covariance terms are correct. Now, $\widehat{\mathbf{w}}$ is an estimate of the true parameter value $\mathbf{w}$. Computing the expectation (Section 2.2.8) of $\widehat{\mathbf{w}}$ with respect to the generating distribution will tell us what we expect $\widehat{\mathbf{w}}$ to be, on average:

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\} = \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) d\mathbf{t}.$$

Substituting $\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$ into this expression allows us to evaluate the integral:

$$\begin{aligned} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} \int \mathbf{t} p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) d\mathbf{t} \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\} \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} \\ &= \mathbf{w}, \end{aligned} \tag{2.40}$$

where we have used the fact that the expected value of a normally distributed random variable is equal to its mean ($\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\} = \mathbf{X}\mathbf{w}$ because $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$).

This result tells us that the expected value of our approximation $\widehat{\mathbf{w}}$ is the true parameter value. We will consider this in more detail later in the chapter, but it means that our estimator is **unbiased** – it is not, on average, too big or too small.

This potential variability in the estimate of $\widehat{\mathbf{w}}$ is encapsulated in its *covariance matrix*. For our purposes, this covariance matrix provides us with two useful pieces of information. The diagonal elements (the variances of the individual elements in $\widehat{\mathbf{w}}$) tell us how much variability we might expect in the individual parameters – i.e. how well they are defined by the data. In our experiment above, the parameters appeared to vary quite a lot, suggesting that they were not defined very well by the data. The off-diagonal elements tell us how the parameters co-vary – if the values are high and positive, it tells us that increasing one will require an increase in the other to maintain a *good* model. Large negative values tell us the opposite – increasing

one will cause a decrease in the other. Values close to zero tell us that the parameters are not dependent on one another. For the example described above, it looks (see Figure 2.13) like increasing w_1 causes a decrease in w_0 so we might expect the off-diagonal elements in the covariance matrix to be negative.

In Section 2.2.8, we derived a general expression for the covariance matrix (Equation 2.16). Substituting $\mathbf{t}$ and $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$ into this expression, and using the previous result, $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\} = \mathbf{w}$, gives us

$$\begin{aligned}\mathsf{cov}\{\widehat{\mathbf{w}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T}\right\} - \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\}\,\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\}^\mathsf{T} \\ &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T}\right\} - \mathbf{w}\mathbf{w}^\mathsf{T} \qquad (2.41)\end{aligned}$$

where we have used the expectation of $\widehat{\mathbf{w}}$ that we derived above. To compute this quantity, we will start with the first term. It can expanded by substituting $\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$ and all of the terms that do not involve $\mathbf{t}$ can be removed from the expectation:

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T}\right\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t})((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t})^\mathsf{T}\right\} \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\,\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\mathsf{T}\right\}\,\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}. \qquad (2.42)\end{aligned}$$

Now, $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$. Therefore, the covariance of $\mathbf{t}$ is, by definition, $\sigma^2\mathbf{I}$ and its mean is $\mathbf{X}\mathbf{w}$. By the same line of derivation that allowed us to reach Equation 2.41, we have

$$\mathsf{cov}\{\mathbf{t}\} = \sigma^2\mathbf{I} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\mathsf{T}\right\} - \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}\,\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}^\mathsf{T}.$$

Therefore, we can rearrange this expression to obtain an expression for $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\mathsf{T}\right\}$:

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\mathsf{T}\right\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}\,\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}^\mathsf{T} + \sigma^2\mathbf{I} \\ &= \mathbf{X}\mathbf{w}(\mathbf{X}\mathbf{w})^\mathsf{T} + \sigma^2\mathbf{I} \\ &= \mathbf{X}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T} + \sigma^2\mathbf{I}.\end{aligned}$$

Substituting this into Equation 2.42 gives

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T}\right\} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \\ &\quad + \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \\ &= \mathbf{w}\mathbf{w}^\mathsf{T} + \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}. \qquad (2.43)\end{aligned}$$

Finally, substituting this into Equation 2.41 gives the expression for the covariance of $\widehat{\mathbf{w}}$:

$$\begin{aligned}\mathsf{cov}\{\widehat{\mathbf{w}}\} &= \mathbf{w}\mathbf{w}^\mathsf{T} + \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} - \mathbf{w}\mathbf{w}^\mathsf{T} \\ &= \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \qquad (2.44)\end{aligned}$$

which is the negative of the inverse of the Hessian matrix of second derivatives derived previously (Equation 2.38), i.e.

$$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}}\right)^{-1}. \tag{2.45}$$

This result tells us that the certainty/uncertainty in the parameters (as described by $\text{cov}\{\widehat{\mathbf{w}}\}$) is directly linked to the second derivative of the log-likelihood. The second derivative of the log-likelihood tells us about the curvature of the likelihood function. Therefore, low curvature corresponds to a high level of uncertainty in parameters and high curvature to a low level. In other words, we have an expression that tells us how much **information** our data gives us regarding our parameter estimates. In fact, our matrix, $\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, is the inverse of something called the **Fisher Information Matrix** ($\mathcal{I}$). The Fisher Information Matrix is computed as the expected value of the martrix of second derivatives of the log-likelihood:

$$\mathcal{I} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{-\frac{\partial^2 \log p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}}\right\}.$$

