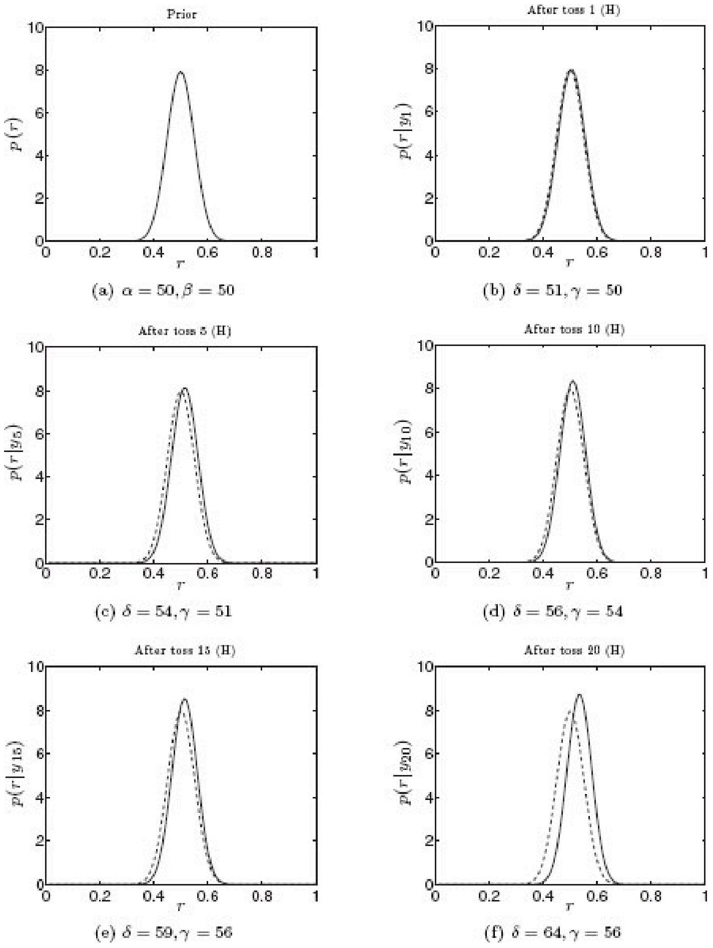
decreasing to such an extent that there was a single value of r that had probability 1 of occurring with $p(r|y_N)$ being zero everywhere else. The expectation we are calculating is



(a) $\alpha = 50, \beta = 50$

(b) $\delta = 51, \gamma = 50$

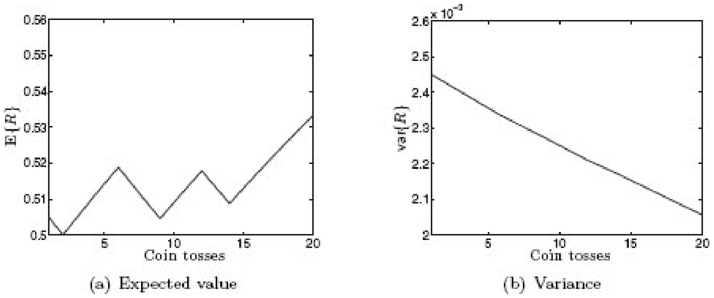
(c) $\delta = 54, \gamma = 51$

(d) $\delta = 56, \gamma = 54$

(e) $\delta = 59, \gamma = 56$

(f) $\delta = 64, \gamma = 56$

FIGURE 3.10 Evolution of the posterior $p(r|y_N)$ as more coin tosses are observed for the fair coin scenario. The dashed line shows the prior density.



(a) Expected value

(b) Variance

FIGURE 3.11 Evolution of $\mathbf{E}_{p(r|y_N)}\{R\}$ (a) and var{R} (b) as the 20 coin tosses are observed for the fair coin scenario.

$$\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6\middle|r)\right\} = \int_{r=0}^{r=1} P(Y_{new} \leq 6|r)p(r|y_N)dr.$$

If $p(r|y_N)$ is zero everywhere except at one specific value (say $\hat{r}$), this becomes

$$\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} \leq 6|r)\} = P(Y_{\text{new}} \leq 6|\hat{r}).$$

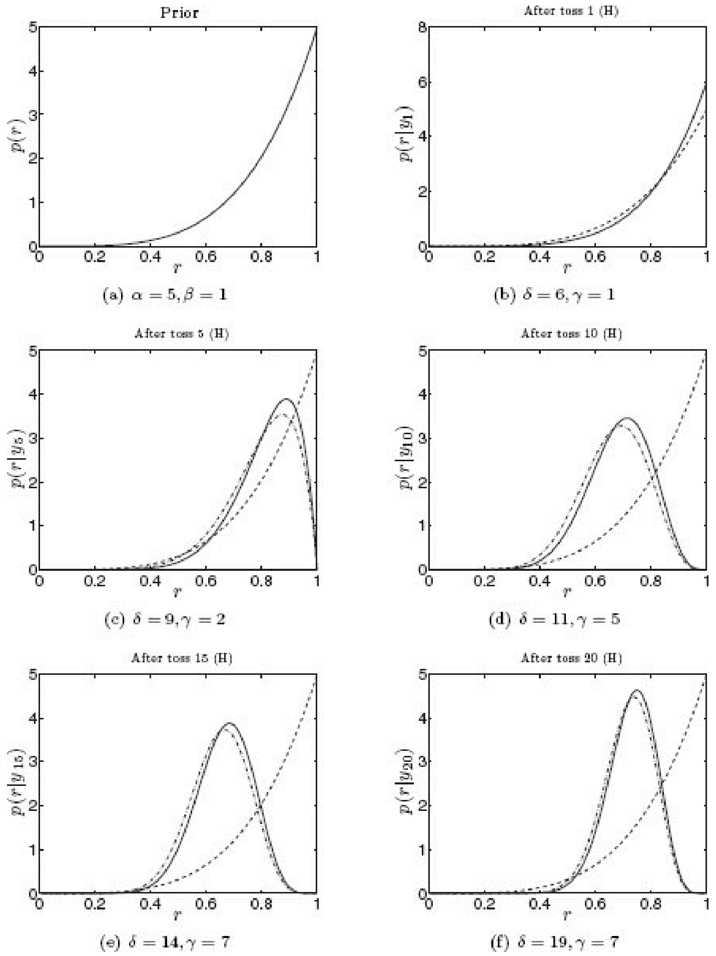
In other words, as the variance decreases, $P\ (Y_{\text{new}} \leq 6|\hat{\boldsymbol{r}})$ becomes a better and better approximation to the true expectation. This is not specific to this example – as the quantity of data increases (and uncertainty about parameters subsequently decreases), point approximations become more reliable.

