6 Parameter estimation: multiple parameters

In the previous chapter we looked at single parameter inference problems, as well as problems with conjugate priors that had analytic solutions. Now we move on to problems with multiple parameters, and will learn how to compute the joint, conditional, and marginal distributions. We will see how some well-known results concerning Gaussian distributions arise from the inference process.

6.1 Conditional and marginal distributions

When we have more than one parameter, the full posterior PDF becomes a multivariate distribution. Yet often we still want to find a one-dimensional distribution over one of its parameters. As introduced at the end of section 1.6.2, two such distributions are of particular interest: the conditional distribution and the marginal distribution.

Suppose a model has two parameters (a, b). If we fix one of the parameters, say $b = b_0$, then we can infer the one-dimensional posterior PDF for a, $P(a\,|\,b = b_0, D)$. From Bayes' theorem this is

$$P(a\,|\,b = b_0, D) \;=\; \frac{1}{Z_a} P(D\,|\,a, b = b_0) P(a\,|\,b = b_0) \tag{6.1}$$

where Z_a is a normalization constant. This is a *conditional posterior distribution* because it is conditioned on a fixed value of b. It corresponds to taking a slice through the full distribution at $b = b_0$. If the priors on a and b are independent then $P(a\,|\,b) = P(a)$.

If, on the other hand, b is not fixed and so must also be determined from the data, then we infer the two-dimensional distribution $P(a, b\,|\,D)$ from the data and the priors on a and b using Bayes' theorem

$$P(a, b\,|\,D) \;=\; \frac{1}{Z_{ab}} P(D\,|\,a, b) P(a, b) \tag{6.2}$$

where Z_{ab} is a normalization constant. If we then want to get the posterior of just a, we *marginalize* (integrate) over b to give

$$P(a\,|\,D) \;=\; \int P(a, b\,|\,D)\, db \tag{6.3}$$

$$\;=\; \frac{1}{Z_{ab}} \int P(D\,|\,a, b) P(a, b)\, db. \tag{6.4}$$

If the priors are independent then $P(a,b) = P(a)P(b)$. Marginalizing is like projecting the distribution along an axis.

Note the fundamental difference between $P(a\,|\,b,D)$ and $P(a\,|\,D)$. The former includes more given information, namely the value of b. Assuming that b determines the data to some degree (i.e. it is not an irrelevant parameter), then $P(a\,|\,b,D)$ must be narrower than $P(a\,|\,D)$, because by fixing b the data are used entirely to constrain a.

Marginalization is a powerful feature of probability analysis because it allows us to include parameters which are an essential part of the model, but which we may not actually be interested in. We marginalize over them to get the posterior PDF for the parameters of interest.

It is worth noting in this context the distinction between accuracy and precision (which I defined in section 2.7) in the context of posterior distributions. The width of the posterior PDF over a parameter is a measure of the *precision* of the estimate of that parameter. A narrower distribution means a higher precision. *Accuracy*, in contrast, is a measure of how close a point estimate of the PDF (e.g. the mode) is to the true value. We can have accurate but imprecise results: the point estimate of our PDF is close to the truth, but the PDF itself is very broad. And we can have precise but inaccurate results: the PDF is very narrow, but centred on a value that lies far from the truth. As we don't normally know the truth (that's why we're doing the inference), the posterior only tells us about the precision.

6.2 Inferring the parameters of a Gaussian

We turn now to a two-parameter problem, namely the inference of the mean and standard deviation of a Gaussian distribution. We will find the conditional, joint, and marginal posteriors for these two parameters.

We have a set of N data points $D = \{x_i\}$ drawn independently from a Gaussian with mean μ and standard deviation σ. The likelihood for these data is therefore

$$\begin{aligned} P(D\,|\,\mu,\sigma) &= \prod_{i=1}^{N} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right]. \end{aligned} \tag{6.5}$$

Given only D, we would like to infer one or both of the parameters of the Gaussian. The physical situation might be that we have made N measurements of some quantity, whereby μ is its unknown value and σ is the (possibly also unknown) measurement error. The maths

will be easier if we write the summation in the likelihood as

$$\sum_{i=1}^{N}(x_i-\mu)^2 \;=\; N(\overline{x}-\mu)^2 + NV_x \quad \text{where} \tag{6.6}$$

$$\overline{x} \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and}$$

$$V_x \;=\; \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2$$

which can be shown in a few lines of algebra. V_x is approximately the variance in the data (see equation 1.33). We will now look at three different cases according to what is known.