We already know what the bit in the brackets is – it is the Hessian matrix we calculated earlier – so

$$\mathcal{I} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}\right\}$$

which, because the argument of the expectation is a constant is just

$$\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}. \tag{2.46}$$

The elements of $\mathcal{I}$ tell us how much information (the more negative the information value is, the more information is present) the data provides about a particular parameter (diagonal elements) or pairs of parameters (off-diagonal elements). Intuitively, if our data is very noisy, the information content is lower. In general, if the information content is high, the data can inform a very accurate parameter estimate and the covariance of $\widehat{\mathbf{w}}$ will be low ($\text{cov}\{\widehat{\mathbf{w}}\} = \mathcal{I}^{-1}$). If the information content is low, the covariance will be high (see Exercises 2.13 and 2.14).

As an example, look at the top line in Figure 2.15. The left hand plot shows the data and the true function ($t = 3x - 2$) and the right hand plot shows the likelihood as a function of the two parameters. We can see that the likelihood function has a low curvature (contour lines are reasonably far apart) because of the large noise level, and, as such, many sets of parameters will result in a reasonable model. A low curvature should, from Equation 2.45, correspond to high covariance in $\widehat{\mathbf{w}}$. The Fisher information and covariance matrices are

$$\mathcal{I} = \begin{bmatrix} 50.0000 & 24.3311 \\ 24.3311 & 15.8953 \end{bmatrix}, \ \text{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.0784 & -0.1200 \\ -0.1200 & 0.2466 \end{bmatrix}.$$

It is difficult to know if these correspond to high or low information and covariance without context. This can be provided by comparing them with those obtained from the second dataset (second row in Figure 2.15). This dataset has much less noise and the corresponding likelihood curvature is much higher (the contour lines are closer together). In this case, the information and covariance matrices are

$$\mathcal{I} = \begin{bmatrix} 1.2500 \times 10^3 & 0.6083 \times 10^3 \\ 0.6083 \times 10^3 & 0.3974 \times 10^3 \end{bmatrix}, \ \text{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.0031 & -0.0048 \\ -0.0048 & 0.0099 \end{bmatrix}$$

which have significantly higher (in $\mathcal{I}$) and lower (in $\text{cov}\{\widehat{\mathbf{w}}\}$) values.

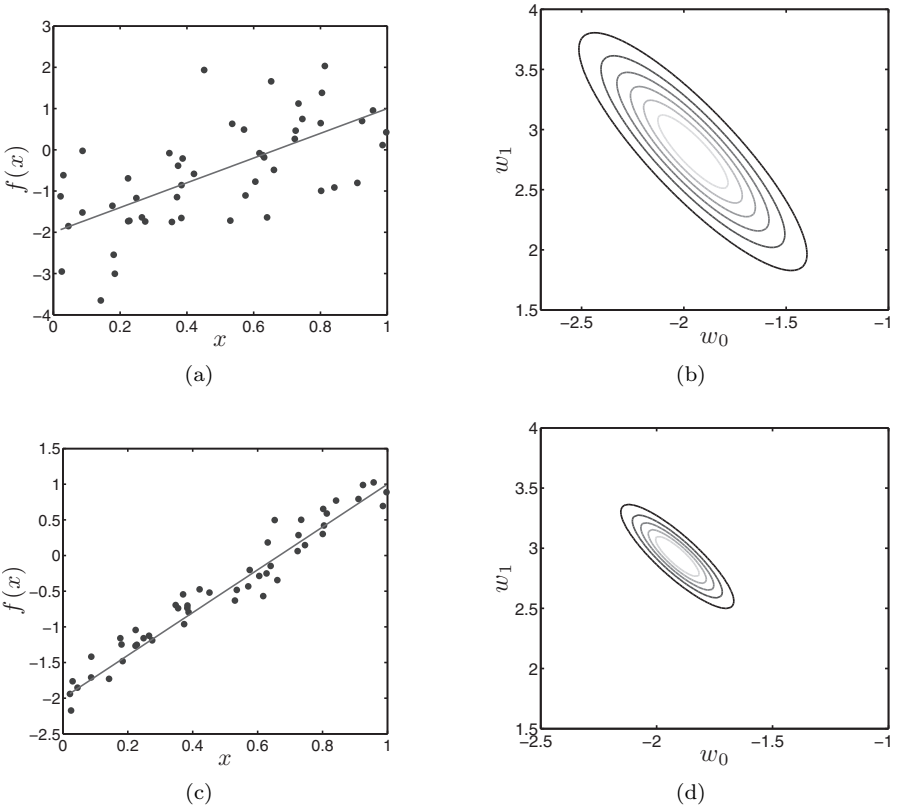


FIGURE 2.15 Two example datasets with different noise levels and the corresponding likelihood function.