3.3.3 A biased coin

In the final scenario we assume that the coin (and therefore the stall owner) is biased to generate more heads than tails (MATLAB script: **coin_scenario3.m**). This is encoded through a beta prior with parameters $\alpha = 5$, $\beta = 1$. The expected value is

$$\mathbf{E}_{p(r)}\left\{r\right\} = 5/6,$$

five coins out of every six will come up heads. Just as for scenario 2, Figure 3.12(a) shows the prior density and Figures 3.12(b), 3.12(c), 3.12(d), 3.12(e) and 3.12(f) show the posterior after 1, 5, 10, 15 and 20 tosses, respectively. Given what we've already seen, there is nothing unusual here. The posterior moves quite rapidly away from the prior (the prior effectively has only the influence of $\alpha + \beta = 6$ data points). Figure 3.13 shows the evolution of expected value and variance. The variance curve has several bumps corresponding to tosses resulting in tails. This is because of the strong prior bias towards a high r value. We don't expect to see many tails under this assumption and so when we do, the model becomes less certain. Once again, we calculate the true quantity of interest, $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\mathbf{new}} \leq 6|r)\right\}$. The final posterior parameter values are $\delta = \alpha + y_N = 5 + 14 = 19$, $\gamma = 1 + N - y_N = 1 + 20 - 14 = 7$. Plugging these in,



(a) $\alpha = 5, \beta = 1$

(b) $\delta = 6, \gamma = 1$

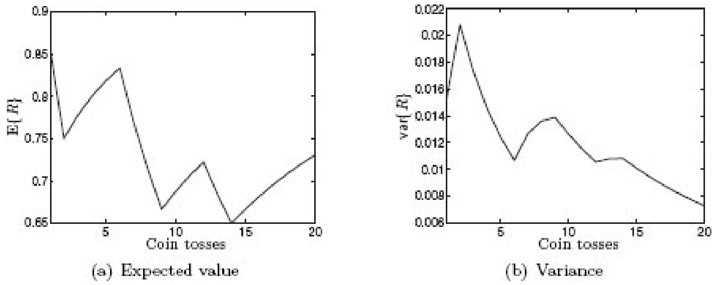
(c) $\delta = 9, \gamma = 2$

(d) $\delta = 11, \gamma = 5$

(e) $\delta = 14, \gamma = 7$

(f) $\delta = 19, \gamma = 7$

FIGURE 3.12 Evolution of the posterior $p(r|y_N)$ as more coin tosses are observed for the biased coin scenario. The dashed line shows the prior density and in the last four plots, the dash-dot line shows the previous posterior (i.e. the posterior after 4, 9, 14 and 19 tosses).



(a) Expected value

(b) Variance

FIGURE 3.13 Evolution of $\mathbf{E}_{p(r|y_N)}\{R\}$ (a) and var{*R*} (b) as the 20 coin tosses are observed for the biased coin scenario.

$$\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.2915.$$

The approximation, noting that $\hat{r}$ = 19/(19 + 7) = 0.7308 is

$$P(Y_{\text{new}} \leq 6|\hat{r}) = 0.2707$$

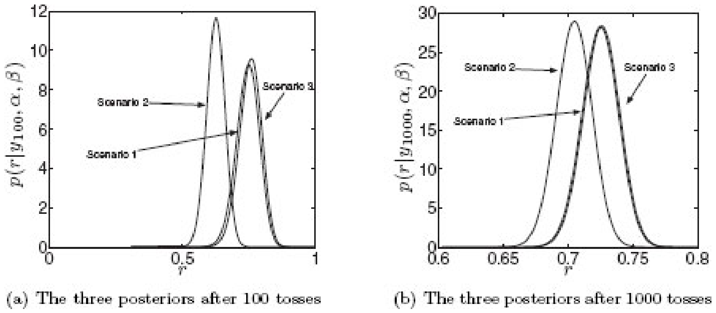
Both values suggest we will lose money on average.

3.3.4 The three scenarios – a summary

Our three different scenarios have given us different values for the expected probability of winning:

1. No prior knowledge: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.4045$.
2. Fair coin: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.7579.$
3. Biased coin: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.2915.$

Which one should we choose? We could choose based on which of the prior beliefs seems most plausible. Given that the stall holder doesn't look like he is about to go out of business, scenario 3 might be sensible. We might decide that we really do not know anything about the stall holder and coin and look to scenario 1. We might believe that an upstanding stall holder would never stoop to cheating and go for scenario 2. It is possible to justify any of them. What we have seen is that the Bayesian technique allows you to combine the data observed (20 coin tosses) with some prior knowledge (one of the scenarios) in a principled way. The posterior density explicitly models the uncertainty that remains in *r* at each stage and can be used to make predictions (see Exercises 3.7 and 3.8).



(a) The three posteriors after 100 tosses

(b) The three posteriors after 1000 tosses

FIGURE 3.14 The posterior densities for the three scenarios after 100 coin tosses (left) and 1000 coin tosses (right).

3.3.5 Adding more data

Before we move on, it is worth examining the effect of adding more and more data. We have seen in each of our scenarios that the addition of more data results in the posterior diverging from the prior – usually through a decrease in variance. In fact, if we continue adding more data, we will find that the posteriors for all three scenarios start to look very similar. In Figure 3.14 we see the posteriors for the three scenarios after 100 and 1000 tosses. Compared with the posteriors for the three scenarios after small numbers of tosses have been observed (Figures 3.6(f), 3.10(d) and 3.12(d)), we notice that the posteriors are becoming more and more similar. This is particularly noticeable for scenarios 1 and 3 – by 1000 tosses they are indistinguishable. The difference between these two and the posteriors for scenario 2 is due to the high strength (low variance) of the prior for scenario 2 – the prior corresponds to a very strong belief and it will take a lot of contradictory data to remove that influence.

The diminishing effect of the prior as the quantity of data increases is easily explained if we look at the expression used to compute the posterior. Ignoring the normalising marginal likelihood term, the posterior is proportional to the likelihood multiplied by the prior. As we add more data, the prior is unchanged but the likelihood becomes a product (if the normal independence assumptions are made) of individual likelihood for more and more observations. This increase will gradually swamp the single contribution from the prior. It is also very intuitive – as we observe more and more data, beliefs we had before seeing any become less and less important.

3.4 MARGINAL LIKELIHOODS

Fortunately, subjective beliefs are not the only option for determining which of our three scenarios is best. Earlier in this chapter, when discussing the terms in Equation 3.3, we showed how the denominator $p(y_N)$ could be considered to be related to r as follows:

$$\begin{aligned} p(y_N) &= \int_{r=0}^{r=1} p(r, y_N) \\ &= \int_{r=0}^{r=1} p(y_N|r)p(r)dr. \end{aligned} \tag{3.12}$$

Now when considering different choices of $p(r)$, we need to be more strict about our conditioning. $p(r)$ should actually be written as $p(r|\alpha, \beta)$ as the density is conditioned on a particular pair of α and β values. Extending this conditioning through Equation 3.12 gives

$$p\left(y_N \middle| \alpha, \beta\right) = \int_{r=0}^{r=1} p(y_N|r)p\left(r \middle| \alpha, \beta\right)dr. \tag{3.13}$$

The marginal likelihood (so called because r has been marginalised), $p(y_N\ |\alpha, \beta)$, is a very useful and important quantity. It tells us how likely the data (y_N) is given our choice of prior parameters α and β. The higher $p(y_N\ |\alpha, \beta)$, the better our evidence agrees with the prior specification. Hence, for our dataset, we could use $p(y_N\ |\alpha, \beta)$ to help choose the best scenario: select the scenario for which $p(y_N\ |\alpha, \beta)$ is highest.