6.2.1 Standard deviation known

Suppose we know σ and want to get the posterior over μ. This is the conditional posterior, equation 6.1. As μ is a location parameter, I adopt an improper uniform prior (see section 5.3.1). The posterior is therefore just proportional to the above likelihood, so

$$\begin{aligned} P(\mu\,|\,D,\sigma) \;&\propto\; \exp\left[-\frac{1}{2\sigma^2}\left(N(\overline{x}-\mu)^2 + NV_x\right)\right] \\ &\propto\; \exp\left[-\frac{(\overline{x}-\mu)^2}{2\sigma^2/N}\right] \end{aligned} \tag{6.7}$$

where I have absorbed terms that are independent of μ into the proportionality. This posterior is a univariate Gaussian with mean $\overline{x}$ and standard deviation $\sigma/\sqrt{N}$. This is the well-known result that the "best" estimate of μ is the mean of the data, and its uncertainty is $\sigma/\sqrt{N}$. (This is not equal to the standard error in the mean – equation 2.13 – because that arises when we estimate σ from the data, whereas here σ is known.) Compare and contrast this with what the central limit theorem (section 2.3) says. It gets the same result, but for large N and for data drawn independently from any distribution (with finite mean and variance). Here we instead have data drawn from a Gaussian distribution, but the resulting posterior is Gaussian for any N.

6.2.2 Mean known

Suppose we now know μ and want to get the posterior over σ (admittedly not so common in practice). We saw in section 5.3.1 that a suitable prior for a scale parameter is the Jeffreys prior, so I adopt $P(\sigma) \propto \sigma^{-1}$. Multiplying this prior by the likelihood we get the univariate posterior

$$P(\sigma\,|\,D,\mu) \;\propto\; \frac{1}{\sigma^{N+1}}\exp\left[-\frac{1}{2\sigma^2}\left(N(\overline{x}-\mu)^2 + NV_x\right)\right]. \tag{6.8}$$

I defer discussion of this distribution because we will encounter a rather similar one in the general case.

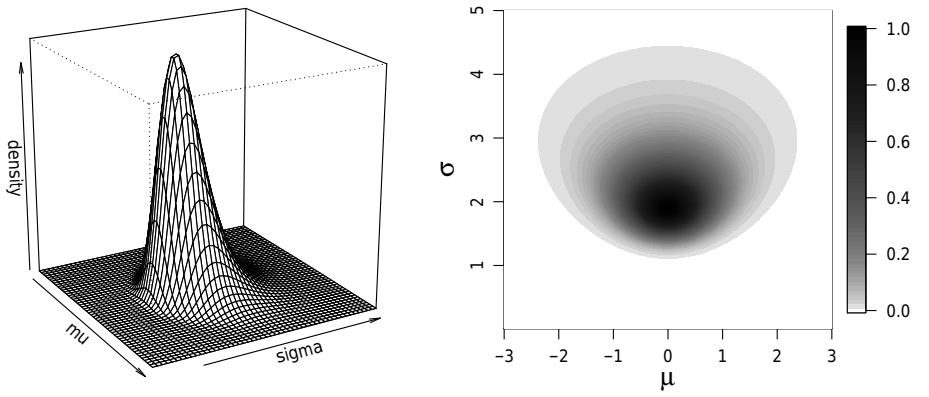

Fig. 6.1 The two-dimensional posterior $P(\mu, \sigma \,|\, D)$ in equation 6.9, which is for the parameters of a Gaussian using a uniform prior on μ and a Jeffreys prior on σ. This distribution is show in the left panel as a three-dimensional perspective mesh. The right panel indicates the value of the posterior density using a grey scale which has been been scaled so that its maximum is one. The distribution is shown in both cases for $\overline{x} = 0$, $V_x = 2^2$, and $N = 10$.