2.10.2 Comparison with empirical values

At the start of Section 2.10 we generated many sets of responses for a set of attributes using the model (and associated noise distribution) given in Equation 2.39. If we use $\widehat{\mathbf{w}}_s$ to describe the parameters obtained from the sth dataset, the empirical covariance matrix can be computed as

$$\widehat{\mathsf{cov}\{\widehat{\mathbf{w}}\}} = \frac{1}{S}\sum_{s=1}^{S}(\widehat{\mathbf{w}}_s - \widehat{\boldsymbol{\mu}})(\widehat{\mathbf{w}}_s - \widehat{\boldsymbol{\mu}})^{\mathsf{T}}$$

where

$$\widehat{\boldsymbol{\mu}} = \frac{1}{S}\sum_{s=1}^{S}\widehat{\mathbf{w}}_s.$$

Using the values shown in Figure 2.13, the empirical covariance matrix is

$$\widehat{\mathsf{cov}\{\widehat{\mathbf{w}}\}} = \begin{bmatrix} 0.0627 & -0.0809 \\ -0.0809 & 0.1301 \end{bmatrix}.$$

Using Equation 2.44 and the true value of $\sigma^2 = 0.5^2$, the theoretical covariance matrix is

$$\mathsf{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.0638 & -0.0821 \\ -0.0821 & 0.1317 \end{bmatrix}$$

which is very close to our empirical value. Normally, we do not have access to a bottomless supply of data and so we can use the theoretical covariance matrix to help understand the variability present in our data. The off-diagonal elements are negative – increasing one of the parameters forces the other to decrease.

To compute the theoretical covariance matrix, we have used the true noise variance. If we take one arbitrary dataset, we can estimate the variance (using Equation 2.35) as $\sigma^2 = 0.2080$ (the true value is $\sigma^2 = 0.25$). The covariance matrix using the estimated variance is

$$\mathsf{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.0530 & -0.0683 \\ -0.0683 & 0.1095 \end{bmatrix}.$$

Because the estimated value of σ^2 is lower than the true value, the values in this matrix are lower than those when the true noise value is used. This suggests that the uncertainty is underestimated and our predictions will be overconfident. The systematic underestimation of noise variance in maximum likelihood is discussed more thoroughly in Section 2.11.2.

At the start of Section 2.10 we saw that changes in the exact values of the noise changed the parameter estimates. In reality we cannot generate many datasets with which to estimate this uncertainty in parameter values. However, we have derived an expression for the covariance of $\widehat{\mathbf{w}}$ that can be used to approximate the uncertainty in parameters. Before we move on to variability in predictions, we will look at the uncertainty present in the maximum likelihood estimations from the Olympic data.

2.10.3 Variability in model parameters – Olympic data

Using the now familiar men's Olympic 100 m data, and the standard linear function

$$f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^{\mathsf{T}}\mathbf{x},$$

we know that the maximum likelihood value of $\mathbf{w}$, $\widehat{\mathbf{w}}$, will be $[36.4165, -0.0133]^{\mathsf{T}}$ (from Equation 2.33). The maximum likelihood variance value, $\widehat{\sigma^2}$, can be computed using Equation 2.37 and is $\widehat{\sigma^2} = 0.0503$. Using Equation 2.44, and using $\widehat{\sigma^2}$ as an estimate of σ^2, we can compute the covariance matrix of the estimate:

$$\mathsf{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 5.7972 & -0.0030 \\ -0.0030 & 1.5204e-06 \end{bmatrix}.$$

Taking the diagonal elements, we can see that the variance of $\widehat{w_0}$ (5.7972) is much higher than the variance in $\widehat{w_1}$ ($1.5204e-06$), suggesting that we could tolerate bigger changes in $\widehat{w_0}$ than $\widehat{w_1}$ and still be left with a reasonably good model. Partly, this can be explained by the fact that $\widehat{w_0}$ has a much higher absolute value. The negativity of the off-diagonal elements tell us that, if we were to slightly increase either $\widehat{w_0}$ or $\widehat{w_1}$, we would have to slightly decrease the other. This is relatively intuitive – if we were to slightly increase $\widehat{w_0}$, the whole line would move up and the best value of $\widehat{w_1}$ would have to be decreased slightly (thereby producing a steeper negative gradient) to pass as close as possible to all of the data points.

Another way to get a feeling for the meaning of $\mathsf{cov}\{\widehat{\mathbf{w}}\}$ is to look at the variability in models that it suggests. To do this, we can assume that $\widehat{\mathbf{w}}$ is a random variable with a Gaussian distribution

$$\mathbf{w} \sim \mathcal{N}(\widehat{\mathbf{w}}, \mathsf{cov}\{\widehat{\mathbf{w}}\}). \tag{2.47}$$

From this density, we can sample several instances of $\mathbf{w}$ and plot the resulting models. An example of ten instances is shown in Figure 2.16. We can see that there is very little change in gradient (w_1) across the ten samples but that this small gradient change would, if we extrapolated back to year zero, result in quite a large change of w_0. This is reflected by the values in $\mathsf{cov}\{\widehat{\mathbf{w}}\}$, as already discussed. The idea of having a distribution over model parameters rather than a single *best* value is very important in Machine Learning and is introduced in the next chapter.