To compute this quantity, we need to evaluate the following integral:

$$\begin{aligned} p(y_N|\alpha, \beta) &= \int_{r=0}^{r=1} p(y_N|r)p\left(r \middle| \alpha, \beta\right)dr \\ &= \int_{r=0}^{r=1} \binom{N}{y_N} r^{y_N}(1-r)^{N-y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}dr \\ &= \binom{N}{y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{r=0}^{r=1} r^{\alpha+y_N-1}(1-r)^{\beta+N-y_N-1}dr. \end{aligned}$$

This is of exactly the same form as Equation 3.10. The argument inside the integral is an unnormalised beta density and so we know that by integrating it we will get the inverse of the normal beta normalising constant. Therefore,

$$p\left(y_N \middle| \alpha, \beta\right) = \binom{N}{y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)}{\Gamma(\alpha+\beta+N)}. \tag{3.14}$$

In our example, $N = 20$ and $y_N = 14$ (there were a total of 14 heads in the 2 sets of 10 tosses). We have three different possible pairs of α and β values. Plugging these values into Equation 3.14 gives

1. No prior knowledge, $\alpha = \beta = 1$, $p(y_N\ |\alpha, \beta) = 0.0476$.
2. Fair coin, $\alpha = \beta = 50$, $p(y_N\ |\alpha, \beta) = 0.0441$.
3. Biased coin, $\alpha = 5$, $\beta = 1$, $p(y_N\ |\alpha, \beta) = 0.0576$.

The prior corresponding to the biased coin has the highest marginal likelihood and the fair coin prior has the lowest. In the previous section we saw that the probability of winning under that scenario was $\mathbf{E}_{p(r|y_N,\alpha,\beta)}\left\{P\left(Y_{\text{new}} \leq 6 \middle| r\right)\right\} = 0.2915.$ (note that we're now conditioning the posterior on the prior parameters – $p(r|y_N, \alpha, \beta)$).

A word of caution is required here. Choosing priors in this way is essentially choosing the prior that best agrees with the data. The prior no longer corresponds to our beliefs *before* we observe any data. In some applications this may be unacceptable. What it does give us is a single value that tells us how much the data backs up the prior beliefs. In the above example, the data suggests that the biased coin prior is best supported by the evidence.

3.4.1 Model comparison with the marginal likelihood

It is possible to extend the prior comparison in the previous section to using the marginal likelihood to optimise α and β. Assuming that α and β can take any value in the ranges

$$\begin{aligned} 0 \leq \ &\alpha \ \leq 50 \\ 0 \leq \ &\beta \ \leq 30, \end{aligned}$$

we can search for the values of *α* and *β* that maximise $p(y_N \mid \alpha, \beta)$.

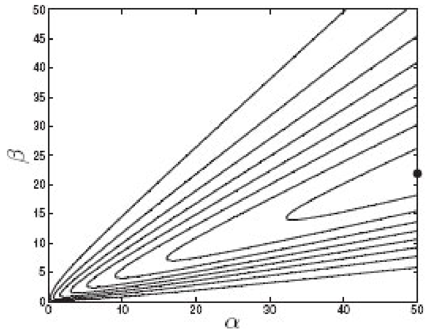


FIGURE 3.15 Marginal likelihood contours (as a function of the prior parameters, *α* and *β*) for the coin example. The circle towards the top right shows the optimum.

Figure 3.15 shows the marginal likelihood as *α* and *β* are varied in their respective ranges. The optimum value is *α* = 50, *β* = 22, resulting in a marginal likelihood of 0.1694. Choosing parameters in this way is known as Type II Maximum Likelihood (to distinguish it from standard (i.e. Type I) Maximum Likelihood, introduced in Chapter 2).

3.5 HYPERPARAMETERS

The Bayesian analysis presented thus far has all been based on the idea that we can represent any quantities of interest as random variables (e.g. *r*, the probability of a coin landing heads). *r* is not the only parameter of interest in our example. *α* and *β* are also parameters – could we do the same thing with them? In some cases we can be directed towards particular values based on our knowledge of the problem (we might know that the coin is biased). Often we will not know the exact value that they should take and should therefore treat them as random variables. To do so, we need to define a prior density over all random variables – $p(r, \alpha, \beta)$. This factorises as (see Section 2.2.5)

$$p(r, \alpha, \beta) = p(r|\alpha, \beta)p(\alpha, \beta).$$

In addition, it will often be useful to assume that *α* and *β* are independent: $p(\alpha, \beta) = p(\alpha)p(\beta)$. The quantity in which we are interested is the posterior over all parameters in the model:

$$p(r, \alpha, \beta|y_N).$$

Applying Bayes' rule, we have

$$\begin{aligned} p(r, \alpha, \beta|y_N) &= \frac{p(y_N|r,\alpha,\beta)p(r,\alpha,\beta)}{p(y_N)} \\ &= \frac{p(y_N|r)p(r,\alpha,\beta)}{p(y_N)} \\ &= \frac{p(y_N|r)p(r|\alpha,\beta)p(\alpha,\beta)}{p(y_N)}. \end{aligned}$$

Note that, in the second step, we removed *α* and *β* from the likelihood $p(y_N \mid r)$. This is another example of conditional independence (see Section 2.8.1). The distribution over y_N depends on *α* and *β* but only through their influence on *r*. Conditioned on a particular value of *r*, this dependence is broken.

$p(\alpha, \beta)$ will normally require some additional parameters – i.e. $p(\alpha, \beta|\kappa)$ where *κ* controls the density in the same way that *α* and *β* control the density for *r*. *κ* is known as a **hyper-parameter** because it is a parameter controlling the prior on the parameters controlling the prior on *r*. When computing the marginal likelihood, we integrate over all random variables and are just left with the data conditioned on the hyperparameters:

$$p(y_N \Big| \kappa) = \iiint p(y_N|r)p(r|\alpha, \beta)p(\alpha, \beta|\kappa)dr\, d\alpha\, d\beta.$$

Unfortunately, adding this extra complexity to the model often means that computation of the quantities of interest – the posterior $p(r, \alpha, \beta|y_N, \kappa)$ (and any predictive expectations) and the marginal likelihood $p(y_N \mid \kappa)$ – is analytically intractable and requires one of the approximation methods that we will introduce in Chapter 4.

At this point, one could imagine indefinitely adding layers to the model. For example, *κ* could be thought of as a random variable that comes from a density parameterised by other random variables. The number of levels in the hierarchy (how far we go before we fix one or more parameters) will be dictated by the data we are trying to model (perhaps we can specify exact values at some level) or how much computation we can tolerate. In general, the more layers we add the more complex it will be to compute posteriors and predictions.