6.2.3 Neither standard deviation nor mean known

We now turn to the general case of neither μ nor σ known. We adopt the same priors as before (they are independent of each other), so $P(\mu, \sigma) \propto \sigma^{-1}$. The bivariate posterior is therefore

$$P(\mu, \sigma \,|\, D) \propto \frac{1}{\sigma^{N+1}} \exp\left[-\frac{1}{2\sigma^2}\left(N(\overline{x} - \mu)^2 + NV_x\right)\right] \tag{6.9}$$

to within a constant that does not depend on the parameters. This is a non-trivial function of both μ and σ, and is plotted in figure 6.1 for the data $\overline{x} = 0$, $V_x = 2^2$, $N = 10$. Note that the individual measurements do not appear in the posterior: these three statistics are sufficient. The code to make the plots is at the end of this section. You can use it to investigate how the shape of the posterior changes when changing the data, in particular V_x and N. Increasing N while keeping V_x constant, for example, results in a more compact PDF, i.e. we achieve a more precise determination of μ and σ, as we would expect.

If we now fix μ we get the conditional posterior over σ, equation 6.8. That this has exactly the same form as equation 6.9 follows from $P(\mu, \sigma \,|\, D) = P(\sigma \,|\, D, \mu)P(\mu)$ with $P(\mu)$ as a delta function. Note that I have not normalized either posterior. The normalized posteriors will have different units: $P(\sigma \,|\, D, \mu)$ has units σ^{-1}, whereas $P(\mu, \sigma \,|\, D)$ has units $(\mu\sigma)^{-1}$.

If we instead fix σ then we get the conditional posterior of μ, which is equation 6.7. We can visualize the conditional distribution as that obtained when taking a slice through the two-dimensional posterior. Equation 6.7 tells us that any horizontal slice through the right panel of figure 6.1 is a Gaussian.

Marginal posterior for μ

When both μ and σ are unknown, the posterior for one of the parameters is found by marginalizing over the other (equation 6.4). Let us first find $P(\mu|D)$ by marginalizing over σ. This can be thought of as projecting the two-dimensional posterior onto the μ-axis. Using the change of variables $\sigma = 1/z$ ($d\sigma = -dz/z^2$) and the standard integral

$$\int_0^\infty z^n \exp(-\beta z^m)\, dz \;=\; \frac{\Gamma(\gamma)}{m}\beta^{-\gamma} \quad \text{where} \quad \gamma = \frac{n+1}{m} \quad \text{and} \quad (m, n, \beta) > 0 \tag{6.10}$$

(Γ is the gamma function defined by equation 1.62) we get

$$\begin{aligned} P(\mu|D) \;&=\; \int_0^\infty P(\mu, \sigma|D)\, d\sigma \\ &\propto \int_0^\infty z^{N-1} \exp\left[-z^2\left(\frac{N(\overline{x}-\mu)^2 + NV_x}{2}\right)\right] dz \\ &\propto \left[1 + \frac{(\overline{x}-\mu)^2}{V_x}\right]^{-N/2} \end{aligned} \tag{6.11}$$

where I have absorbed factors independent of μ into the proportionality constant in the second line, and have absorbed an additional factor of $(NV_x/2)^{-N/2}$ in going to the third line. This density function is called a *Student's t distribution* (or just "t distribution"). The standard, normalized form for this distribution is

$$P(t) \;=\; \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \nu > 0 \tag{6.12}$$

where ν, the *degrees of freedom*, is its sole parameter. Examples of the distribution are shown in figure 6.2. It is symmetric about $t = 0$ and becomes increasingly like a standardized Gaussian for larger values of ν. Comparing equation 6.12 with equation 6.11, we see that our marginal posterior is a t distribution with $\nu = N - 1$ degrees of freedom and

$$t \;=\; \frac{(\overline{x} - \mu)}{\sqrt{V_x/(N-1)}}. \tag{6.13}$$

Here t is the difference between the sample mean and the true mean, scaled by a quantity which, from reference to equation 2.13, is the standard error in the mean $\hat{\sigma}/\sqrt{N}$, where $\hat{\sigma} = \sqrt{V_x N/(N-1)}$ is an estimate of the standard deviation of the data. The best estimate of μ is thus $\overline{x}$ (which is the mean, mode, and median of the posterior), and its uncertainty is the standard error in the mean. We stated this classic statistical result in section 2.4, but now we have arrived at it via an inference procedure for data drawn from a Gaussian likelihood when we adopt Jeffreys priors on the mean and standard deviation. Previously we did not state what the distribution over the mean was; now we have shown it to be a t distribution.