2.11 VARIABILITY IN PREDICTIONS

In Chapter 1 we made some predictions about 100 m winning times in future Olympics. We argued that these predictions were not very useful, as they took the form of exact values. It would seem more sensible to predict a range of values in which we think the winning time might fall. If we are quite certain about our prediction, this range might be small; if we are less certain, it might be large. So, as well as obtaining an indication of the variability of our parameter estimate, $\widehat{\mathbf{w}}$, it makes sense to provide indications of any variability or uncertainty in our predictions. Suppose we observe a new set of attributes, $\mathbf{x}_{\mathsf{new}}$. We would like to predict the output, t_{new}, and the variability associated with this output, σ^2_{new}.

To predict t_{new}, we multiply $\mathbf{x}_{\mathsf{new}}$ by the best set of model parameters, $\widehat{\mathbf{w}}$:

$$t_{\mathsf{new}} = \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\mathsf{new}}. \tag{2.48}$$

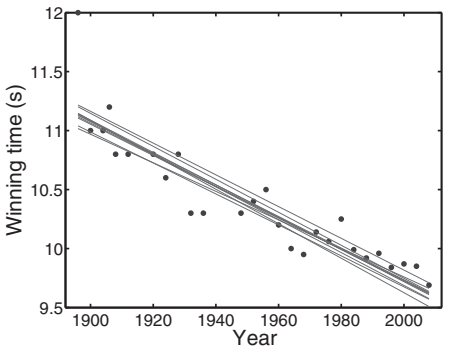


FIGURE 2.16 Ten samples of $\mathbf{w}$ using the distribution given in Equation 2.47.

To check that this is sensible, we can compute its expectation:

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{t_{\text{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\ &= \mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}}\end{aligned}$$

where we have used Equation 2.40. The expected value of our prediction is the new data attribute multiplied by the *true* $\mathbf{w}$. In Section 2.2.8 we derived a general expression for variance. In our case, this is

$$\sigma^2_{\text{new}} = \mathsf{var}\{t_{\text{new}}\} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t^2_{\text{new}}\right\} - (\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{t_{\text{new}}\})^2.$$

To evaluate this expression, we need to first substitute $t_{\text{new}} = \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}}$:

$$\begin{aligned}\mathsf{var}\{t_{\text{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}})^2\right\} - (\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}})^2 \\ &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{x}^{\mathsf{T}}_{\text{new}}\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}}\right\} - \mathbf{x}^{\mathsf{T}}_{\text{new}}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}}.\end{aligned}$$

Substituting our now familiar expression for $\widehat{\mathbf{w}}$,

$$\mathsf{var}\{t_{\text{new}}\} = \mathbf{x}^{\mathsf{T}}_{\text{new}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\right\}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}} - \mathbf{x}^{\mathsf{T}}_{\text{new}}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}}.$$

Using the expression for the $\mathsf{cov}\{\mathbf{t}\}$ (Equation 2.10.1) allows us to compute the expectation and simplify the expression:

$$\begin{aligned}\mathsf{var}\{t_{\text{new}}\} &= \mathbf{x}^{\mathsf{T}}_{\text{new}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}})\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}} - \mathbf{x}^{\mathsf{T}}_{\text{new}}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\ &= \sigma^2\mathbf{x}^{\mathsf{T}}_{\text{new}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}} + \mathbf{x}^{\mathsf{T}}_{\text{new}}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}} - \mathbf{x}^{\mathsf{T}}_{\text{new}}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\ &= \sigma^2\mathbf{x}^{\mathsf{T}}_{\text{new}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}.\end{aligned}$$

Note that, by substituting our expression for $\mathsf{cov}\{\widehat{\mathbf{w}}\}$ (Equation 2.41), this expression can be rewritten as

$$\sigma^2_{\text{new}} = \mathbf{x}^{\mathsf{T}}_{\text{new}}\mathsf{cov}\{\widehat{\mathbf{w}}\}\mathbf{x}_{\text{new}}.$$

To summarise, our prediction and associated variance are given as

$$t_{\text{new}} = \mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} = \mathbf{x}_{\text{new}}^{\mathsf{T}}\widehat{\mathbf{w}}, \tag{2.49}$$

$$\sigma_{\text{new}}^2 = \sigma^2\mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}. \tag{2.50}$$

σ^2 is the *true* variance of the dataset noise. In its place, we can use our estimate, $\widehat{\sigma^2}$.

2.11.1 Predictive variability – an example

Figure 2.17(a) shows the function $f(x) = 5x^3 - x^2 + x$ and data points sampled from this function and corrupted by Gaussian noise with mean zero and variance 1000. In Figures 2.17(b), 2.17(c) and 2.17(d) we can see $t_{\text{new}} \pm \sigma_{\text{new}}^2$ for linear, cubic and sixth-order models, respectively (MATLAB script: `predictive_variance_example.m`).