3.6 GRAPHICAL MODELS

When adding extra layers to our model (hyperparameters, etc.), they can quickly become unwieldy. It is popular to describe them graphically. A **graphical model** is a network where nodes correspond to random variables and edges to dependencies between random variables. For example, in Section 2.2.4 we introduced various properties of random variables through a model that consisted of two random variables – one representing the toss of a coin (X) and one representing how I say the coin landed (Y). The model is defined through the conditional distribution $P(Y = y|X = x)$ and is represented graphically in Figure 3.16(a). The two nodes are joined by an aarrow to show that Y is defined as being conditioned on X. Note also that the node for Y is shaded. This is because, as far as the listener is concerned, this variable is *observed*. The listener does not see the coin actually landing and so doesn't observe X. Imagine that the procedure was repeated N times; we now have $2N$ random variables, X_1, ..., X_N and Y_1, ..., Y_N. Drawing all of these would be messy. Instead we can embed the nodes within a **plate**. Plates are rectangles that tell us that whatever is embedded within them is repeated a number of times. The number of times is given in the bottom right corner, as shown in Figure 3.16(b).

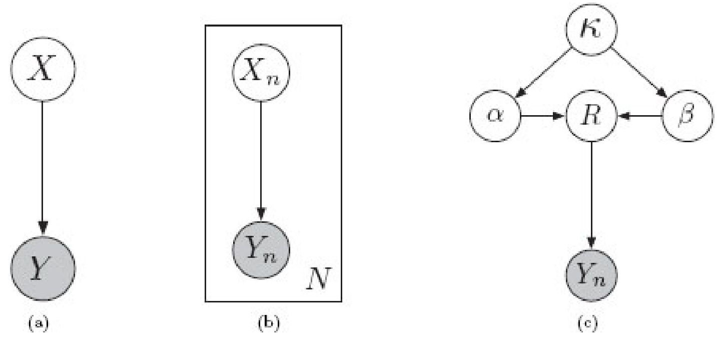


FIGURE 3.16 Graphical model examples. Nodes correspond to random variables, with the shaded nodes corresponding to things that we observe. Arrows describe the dependencies between variables and the plates describe multiple instances. For example, in (b), there are N random variables Y_n ($n = 1,..., N$) and each is dependent on a random variable X_n. (c) is a graphical representation of the model used in the coin example with the addition of a prior on α and β parameterised by κ.

Figure 3.16(c) shows a graphical representation of our coin toss model. It has a single (observed) random variable that represents the number of heads in N tosses, y_N. This is conditioned on a random variable R, which depends on random variables α and β. Finally, α and β are dependent on the hyper-parameter κ.

More information on graphical models can be found in the suggested reading at the end of the chapter.

3.7 SUMMARY

In the previous sections we have introduced many new concepts. Perhaps the most important is the idea of treating all quantities of interest as random variables. To do this we must define a prior distribution over the possible values of these quantities and then use Bayes' rule (Equation 3.3) to see how the density changes as we incorporate evidence from observed data. The resulting posterior density can be examined and used to compute interesting expectations. In addition, we have shown how the marginal likelihood (the normalisation constant in Bayes' rule) can be used to compare different models – for example, choosing the most likely prior in our coin tossing example – and discussed the possible pitfalls and objections to such an approach. Finally, we have shown how the Bayesian method can be extended by treating parameters that define the priors over other parameters as random variables. Additions to the hierarchy such as this often make analytical computations intractable and we have to resort to sampling and approximation based techniques, which are the subject of the next chapter.

3.8 A BAYESIAN TREATMENT OF THE OLYMPIC 100 m DATA

We now return to the Olympic 100m data. In the previous chapters we fitted a linear (in the parameters) model by minimising the squared loss and then incorporated an explicit noise model and found optimal parameter values by maximising the likelihood. In this section, we will give the data a Bayesian treatment with the aim of making a prediction for the 2012 Olympics in London. This will involve several steps. Firstly, we will need to define the prior and likelihood (as we did in the coin example) and use these to compute the posterior density over the parameters of our model, just as we computed the posterior over r in the coin example. Once we've computed the posterior, we can use it to make predictions for new Olympic years.

3.8.1 The model

We will use the kth order polynomial model that was introduced in Chapter 1 with the Gaussian noise model introduced in Chapter 2:

$$t_n = w_0 + w_1 x_n + w_2 x_n^2 + \cdots + w_K x_n^K + \epsilon_n,$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$. In vector form, this corresponds to

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

where **w** = [w_0, ..., w_K]$^\mathrm{T}$ and $\mathbf{x}_n = [1, x_n, x_n^2, \dots, x_n^K]^\mathbf{T}$. Stacking all of the responses into one vector **t** = [t_1, ..., t_N]$^\mathrm{T}$ and all of the inputs into a single matrix, **X** = [$\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$]$^\mathrm{T}$ (just as in Equation 1.18), we get the following expression for the whole dataset:

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \epsilon,$$

where = [$_1$, ..., $_N$]$^\mathrm{T}$.

In this example, we are going to slightly simplify matters by assuming that we know the true value of σ^2. We could use all of the methods introduced in this chapter to treat σ^2 as a random variable and we could get analytical results for the posterior distribution but the maths is messier, which could detract from the main message. Substituting these various symbols into Bayes' rule gives

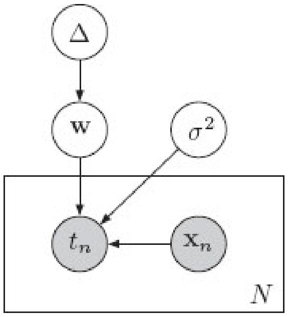


FIGURE 3.17 Graphical model for the Bayesian model of the Olympic men's 100m data.

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2,\Delta)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X},\sigma^2,\Delta)} \\ &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X},\sigma^2,\Delta)} \end{aligned}$$

where Δ corresponds to some set of parameters required to define the prior over **w** that will be defined more precisely below. The graphical model can be seen in Figure 3.17. Expanding the marginal likelihood we have

$$p(\mathbf{w}\Big|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)}{\int p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)\ d\mathbf{w}}. \tag{3.15}$$

We are interested in making predictions which will involve taking an expectation with respect to this posterior density. In particular, for a set of attributes $\mathbf{x}_{\text{new}}$ corresponding to a new Olympic year, the density over the associated winning time t_{new} is given by

$$p(t_{\text{new}}\Big|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2, \Delta) = \int p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)d\mathbf{w}. \tag{3.16}$$

Notice again the conditioning on the right hand side. The posterior density of **w** does not depend on $\mathbf{x}_{\text{new}}$ and so it does not appear in the conditioning. Similarly, when we make predictions, we will not be using Δ and so it doesn't appear in $p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2)$. Predictions could also take the form of probabilities. For example, we could compute the probability that the winning time will be under 9.5 seconds:

$$P(t_{\text{new}} < 9.5\Big|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2, \Delta) = \int P(t_{\text{new}} > 9.5|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)d\mathbf{w}. \tag{3.17}$$

3.8.2 The likelihood

The likelihood $p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)$ is exactly the quantity that we maximised in the previous chapter. Our model tells us that

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_N)$. This is a Gaussian random variable ( ) plus a constant. We showed in Section 2.8 that this is equivalent to the Gaussian random variable with the constant added to the mean. This gives us our likelihood

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

an *N*-dimensional Gaussian density with mean **Xw** and variance $\sigma^2\mathbf{I}_N$. The analogous expression in the coin example is the binomial likelihood given in Equation 3.2.