As is apparent from figure 6.2, the t distribution has heavier tails than the standardized Gaussian. This reflects the fact that when σ is unknown, the data are not as informative about the mean as they are when σ is known (in which case the posterior is a Gaussian,

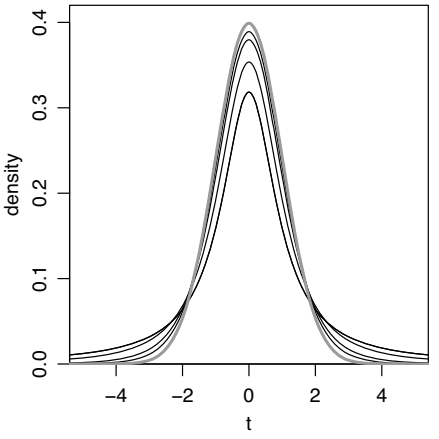

Fig. 6.2 The Student's t distribution for 1, 2, 5, and 10 degrees of freedom (dof). The larger the dof, the higher the central peak. The thick grey line shows a standardized Gaussian for comparison, to which the t distribution converges for infinite degrees of freedom.

as we saw in section 6.2.1). The more data we have, the better the determination of the parameters, the closer the t distribution gets to a Gaussian, and $\hat{\sigma}$ asymptotes to $\sqrt{V_x}$.

If we adopted a uniform prior on σ, $P(\sigma) \propto \sigma^0$, rather than a Jeffreys prior ($\propto \sigma^{-1}$), we see from inspection of the above equations that the marginal posterior $P(\mu\,|\,D)$ would again be a t distribution, but with $N-2$ degrees of freedom. For large N the distribution hardly depends on the value of N, so the difference between these priors then becomes irrelevant. But given that σ cannot be negative, it would be illogical to put a uniform prior on it.

We will encounter the t distribution again in section 10.2.2.

Marginal posterior for σ and σ^2

The marginal posterior for σ is found by integrating equation 6.9 over μ

$$\begin{aligned} P(\sigma\,|\,D) &= \int_{-\infty}^{\infty} P(\mu,\sigma\,|\,D)\,d\mu \\ &\propto \frac{1}{\sigma^{N+1}} \exp\left[-\frac{NV_x}{2\sigma^2}\right] \int_{-\infty}^{\infty} \exp\left[-\frac{N(\overline{x}-\mu)^2}{2\sigma^2}\right] d\mu \\ &\propto \frac{1}{\sigma^{N}} \exp\left[-\frac{NV_x}{2\sigma^2}\right] \end{aligned} \tag{6.14}$$

as the Gaussian integral is just $\sqrt{2\pi\sigma^2/N}$. Plots of this for $V_x = 2^2$ and various values of N are shown in figure 6.3. Note that equation 6.14 is similar to the conditional posterior for σ in equation 6.8 if we set $\overline{x} = \mu$. They then only differ by a factor of σ, with the marginal distribution being slightly broader on account of the mean not being known.

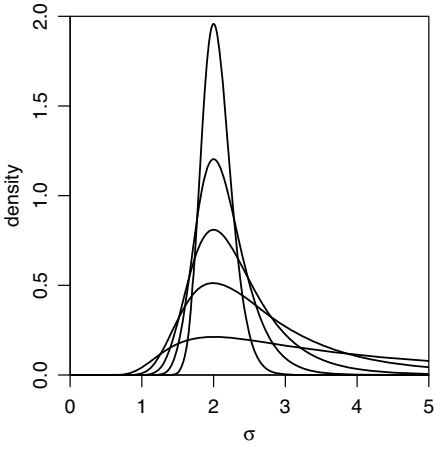

Fig. 6.3 The posterior $P(\sigma\,|\,D)$ (equation 6.14) for $V_x = 2^2$ and $N = 2, 5, 10, 20, 50$, whereby the higher peaks correspond to the larger N.