The linear model has very high predictive variance. It is unable to model the deterministic trend in the data very well, and much of the variability of the data is assumed to be noise. The cubic model is better able to model the trend (it is the correct order) and this is reflected in its much more confident predictions. The sixth-order model is over-complex – it has too much freedom and can therefore fit the data well for quite a large range of parameter values. This uncertainty in $\widehat{\mathbf{w}}$ feeds through to increased predictive variability – if we are less sure on the parameter values, we're going to be less sure of the predictions too. This point can be demonstrated by computing $\mathsf{cov}\{\widehat{\mathbf{w}}\}$ for the third- and sixth-order models and then sampling functions just as we did in Section 2.10.3. Figure 2.18 shows 20 functions drawn from a Gaussian with mean $\widehat{\mathbf{w}}$ and covariance $\mathsf{cov}\{\widehat{\mathbf{w}}\}$ for the third- and sixth-order models (the plot is zoomed into a small region of x and the darker line shows the true function) (MATLAB script: `predictive_variance_example.m`). The increased variability in possible functions caused by the increase in parameter uncertainty is clear for the sixth-order model.

A final interesting point is that, for all models, the predictive variance increases as we move towards the edge of the data. The model is less confident in areas where it has less data – an appealing property. In Chapter 1 we pointed out that making point indefinitely into the future (i.e. beyond the range of the training data) was not very sensible. We now have a model that will make predictions beyond the range of the training data, but will do so with increasing uncertainty, which is likely to be more useful. We also observe this effect towards the centre of the data (particularly Figure 2.17(d)) where there is a small gap (not many instances at around $x = 1$). Exercise 2.12 gives you the opportunity to investigate this effect further.

2.11.2 Expected values of the estimators

In Section 2.10.1 we computed the expected value of our estimate $\widehat{\mathbf{w}}$. This expectation is taken with respect to the generating density $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$

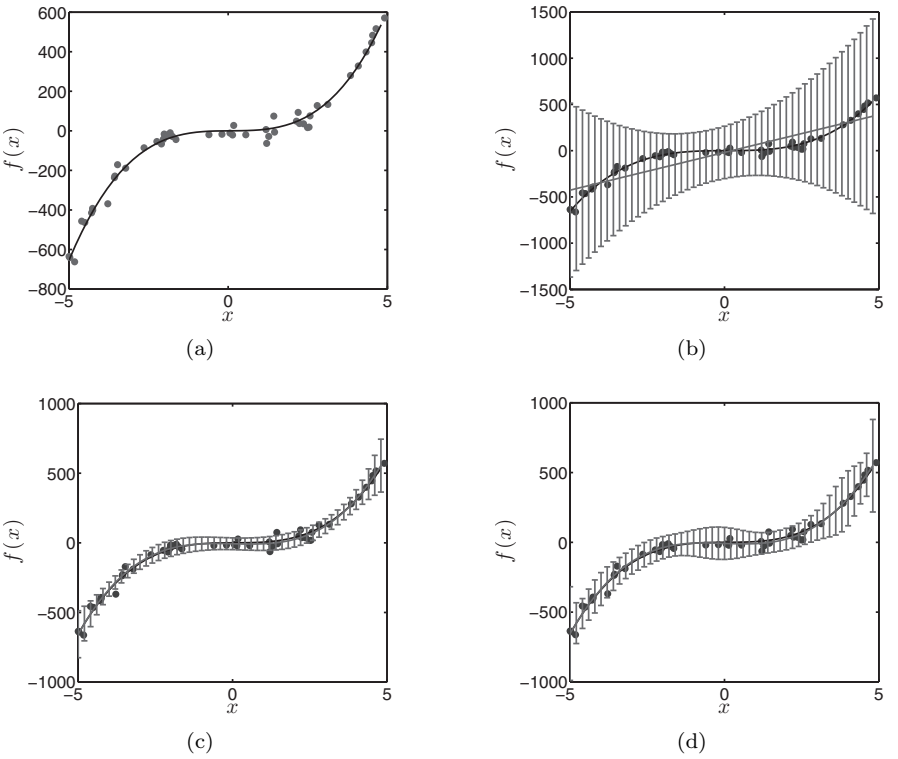


FIGURE 2.17 (a) Example data set. (b), (c) and (d) predictive error bars for a linear, cubic and sixth-order model, respectively.

and is repeated here:

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X})}\left\{(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\right\} \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X})}\{\mathbf{t}\} \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} \\ &= \mathbf{I}\mathbf{w} = \mathbf{w}\end{aligned}$$

where we have used the expression for $\widehat{\mathbf{w}}$ ($\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$) and the fact that the expected value of a Gaussian random variable ($\mathbf{t}$) is equal to the mean of the Gaussian ($\mathbf{X}\mathbf{w}$). So, the expected value of our estimate, $\widehat{\mathbf{w}}$, is the true value, $\mathbf{w}$. This is an important property of $\widehat{\mathbf{w}}$ – it tells us that $\widehat{\mathbf{w}}$ is an unbiased estimator – it is neither consistently too high nor too low. Another way of thinking about this is to think back to the experiment at the start of Section 2.10. There, for a set of attributes $x_1, \ldots, x_N$, we generated many sets of responses and looked at how much influence different particular noise values had on $\widehat{\mathbf{w}}$. Because $\widehat{\mathbf{w}}$ is unbiased, it should, on average, be correct. So, if we took the average of all of the different $\widehat{\mathbf{w}}$ values obtained in our experiment, it should be very close to the truth. In fact,