3.8.3 The prior

Because we are interested in being able to produce an exact expression for our posterior, we need to choose a prior, $p(\mathbf{w}|\Delta)$, that is conjugate to the Gaussian likelihood. Conveniently, a Gaussian prior is conjugate to a Gaussian likelihood. Therefore, we will use a Gaussian prior for **w**. In particular,

$$p(\mathbf{w}|\mu_0, \textstyle\sum_0) = \mathcal{N}(\mu_0, \textstyle\sum_0),$$

where we will choose the parameters μ_0 and $\mathbf{\Sigma}_0$ later. This is analogous to Equation 3.4 in the coin example. From now on we will not always explicitly condition on μ_0 and $\mathbf{\Sigma}_0$ in our expressions. For example, for brevity, instead of writing $p(\mathbf{w}\,|\,\mathbf{t}, \mathbf{X}, \sigma^2, \mu_0, \mathbf{\Sigma}_0)$ we will use $p(\mathbf{w}\,|\,\mathbf{t}, \mathbf{X}, \sigma^2)$ (see Exercise 3.10).

3.8.4 The posterior

We now turn our attention to computing the posterior. As in the coin example, we will use the fact that we *know* that the posterior will be Gaussian. This allows us to ignore the marginal likelihood in Equation 3.15 and just manipulate the likelihood and prior until we find something that is proportional to a Gaussian. As a first step, we can collect the terms in **w** together and ignore any term that does not include **w**:

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) &\propto p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\boldsymbol{\mu}_0, \textstyle\sum_0) \\
&= \frac{1}{(2\pi)^{N/2}|\sigma^2\mathbf{I}|^{1/2}}\exp\left(-\tfrac{1}{2}(\mathbf{t}-\mathbf{Xw})^{\mathrm{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{t}-\mathbf{Xw})\right) \times \frac{1}{(2\pi)^{N/2}|\sum_0|^{1/2}}\exp\left(-\tfrac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathrm{T}}\textstyle\sum_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right) \\
&\propto \exp\left(-\tfrac{1}{2\sigma^2}(\mathbf{t}-\mathbf{Xw})^{\mathrm{T}}(\mathbf{t}-\mathbf{Xw})\right) \times \exp\left(-\tfrac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathrm{T}}\textstyle\sum_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right) \\
&= \exp\left\{-\tfrac{1}{2}\left(\tfrac{1}{\sigma^2}(\mathbf{t}-\mathbf{Xw})^{\mathrm{T}}(\mathbf{t}-\mathbf{Xw}) + (\mathbf{w}-\boldsymbol{\mu}_0)^{\mathrm{T}}\textstyle\sum_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right)\right\}.
\end{aligned}$$

Multiplying the terms in the bracket out and once again removing any that don't involve **w** gives

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) \propto \exp\left\{-\tfrac{1}{2}\left(-\tfrac{2}{\sigma^2}\mathbf{t}^{\mathrm{T}}\mathbf{Xw} + \tfrac{1}{\sigma^2}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Xw} + \mathbf{w}^{\mathrm{T}}\textstyle\sum_0^{-1}\mathbf{w} - 2\boldsymbol{\mu}_0^{\mathrm{T}}\textstyle\sum_0^{-1}\mathbf{w}\right)\right\}.$$

We know that the posterior will be Gaussian. Therefore we can remove the constants (i.e. terms not involving **w**) and rearrange an expression for a multivariate Gaussian to make it look something like the expression we have above:

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}, \mathbf{X} < \sigma^2) &= \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}, \textstyle\sum_{\mathbf{w}}) \\
&\propto \exp\left(-\tfrac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})^{\mathrm{T}}\textstyle\sum_{\mathbf{w}}^{-1}(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})\right) \\
&\propto \exp\left\{-\tfrac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\textstyle\sum_{\mathbf{w}}^{-1}\mathbf{w} - 2\boldsymbol{\mu}_{\mathbf{w}}^{\mathrm{T}}\textstyle\sum_{\mathbf{w}}^{-1}\mathbf{w}\right)\right\}.
\end{aligned} \tag{3.18}$$

The terms linear and quadratic in **w** in Equation 3.8.4 must be equal to those in Equation 3.18. Taking the quadratic terms, we can solve for $\mathbf{\Sigma}_{\mathrm{W}}$:

$$\begin{aligned}
\mathbf{w}^{\mathrm{T}}\textstyle\sum_{\mathbf{w}}^{\text{-}1}\mathbf{w} &= \tfrac{1}{\sigma^2}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Xw} + \mathbf{w}^{\mathrm{T}}\textstyle\sum_0^{-1}\mathbf{w} \\
&= \mathbf{w}^{\mathrm{T}}\left(\tfrac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \textstyle\sum_0^{-1}\right)\mathbf{w}
\end{aligned}$$

$$\textstyle\sum_{\mathbf{w}} = \left(\tfrac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \textstyle\sum_0^{-1}\right)^{-1}.$$

Similarly, equating the linear terms from Equations 3.8.4 and 3.18 (and using our new expression for $\mathbf{\Sigma}_{\mathrm{W}}$) we can get an expression for μ_{W}:

$$
\begin{aligned}
-2\boldsymbol{\mu}_{\mathbf{w}}^{\mathrm{T}} \textstyle\sum_{\mathbf{w}}^{-1} \mathbf{w} &= -\tfrac{1}{\sigma^2}\mathbf{t}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\boldsymbol{\mu}_{0}^{\mathrm{T}} \textstyle\sum_{0}^{-1} \mathbf{w} \\
\boldsymbol{\mu}_{\mathbf{w}}^{\mathrm{T}} \textstyle\sum_{\mathbf{w}}^{-1} \mathbf{w} &= \tfrac{1}{\sigma^2}\mathbf{t}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \boldsymbol{\mu}_{0}^{\mathrm{T}} \textstyle\sum_{0}^{-1} \mathbf{w} \\
\boldsymbol{\mu}_{\mathbf{w}}^{\mathrm{T}} \textstyle\sum_{\mathbf{w}}^{-1} &= \tfrac{1}{\sigma^2}\mathbf{t}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\mu}_{0}^{\mathrm{T}} \textstyle\sum_{0}^{-1} \\
\boldsymbol{\mu}_{\mathbf{w}}^{\mathrm{T}} \textstyle\sum_{\mathbf{w}}^{-1} \textstyle\sum_{\mathbf{w}} &= \left(\tfrac{1}{\sigma^2}\mathbf{t}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\mu}_{0}^{\mathrm{T}} \textstyle\sum_{0}^{-1}\right) \textstyle\sum_{\mathbf{w}} \\
\boldsymbol{\mu}_{\mathbf{w}}^{\mathrm{T}} &= \left(\tfrac{1}{\sigma^2}\mathbf{t}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\mu}_{0}^{\mathrm{T}} \textstyle\sum_{0}^{-1}\right) \textstyle\sum_{\mathbf{w}}
\end{aligned}
$$