We can write equation 6.14 in terms of the variance σ^2 using a transformation of variables (section 1.9.1) $P(\sigma\,|\,D) = 2\sigma P(\sigma^2\,|\,D)$. This gives

$$P(\sigma^2\,|\,D) \;\propto\; \frac{1}{(\sigma^2)^{(N+1)/2}} \exp\left[-\frac{NV_x}{2\sigma^2}\right] \tag{6.15}$$

which is an *inverse gamma distribution*. Examples of this are shown in figure 6.4. For a variable z it is usually written in terms of the two parameters, the shape α and scale β,

$$P(z) \;=\; \frac{\beta^\alpha}{\Gamma(\alpha)}\, z^{-(\alpha+1)} e^{-\beta/z} \quad \text{where} \quad \alpha > 0,\ \beta > 0,\ z > 0. \tag{6.16}$$

In our application $z = \sigma^2$, $\alpha = (N-1)/2$, and $\beta = NV_x/2$. To satisfy the condition on α the distribution is only defined for $N \geq 2$. And to satisfy the condition for β there need to be at least two *different* data points, to ensure $V_x > 0$. This makes sense. From a logical point of view we don't expect to be able to make an inference on σ with just one data point, because we used improper priors. From a mathematical point of view the distribution is not defined for $N = 1$, because it cannot be normalized. As we used improper priors there is no guarantee that we would get a proper (normalizable) posterior, and here for $N = 1$ we do not. As soon as we have at least two (finite and different) data points, the likelihood ensures we get convergence of the posterior distribution. For $N = 0$ we have no data, so the posterior is identical to the prior anyway.

I wrote in section 5.4 that the inverse gamma distribution was the conjugate prior for the variance of a Gaussian likelihood of *known* mean μ. This is now evident from the form of the dependence on $z = \sigma^2$ of equation 6.16, which is

$$P(\sigma^2) \;\propto\; (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}. \tag{6.17}$$

Adopting this as our prior and multiplying it by the likelihood in equation 6.5, which we

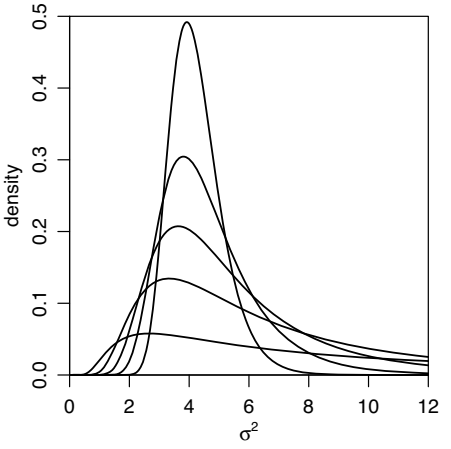

Fig. 6.4 The inverse gamma distribution of equation 6.16 with $V_x = 2^2$ and $N = 2, 5, 10, 20, 50$, whereby the higher peaks correspond to the larger N. The mode is at $V_x(N/N-1)$, so as $N \to \infty$ the mode tends towards V_x.

can write as

$$P(D\,|\,\mu,\sigma^2) \;\propto\; (\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right] , \tag{6.18}$$

we see from inspection that the posterior has the same functional form as the prior, with shape parameter $\alpha + N/2$ and scale parameter $\beta + (1/2)\sum_i (x_i - \mu)^2$.

The following code produces the density plot in the right panel of figure 6.1. I compute the (unnormalized) two-dimensional posterior on a dense grid, then use the function `image.plot` in the `fields` package to plot this as a greyscale image along with a bar to indicate the scale. The palette is made with the `RColorBrewer` package, which of course produces colour scales too (replace `Greys` with `YlOrRd` to produce a red/yellow scale, for example). To produce the mesh plot in the left panel of figure 6.1 use

```
persp(x=mu, y=sigma, z=postDen, phi=20, theta=60, d=5,
  xlab=expression(mu), ylab=expression(sigma), zlab="density")
```

but use only 50 elements in `mu` and `sigma` (i.e. set `length.out=50` in their constructions) so that the shading works out nicely.