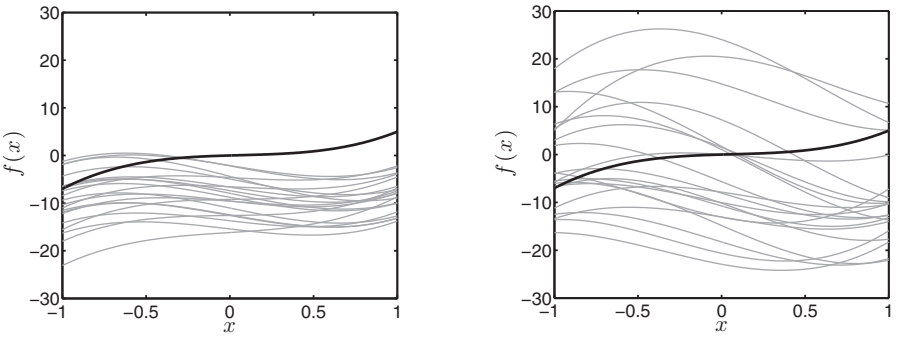


FIGURE 2.18 Examples of functions drawn with parameters from a Gaussian with mean $\widehat{\mathbf{w}}$ and covariance $\mathsf{cov}\{\widehat{\mathbf{w}}\}$ for the example dataset shown in Figure 2.17(a).

taking this average, we get $\widehat{w_0} = -2.0007$ and $\widehat{w_1} = 3.0008$ which are both very close to the true values: $w_0 = -2, w_1 = 3$.

We can do the same for the estimate of the noise variance, $\widehat{\sigma^2}$. Recall the expression for $\widehat{\sigma^2}$ from Equation 2.37:

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}}).$$

Taking the expectation with respect to $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$ and doing some manipulation gives

$$\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}}\right\} \\
&= \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\right\} \\
&= \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\mathsf{T}\mathbf{t}\right\} \\
&\quad -\frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\right\}.
\end{aligned} \tag{2.51}$$

Comment 2.8 – Matrix trace: The trace of a square matrix $\mathbf{A}$, denoted $\mathsf{Tr}(\mathbf{A})$, is the sum of the diagonal elements of $\mathbf{A}$. For example, if

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1D} \\ A_{21} & A_{22} & \cdots & A_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ A_{D1} & A_{D2} & \cdots & A_{DD} \end{bmatrix},$$

then

$$\mathsf{Tr}(\mathbf{A}) = \sum_{d=1}^{D} A_{dd}.$$

It follows that, if $\mathbf{A} = \mathbf{I}_D$, i.e. the $D \times D$ identity matrix,

$$\mathsf{Tr}(\mathbf{I}_D) = \sum_{d=1}^{D} 1 = D.$$

A useful identity that we will often use is that

$$\mathsf{Tr}(\mathbf{AB}) = \mathsf{Tr}(\mathbf{BA}).$$

Also, the trace of a scalar is just equal to the scalar value (a scalar could be thought of as a 1×1 matrix), i.e.

$$\mathsf{Tr}(a) = a,$$

or, if $\mathbf{w} = [w_1, \ldots, w_D]^\mathsf{T}$,

$$\mathsf{Tr}(\mathbf{w}^\mathsf{T}\mathbf{w}) = \mathbf{w}^\mathsf{T}\mathbf{w}$$

because the result of $\mathbf{w}^\mathsf{T}\mathbf{w}$ is a scalar.

We have seen the expectation of the form $\mathbf{t}\mathbf{t}^\mathsf{T}$ before but not $\mathbf{t}^\mathsf{T}\mathbf{t}$ ($= \mathbf{t}^\mathsf{T}\mathbf{I}\mathbf{t}$) or $\mathbf{t}^\mathsf{T}\mathbf{A}\mathbf{t}$. When $\mathbf{t}$ is a Gaussian random variable, expectations of the form $\mathbf{t}^\mathsf{T}\mathbf{A}\mathbf{t}$ are given by

$$\begin{aligned} \mathbf{t} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \mathbf{E}_{p(\mathbf{t})}\left\{\mathbf{t}^\mathsf{T}\mathbf{A}\mathbf{t}\right\} &= \mathsf{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\mathsf{T}\mathbf{A}\boldsymbol{\mu}, \end{aligned}$$

where $\mathsf{Tr}()$ is the trace function (see Comment 2.8). For the first term on the right hand side of Equation 2.51, $\mathbf{A} = \mathbf{I}_N$ (note that $\mathbf{t}^\mathsf{T}\mathbf{t} = \mathbf{t}^\mathsf{T}\mathbf{I}_N\mathbf{t}$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix) and in the second term, $\mathbf{A} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$. In both cases, $\boldsymbol{\mu} = \mathbf{X}\mathbf{w}$ and $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_N$. Substituting the necessary values into Equation 2.51 gives

$$\begin{aligned} \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \frac{1}{N}\left(\mathsf{Tr}(\sigma^2\mathbf{I_N}) + \mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}\right) \\ &- \frac{1}{N}\left(\mathsf{Tr}(\sigma^2\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}) + \mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}\right), \end{aligned}$$

since $\mathbf{I}_N\mathbf{I}_N = \mathbf{I}_N$. Now, $\mathsf{Tr}(\sigma^2\mathbf{A}) = \sigma^2\mathsf{Tr}(\mathbf{A})$ and $\mathsf{Tr}(\mathbf{I_N}) = N$ by definition. Using

these, we can simplify the expression to

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \sigma^2 + \frac{1}{N}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} - \frac{\sigma^2}{N}\mathsf{Tr}(\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}) - \frac{1}{N}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} \\ &= \sigma^2 - \frac{\sigma^2}{N}\mathsf{Tr}(\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}) \\ &= \sigma^2\left(1 - \frac{1}{N}\mathsf{Tr}(\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})\right).\end{aligned}$$