$$
\boldsymbol{\mu}_{\mathbf{w}} = \textstyle\sum_{\mathbf{w}} \left(\tfrac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{t} + \textstyle\sum_{0}^{-1} \boldsymbol{\mu}_{0}\right), \tag{3.19}
$$

because $\boldsymbol{\Sigma}_{\mathbf{w}}^{\mathrm{T}} = \mathit{\Sigma}_{\mathbf{w}}$ due to the fact that it must be symmetric. Therefore,

$$
p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}, \textstyle\sum_{\mathbf{w}}) \tag{3.20}
$$

where

$$
\textstyle\sum_{\mathbf{w}} = \left(\tfrac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \textstyle\sum_{0}^{-1}\right)^{-1} \tag{3.21}
$$

$$
\boldsymbol{\mu}_{\mathbf{w}} = \textstyle\sum_{\mathbf{w}} \left(\tfrac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{t} + \textstyle\sum_{0}^{-1} \boldsymbol{\mu}_{0}\right) \tag{3.22}
$$

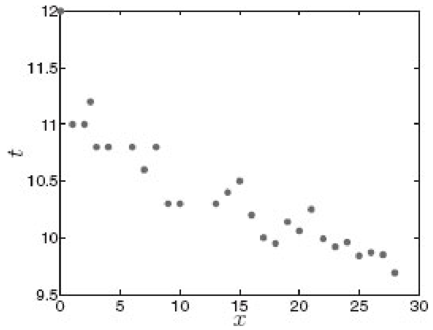


FIGURE 3.18 Olympic data with rescaled *x* values.

(see Exercise 3.12). These expressions do not look too far away from things we have seen before. In particular, compare Equation 3.22 with the regularised least squares solution given in Equation 1.21. In fact, if μ_0 = [0, 0, ..., 0]$^\mathrm{T}$, the expressions are almost identical. Given that the posterior is a Gaussian, the single most likely value of **w** is the mean of the posterior, μ_W. This is known as the **maximum a posteriori** (MAP) estimate of **w** and can also be thought of as the maximum value of the joint density *p*(**w, t**|**X**, σ^2, Δ) (the likelihood multiplied by the prior). We have already seen that the squared loss considered in Chapter 1 is very similar to a Gaussian likelihood and it follows from this that computing the most likely posterior value (when the likelihood is Gaussian) is equivalent to using regularised least squares (see Exercise 3.9). This comparison can often help to provide intuition regarding the effect of the prior.

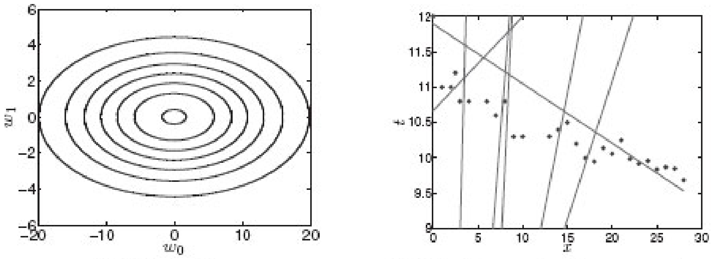
3.8.5 A first-order polynomial

We will illustrate the prior and posterior with a first-order polynomial, as it is possible to visualise densities in the two-dimensional parameter space. The input vectors also have two elements, $\mathbf{x}_n$ = [1, x_n]$^\mathrm{T}$. To aid visualisation, we will rescale the Olympic year by subtracting the year of the first Olympics (1896) from each year and then dividing each number by 4. This means that x_1 is now 0, x_2 is 1, etc. The data with this new *x* scaling is plotted in Figure 3.18.

Returning to the fairground, the first step in our analysis is the choice of prior parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$. For $\boldsymbol{\mu}_0$, we will assume that we don't really know anything about what the parameters should be and choose $\boldsymbol{\mu}_0$ = [0, 0]$^\mathrm{T}$. For the covariance, we will use

$$
\textstyle\sum_0 = \begin{bmatrix} 100 & 0 \\ 0 & 5 \end{bmatrix}.
$$

The larger value for the variance of w_0 is due to the fact that we saw in the maximum likelihood estimate that the optimal value of w_0 was much higher than that for w_1. We have also assumed that the two variables are independent in the prior by setting the off-diagonal elements in the covariance matrix to zero. This does not preclude them from being dependent in the posterior. The contours of this prior density can be seen in Figure 3.19(a). It's hard to visualise what this means in terms of the model. To help, in Figure 3.19(b) we have shown functions corresponding to several sets of parameters drawn from this prior. To create these, we sampled **w** from the Gaussian defined by $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ and then substituted these into our linear model – $t_n = w_0 + w_1 x_n$. The examples show that the prior admits the possibility of many very different models.



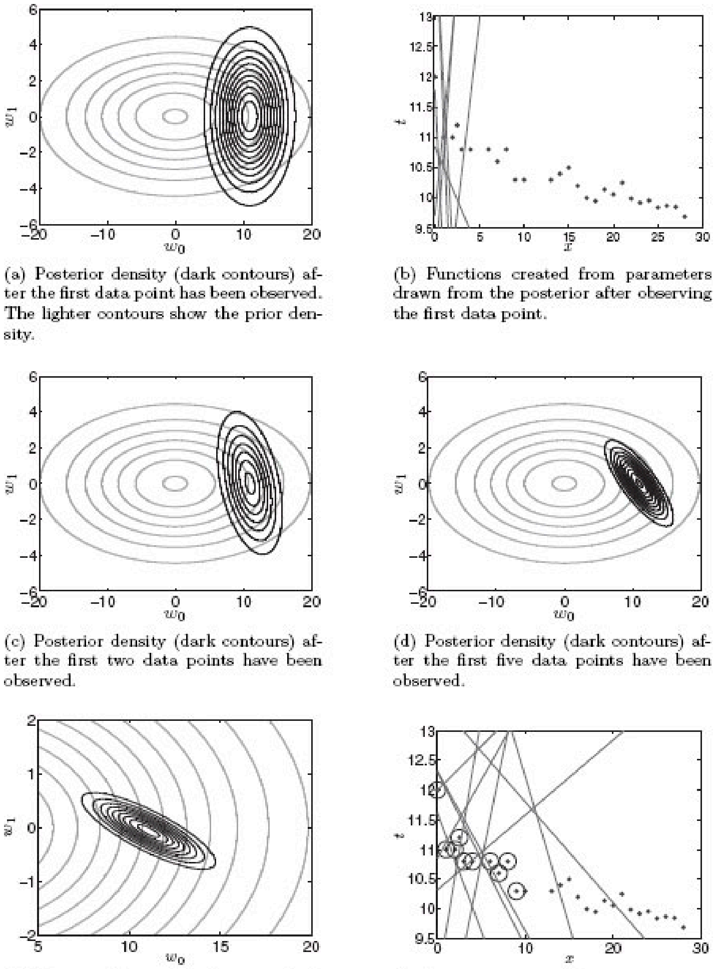
(a) Prior density.