R file: `2D_gaussian_posterior.R`

```
##### Plot 2D posterior over Gaussian mu and sigma
##### for uniform prior on mu and Jeffreys prior on sigma

library(fields) # for image.plot
library(RColorBrewer) # for colorRampPalette
mypalette <- colorRampPalette(brewer.pal(9, "Greys"), space="rgb",
                              interpolate="linear", bias=2.5)
mycols <- mypalette(64)
```

```
# Define function to return the unnormalized posterior
post <- function(mu, sigma, xbar, Vx, N) {
  (1/sigma^(N+1))*exp( (-N/(2*sigma^2)) * ((xbar-mu)^2 + Vx) )
}

# Define data and calculate posterior density on a dense grid
xbar  <- 0
Vx    <- 2^2
N     <- 10
mu    <- seq(from=-3,   to=3, length.out=1e3)
sigma <- seq(from=0.01, to=5, length.out=1e3)
postDen <- matrix(data=NA, nrow=length(mu), ncol=length(sigma))
for(i in 1:length(mu)) {
  for(j in 1:length(sigma)) {
    postDen[i,j] <- post(mu=mu[i], sigma=sigma[j], xbar=xbar, Vx=Vx, N=N)
  }
}
postDen <- postDen/max(postDen) # scale so maximum is one

pdf("2D_gaussian_posterior.pdf", 5, 4)
par(mfrow=c(1,1), mar=c(3.5,3.5,0.5,1), oma=c(0.1,0.1,0.5,0.1),
    mgp=c(2.2,0.8,0), cex=1.0)
image.plot(z=postDen, x=mu, y=sigma, nlevel=1024, xlab=expression(mu),
           ylab=expression(sigma), col=mycols, cex.lab=1.5)
dev.off()
```

6.3 A two-parameter problem: estimating amplitude and background

We now turn to a parameter estimation problem for which the posterior does not have a convenient functional form.[1]

A spectrograph is a device for spreading out light along a detector as a function of wavelength. We use this to measure the intensity – the number of photons d – as a function of the wavelength x. Suppose we know that this spectrum comprises a single emission line on top of a constant background. We would like to determine the amplitude a of the emission line in the presence of both measurement noise and an unknown background level b. Our model for the line is that it has a Gaussian shape centred at $x = x_0$ with standard deviation w. Thus the expected signal (number of photons) at any position x is

$$s = t\left[a\exp\left(-\frac{(x-x_0)^2}{2w^2}\right)+b\right] \tag{6.19}$$

where t is proportional to the exposure time, and so is proportional to the expected number of photons collected. If we expose for longer we expect to get more signal from both source and background, but without changing the shape of either. This model is shown as the grey

[1] This example is based on one in Sivia & Skilling (2006).