Finally, we need to use the fact that $\mathsf{Tr}(\mathbf{AB}) = \mathsf{Tr}(\mathbf{BA})$ and therefore the first $\mathbf{X}$ inside the trace function can be moved to be the last:

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \sigma^2\left(1 - \frac{1}{N}\mathsf{Tr}((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X})\right) \\ &= \sigma^2\left(1 - \frac{1}{N}\mathsf{Tr}(\mathbf{I}_D)\right) \\ &= \sigma^2\left(1 - \frac{D}{N}\right), \qquad (2.52)\end{aligned}$$

where D is the number of attributes (the number of columns in $\mathbf{X}$).

Assuming that $D < N$ (i.e. the number of attributes we measure for each data point is smaller than the number of data points), then our estimate of the variance will, on average, be lower than the true variance:

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} < \sigma^2.$$

Unlike $\widehat{\mathbf{w}}$, this estimator *is* **biased**.

We can see this bias by returning to our synthetic experiment. The average value of $\widehat{\sigma^2}$ over all of the datasets was 0.2264. The true value is $\sigma^2 = 0.5^2 = 0.25$. We can see that the average value is indeed too low. For this example, $D = 2$ and $N = 20$, so our theoretical expected value is

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right) = 0.25\left(1 - \frac{2}{20}\right) = 0.2250$$

which is close to the observed average.

From Equation 2.52, we notice that one way to decrease the bias is to make D/N smaller. D is normally fixed, but we can increase N. In Figure 2.19 we can see the effect of increasing N from 20 to 10,000 (MATLAB script: `w_variation_demo.m`). The theoretical (dashed) curve and the empirical (solid) curve (created by rerunning our previous experiment with different numbers of observations, N) are in close agreement and converge towards the true value of $\sigma^2 = 0.25$ as the amount of data increases.

It is possible to provide an intuitive explanation for the bias in $\widehat{\sigma^2}$. The expression for the ML estimate of σ^2 is

$$\widehat{\sigma^2} = \frac{1}{N}\left(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}}\right). \qquad (2.53)$$

It is possible to rearrange this to be equal to the sum of squared errors between the predictions and the true responses (see Exercise 2.11):

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^\mathsf{T}\widehat{\mathbf{w}})^2.$$

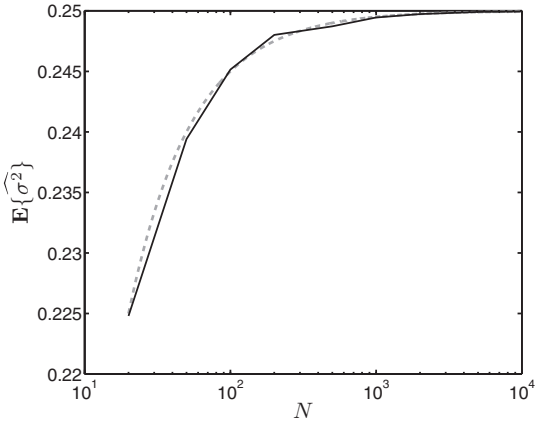


FIGURE 2.19 Evolution of the theoretical (dashed line) and empirical (solid line) estimates of $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ as the number of data points increases.

This tells us that the closer the model gets to the data, the smaller $\widehat{\sigma^2}$. Now imagine the true value of $\mathbf{w}$ and our estimate $\widehat{\mathbf{w}}$. Which will get closer to the data? The maximum likelihood estimate, $\widehat{\mathbf{w}}$, is identical to the minimum loss estimate. It is, by definition, the set of parameters that gets closest to the data and therefore minimises $\widehat{\sigma^2}$. The value of $\widehat{\sigma^2}$ that we would get if we used the true value $\mathbf{w}$ instead of $\widehat{\mathbf{w}}$ in Equation 2.53 would have to be the same or higher than the value we get with $\widehat{\mathbf{w}}$. Because we are finding the value of $\mathbf{w}$ that minimises the noise, we will, on average, end up with a lower level of noise than the true value.