(b) Functions created from parameters drawn from prior.

FIGURE 3.19 Gaussian prior used for the Olympic 100m data (a) and some functions created with samples drawn from the prior (b).

Using $\sigma^2 = 10$ for illustrative purposes (MATLAB script: **olympbayes.m**), we can now compute the posterior distribution when we observe one data point. Using the data point corresponding to the first Olympics, our data is summarised as $\mathbf{x} = [1, 0]^T$, $\mathbf{X} = [1, 0]$, $\mathbf{t} = [12]$. Plugging these values along with our prior parameters and $\sigma^2 = 10$ into Equations 3.20–3.22, we obtain the posterior distribution shown in Figure 3.20(a). The posterior now has much more certainty regarding w_0 but still knows very little about w_1. This makes sense – we've provided a data point at $x = 0$ so this should be highly informative in determining the intercept but tells us very little about the gradient (one data point alone could never tell us much about the gradient). Some functions created with samples from this posterior are shown in Figure 3.20(b). They look quite different from those from the prior – in particular, they all pass quite close to our first data point.

Figures 3.20(c), 3.20(d) and 3.20(e) show the evolution of the posterior after 2, 5 and 10 data points, respectively. Just as in the coin example, we notice that the posterior becomes more condensed (we are becoming more certain about the value of **w**). Also, as it evolves, the posterior begins to tilt. This is indicative of a dependence developing between the two parameters – if we increase the intercept w_0, we must decrease the gradient. Recall that, in the prior, we assumed that the two parameters were independent ($\boldsymbol{\Sigma}_0$ only had non-zero values on the diagonal) so this dependence is coming entirely from the evidence within the data. To help visualise what the posterior means at this stage, Figure 3.20(f) shows a set of functions made from parameters drawn from the posterior. When compared with Figure 3.20(b), we see that the posterior density is beginning to favour parameters that correspond to models suited to our data. Finally, in Figure 3.21(a) we see the posterior after all 27 data points have been included and in Figure 3.21(b) we see functions drawn from this posterior. The functions are really now beginning to follow the trend in our data. There is still a lot of variability though. This is due to the relatively high value of $\sigma^2 = 10$ that we chose to help visualise the prior and posteriors. For making predictions, we might want to use a more realistic value. In Figure 3.22(a) we show the posterior after all data has been observed for $\sigma^2 = 0.05$ (this is roughly the maximum likelihood value we obtained in Section 2.8.2). The posterior is now far more condensed – very little variability remains in **w**, as can be seen by the homogeneity of the set of functions drawn in Figure 3.22(b). We will now turn our attention to making predictions.



(a) Posterior density (dark contours) after the first data point has been observed. The lighter contours show the prior density.

(b) Functions created from parameters drawn from the posterior after observing the first data point.

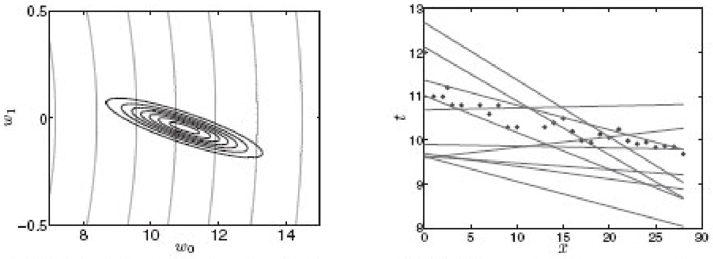
(c) Posterior density (dark contours) after the first two data points have been observed.

(d) Posterior density (dark contours) after the first five data points have been observed.

(e) Posterior density (dark contours) after the first ten data points have been observed. (Note that we have zoomed in.)

(f) Functions created from parameters drawn from the posterior after observing the first ten (highlighted) data points.

FIGURE 3.20 Evolution of the posterior density and example functions drawn from the posterior for the Olympic data as observations are added.



(a) Posterior density (dark contours) after all datapoints have been observed. The lighter contours show the prior density. (Note that we have zoomed in.)

(b) Functions created from parameters drawn from the posterior after observing all data points.

FIGURE 3.21 Posterior density (a) and sampled functions (b) for the Olympic data when all 27 data points have been added.

3.8.6 Making predictions

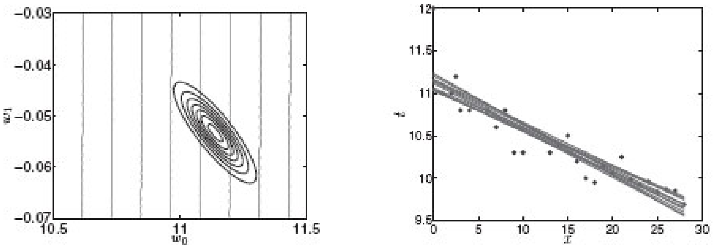
Given a new observation $\mathbf{x}_{\text{new}}$, we are interested in the density

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}},\mathbf{X},\mathbf{t},\sigma^2).$$

Notice that this is not conditioned on **w** – just as in the coin example, we are going to integrate out **w** by taking an expectation with respect to the posterior, $p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2)$. In particular, we need to compute

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

This is analogous to Equation 3.9 in the coin example.



(a) Posterior density (dark contours) after all data points have been observed. The lighter contours show the prior density. (Note that we have zoomed in.)

(b) Functions created from parameters drawn from the posterior after observing all data points.

FIGURE 3.22 Posterior density (a) and sampled functions (b) for the Olympic data when all 27 data points have been added with more realistic noise variance, $\sigma^2 = 0.05$.

$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2)$ is defined by our model as the product of $\mathbf{x}_{\text{new}}$ and $\mathbf{w}$ with some additive Gaussian noise:

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{w},\sigma^2).$$

Because this expression and the posterior are both Gaussian, the result of the expectation is another Gaussian. In general, if $p(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the expectation of another Gaussian density $(\mathcal{N}(\mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{w},\sigma^2))$ is given by

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}},\mathbf{X},\mathbf{t},\sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{w}},\sigma^2 + \mathbf{x}_{\text{new}}^{\mathrm{T}}\textstyle\sum_{\mathbf{w}}\mathbf{x}_{\text{new}}).$$

For the posterior shown in Figure 3.22(a), this is

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}},\mathbf{X},\mathbf{t},\sigma^2) = \mathcal{N}(9.5951,\ 0.0572)$$

and is plotted in Figure 3.23.