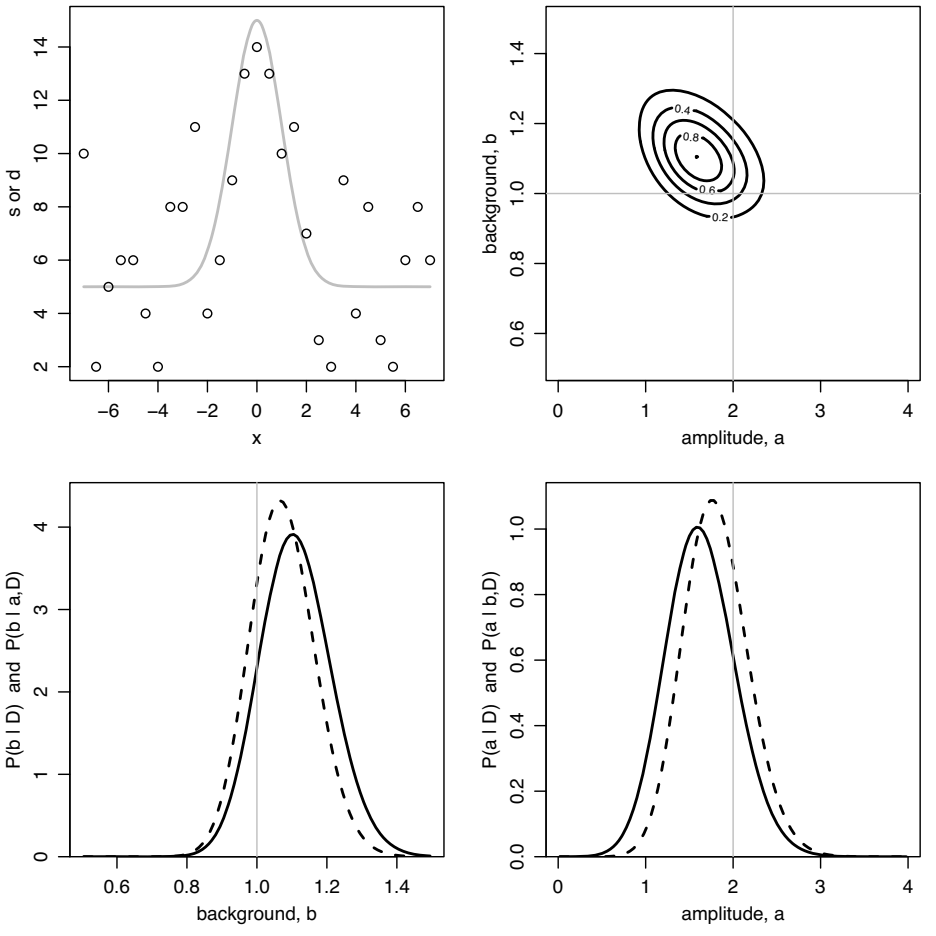

Fig. 6.5 Estimating the amplitude and background of a signal. Top left: true model (grey curve) and observed data (open circles). Top right: two-dimensional unnormalized posterior, scaled to have a maximum of unity. The contours are equal spacing in probability density, labelled by the density relative to the maximum. Bottom two panels: the marginalized posteriors (solid lines), and conditional posteriors (dashed lines) using the true values of the parameters. The vertical grey lines indicate the true parameters.

line in the top-left panel of figure 6.5 for $x_0 = 0, w = 1, t = 5, a = 2, b = 1$. These are the true values of the model parameters which I will use to simulate the data in this example.

We measure this signal at a number of different positions $\{x_i\}$ on the detector. If the expected number of photons at any of these positions is s (which is not generally an integer), then the number of photons actually observed, d, follows a Poisson distribution with mean

s, i.e.

$$P(d|s) = \frac{s^d e^{-s}}{d!}. \tag{6.20}$$

One possible realization of the measurements, and the one which I will use for the subsequent inference, is shown as the open circles in the top-left panel of figure 6.5. There are 29 data points measured uniformly from $x = -7w$ to $x = +7w$ in steps of $0.5w$. Denoting the set of measurements as $D = \{d_i(x_i)\}$, the likelihood of the data is

$$P(D|x_0, w, t, a, b) = \prod_i \frac{s_i^{d_i} e^{-s_i}}{d_i!} \tag{6.21}$$

where $s_i = s(x_i)$ from equation 6.19 and so introduces the dependence on the model parameters.

The model has five parameters, but let us assume that x_0, w, and t are known. We therefore want to infer $P(a, b|D, M)$ from the data, where the model M expresses that we know the shape of the line and the values of the fixed parameters (and that the noise model is Poisson). Let us adopt the minimalistic prior that a and b cannot be negative. I set no upper limit, so this is an improper prior. Provided the posterior drops to very low densities at large a and b, we can later truncate it at some larger values of a and b in order to normalize it. This is equivalent to the prior dropping to zero at these values, although by construction this has no relevant impact on the posterior.[2] The prior $P(a, b)$ is constant when both a and b are positive, and zero otherwise. The posterior is therefore

$$P(a, b|D, M) = \begin{cases} \frac{1}{Z} \prod_i \frac{s_i^{d_i} e^{-s_i}}{d_i!} & \text{if } a \geq 0 \text{ and } b \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.22}$$

and the log posterior is

$$\ln P(a, b|D, M) = \begin{cases} \sum_i d_i \ln(s_i) - s_i + \text{constant} & \text{if } a \geq 0 \text{ and } b \geq 0 \\ -\infty & \text{otherwise} \end{cases} \tag{6.23}$$

where the constant absorbs terms which do not depend on a or b (as they do not effect the shape of the function). Given the $\{x_i\}$ we calculate the expected counts $\{s_i\}$ from the generative model, equation 6.19, which gives the dependence on the parameters. Equation 6.23 has a nonlinear dependence on the parameters, so solving for the maximum, full-width at half-maximum, etc., of the posterior is not analytically simple. But we can plot the posterior just by calculating it on a grid of values of $\{a_j, b_k\}$. This is done by the R code in the following section. I use a regular grid of size $K \times K$ with $K = 100$ over a pre-defined range with spacing δa in a and δb in b. I found the best range by trial and error. The `contour` function in R is then used to find and plot smooth contours of constant