2.12 CHAPTER SUMMARY

In the preceding sections, we have covered a lot of material. An introduction to random variables provided the foundations required to be able to model the errors between the data and the proposed deterministic model. By explicitly modelling these errors, we have seen how the least squares solution from Chapter 1 is equivalent to the solution obtained by maximising a different quantity called the *likelihood* if the noise in the data is assumed to be normally distributed. The benefit of the likelihood approach is the ability to quantify the uncertainty in our parameter estimates and hence also, crucially, in our predictions. This allows us to move away from exact predictions (which will certainly be wrong) to ranges of values (e.g. $t_{\text{new}} \pm \sigma^2_{\text{new}}$). In most applications this will be much more useful. Finally, we looked at some theoretical properties of the maximum likelihood parameter values and saw that, although our estimate $\widehat{\mathbf{w}}$ is unbiased, $\widehat{\sigma^2}$ is, on average, biased to be too low.

2.13 EXERCISES

2.1 Would the errors in the 100 m linear regression (shown in Figure 2.1) be best modelled with a discrete or continuous random variable?

2.2 By using the fact that, when rolling a die, all outcomes are equally likely and by using the constraints given in Equations 2.1 and 2.2, compute the probabilities of the dice landing with each of the six faces facing up.

2.3 Y is a random variable that can take any positive integer value. The likelihood of these outcomes is given by the Poisson pdf

$$p(y) = \frac{\lambda^y}{y!} \exp\{-\lambda\}.$$

By using the fact that, for a discrete random variable, the pdf gives the probabilities of the individual events occurring and that probabilities are additive, (a) compute the probability that $Y \leq 4$ for $\lambda = 5$, i.e. $P(Y \leq 4)$.
(b) Using the result of (a) and the fact that one outcome has to happen, compute the probability that $Y > 4$. (Hint: one of the two events, $Y \leq 4$ and $Y > 4$, *has* to happen.)

2.4 Y is a random variable with a uniform density $p(y) = \mathcal{U}(a, b)$. Derive $\mathbf{E}_{p(y)}\{\sin(y)\}$. Note that $\int \sin(y)\ dy = -\cos(y)$. Compute $\mathbf{E}_{p(y)}\{\sin(y)\}$ for $a = 0, b = 1$. Modify `approx_expected_value.m` to compute a sample-based approximation to this value and observe how the approximation improves with the number of samples drawn.

2.5 Assume that $p(\mathbf{w})$ is the Gaussian pdf for a D-dimensional vector $\mathbf{w}$ given in Equation 2.28. By expanding the vector notation and rearranging, show that using $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ as the covariance matrix assumes independence of the D elements of $\mathbf{w}$. You will need to be aware that the determinant of a matrix that only has entries on the diagonal ($|\sigma^2\mathbf{I}|$) is the product of the diagonal values and that the inverse of the same matrix is constructed by simply inverting each element on the diagonal. (Hint: a product of exponentials can be expressed as an exponential of a sum.)

2.6 Using the same setup as Exercise 2.5 above, see what happens if we use a diagonal covariance matrix with different elements on the diagonal, i.e.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_D^2 \end{bmatrix}.$$

2.7 Show that for a first-order polynomial, the diagonal elements of the Hessian matrix of second derivatives of the log-likelihood is equivalent to (they will differ by a multiplicative constant) the second derivatives in Equation 1.9.

2.8 Assume that a dataset of N values, $x_1, \dots, x_N$, was sampled from a Gaussian distribution. Assuming that the data are IID, find the maximum likelihood estimate of the Gaussian mean and variance. (Hint: start by writing down the combined likelihood of all N data points and note that the product of an exponential function can be written as the exponential of a sum.)

2.9 Assume that a dataset of N binary values, $x_1, \ldots, x_N$, was sampled from a Bernoulli. Compute the maximum likelihood estimate for the Bernoulli parameter.

2.10 Obtain the maximum likelihood estimates of the mean vector and covariance matrix of a multivariate Gaussian density given N observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

2.11 Show that the maximum likelihood estimate of the noise variance in our linear model,

$$\widehat{\sigma^2} = \frac{1}{N}\left(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}}\right),$$

can also be expressed as

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}_n^\mathsf{T}\mathbf{w})^2.$$

(Hint: work backwards from the second expression.)

2.12 Using `predictive_variance_example.m`, generate a dataset and remove all values for which $-1.5 \leq x \leq 1.5$. Observe the effect this has on the predictive variance in this range.

2.13 Compute the Fisher information for the parameter of a Bernoulli distribution.

2.14 Compute the Fisher information matrix for the components of the mean vector in a multivariate Gaussian density.

2.14 FURTHER READING

[1] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

This book is an excellent resource for many machine learning concepts. In particular, it includes a detailed discussion of the bias-variance trade-off.

[2] J. H. McColl. *Probability*. Elsevier, 1995.

A very accessible introduction to probability theory.

[3] Paul Meyer. *Introductory Probability and Statistical Applications*. Addison-Wesley, 1978.

An excellent resource for introductory probability theory.

[4] J. Rosenthal. *A First Look at Rigorous Probabillity Theory*. World Scientific Publishing Company, 2006.

This is a very accessible book to begin exploring measure theory – the branch of mathematics that underpins probability theory.

[5] Michael Tipping and Christopher Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999.

An interesting application of maximum likelihood. Here it is applied to one of the first probabilistic approaches to the the classical statistical problem of Principal Component Analysis.

Taylor & Francis

Taylor & Francis Group

http://taylorandfrancis.com