This density looks rather like the predictive densities we obtained from the maximum likelihood solution in Chapter 2. However, there is one crucial difference. With the maximum likelihood we chose one particular model: the one corresponding to the highest likelihood. To generate the density shown in Figure 3.23, we have averaged over all models that are consistent with our data and prior (we averaged over our posterior). Hence this density takes into account all uncertainty that remains in **w** given a particular prior and the data.

3.9 MARGINAL LIKELIHOOD FOR POLYNOMIAL MODEL ORDER SELECTION

In Section 1.5 we used a cross-validation procedure to select the order of polynomial to be used. The cross-validation procedure correctly identified that the dataset was generated from a third-order polynomial. In Section 3.4 we saw how the marginal likelihood could be used to choose prior densities. We will now see that it can also be used to choose models. In particular, we will use it to determine which order polynomial function to use for some synthetic data.

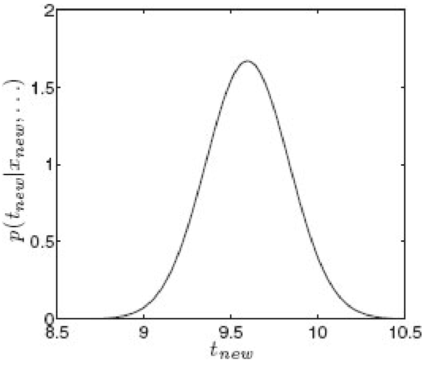


FIGURE 3.23 Predictive distribution for the winning time in the men’s 100m sprint at the 2012 London Olympics.

The marginal likelihood for our Gaussian model is defined as

$$p(\mathbf{t}\Big|\mathbf{X},\boldsymbol{\mu}_0,\textstyle\sum_0) = \displaystyle\int p(\mathbf{t}\Big|\mathbf{X},\mathbf{w},\sigma^2)p(\mathbf{w}\Big|\boldsymbol{\mu}_0,\textstyle\sum_0)d\mathbf{w}.$$

This is analogous to Equation 3.14 in the coin example. It is of the same form as the predictive density discussed in the previous section and is another Gaussian,

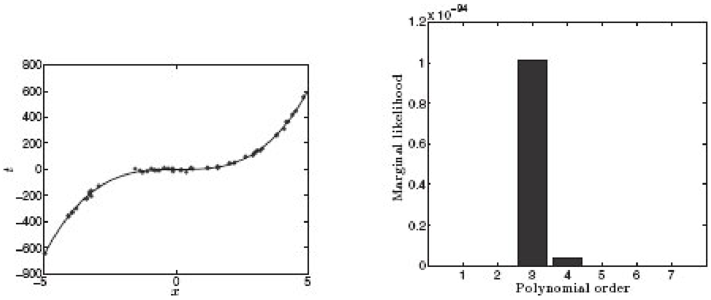
$$p(\mathbf{t}\Big|\mathbf{X},\boldsymbol{\mu}_0,\textstyle\sum_0) = N(\mathbf{X}\boldsymbol{\mu}_0,\sigma^2\mathbf{I}_N + \mathbf{X}\textstyle\sum_0\mathbf{X}^{\mathrm{T}}), \tag{3.23}$$

which we evaluate at **t** – the responses in the training set. Just as in Section 1.5, we will generate data from a noisy third-order polynomial and then compute the marginal likelihood for models from first to seventh-order. For each possible model, we will use a Gaussian prior on **w** with zero mean and an identity covariance matrix. For example, for the first-order model,

$$\boldsymbol{\mu}_0 = [0,0]^{\mathrm{T}}, \textstyle\sum_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and for the fourth-order model

$$\boldsymbol{\mu}_0 = [0,0,0,0,0]^{\mathrm{T}}, \textstyle\sum_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$



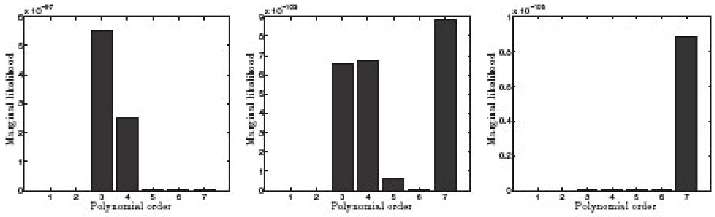
(a) Noisy data from a third-order polynomial.

(b) Marginal likelihood for models of different order.

FIGURE 3.24 Dataset sampled from the function $t = 5x^3 - x^2 + x$ (a) and marginal likelihoods for polynomials of increasing order (b).

The data and true polynomial are shown in Figure 3.24(a) (MATLAB script: **margpoly.m**). The true polynomial is $t = 5x^3 - x^2 + x$ and Gaussian noise has been added with mean zero and variance 150. The marginal likelihood for models from first to seventh order is calculated by plugging the relevant prior into Equation 3.23 and then evaluating this density at **t**, the observed responses. The values are shown in Figure 3.24(b). We can see that the marginal likelihood value is very sharply peaked at the true third-order model. The advantage of this over the cross-validation method is that, for this model, it is computationally undemanding (we don't have to fit several different datasets). We can also use all the data. However, as we have already mentioned, calculating the marginal likelihood is, in general, very difficult and we will often find it easier to resort to cross-validation techniques.

The marginal likelihood is conditioned on the prior parameters and so changing them will have an effect on the marginal likelihood values and possibly the highest scoring model. To show the effect of this, we can define $\mathbf{\Sigma}_0 = \boldsymbol{\sigma}_0^2\mathbf{I}$ and vary $\boldsymbol{\sigma}_0^2$. We have already seen the result for $\boldsymbol{\sigma}_0^2 = 1$. If we decrease $\boldsymbol{\sigma}_0^2$, we see higher-order models performing better. This can be seen in Figure 3.25. Decreasing $\boldsymbol{\sigma}_0^2$ from 1 to 0.3 results in the seventh-order polynomial becoming the most likely model. By decreasing $\boldsymbol{\sigma}_0^2$, we are saying that the parameters have to take smaller and smaller values. For a third order polynomial model to fit well, one of the parameters needs to be 5 (recall that $t = 5x^3 - x^2 + x$). As we decrease $\boldsymbol{\sigma}_0^2$, this becomes less and less likely, and higher-order models with lower parameter values become more likely. This emphasises the importance of understanding what we mean by a model. In this example, the model consists of the order of polynomial *and* the prior specification and we must be careful to choose the prior sensibly (see Exercise 3.11).



(a) $\sigma_0^2 = 0.7$

(b) $\sigma_0^2 = 0.4$

(c) $\sigma_0^2 = 0.3$

FIGURE 3.25 Marginal likelihoods for the third-order polynomial example with $\boldsymbol{\Sigma_0 = \sigma_0^2 \mathbf{I}}$ is $\boldsymbol{\sigma_0^2}$ decreased.