[2] We must be careful to ensure that the probability mass in the posterior beyond the truncated range really is negligible. This can be hard to know in more complex or higher dimensional problems. Of course if the likelihood – seen as a function of the parameters – drops to zero within a finite region, or if it asymptotes fast enough to have a finite integral, then this truncation is not necessary.

probability density. The result is shown in the top-right panel of figure 6.5. We see a clear anticorrelation between a and b. This means we can get almost equally good models of the data if we simultaneously increase b and decrease a slightly (or vice versa). This makes sense, because if we attribute more of the measured signal to the background, we must attribute less to the spectral line.

The bottom two panels of figure 6.5 show the marginal posterior PDFs of the two parameters. These are formally found by integration (equation 6.3). But as I have a grid, I approximate them by simply summing over the other parameter. For a this is

$$P(a_j \,|\, D) \;\simeq\; \delta b \sum_{k=1}^{K} P(a_j, b_k \,|\, D), \tag{6.24}$$

and similarly for b. The posterior is normalized in the program via the rectangle rule using this grid, in the same way as done in section 5.1.1.

We can use these grid evaluations also to estimate the mean μ_a and variance σ_a^2 of a (and likewise for b), as well as the covariance $\mathrm{Cov}(a, b)$ and correlation coefficient ρ (see section 1.6.1). We must ensure that the grid evaluations extend well into the tails of the posterior to ensure that we capture essentially all of the probability. It follows from the definitions of these quantities that

$$\mu_a \;=\; \int a P(a \,|\, D) da \;\simeq\; \delta a \sum_{j=1}^{K} a_j P(a_j \,|\, D) \tag{6.25}$$

$$\sigma_a^2 \;=\; \int (a - \mu_a)^2 \, P(a \,|\, D) da \;\simeq\; \delta a \sum_{j=1}^{K} (a_j - \mu_a)^2 P(a_j \,|\, D) \tag{6.26}$$

$$\begin{aligned} \mathrm{Cov}(a, b) \;&=\; \iint (a - \mu_a)(b - \mu_b) \, P(a, b \,|\, D) \, da \, db \\ &\simeq\; \delta a \delta b \sum_{j=1}^{K} \sum_{k=1}^{K} (a_j - \mu_a)(b_k - \mu_b) P(a_j, b_k \,|\, D) \end{aligned} \tag{6.27}$$

$$\rho \;=\; \frac{\mathrm{Cov}(a, b)}{\sigma_a \sigma_b}. \tag{6.28}$$

Strictly speaking I should have a factor of $K/(K-1)$ in equations 6.26 and 6.27 because the means μ_a and μ_b have been estimated from the data. But with $K = 100$ this introduces a negligible error on top of the approximation from gridding. For the data shown in the top-left panel of figure 6.5 the resulting estimates of the parameters (mean $\pm$ standard deviation) are $a = 1.63 \pm 0.40$ and $b = 1.11 \pm 0.10$, and the correlation coefficient is $\rho = -0.40$. Note that we cannot use the R functions `mean`, `sd`, `cov`, and `cor` to calculate the above, because these functions assume the samples they operate on have equal probability density. That is, they assume the samples have been drawn from the posterior PDF. But that is not the case here: we have evaluated the PDF at pre-defined points. We will learn in chapter 8 how we can sample from arbitrary probability density functions.

If we knew the value of a or b, then we could compute the conditional posteriors (equation 6.1). These are plotted as dashed lines in the lower two panels of figure 6.5, using the

true values of the parameters in each case. Note how the conditional posteriors are slightly narrower and higher than the marginal posteriors. This just reflects the fact that when we fix some parameters, more of the data can be used to determine the other parameters. This requires that the data depend on the fixed parameters, of course. If the fixed parameters were irrelevant, then fixing them would not improve our inference about the other parameters.

6.3.1 R code for fitting the amplitude and background

The code below performs all the computations described above and produces the plots in figure 6.5. The code is documented and should be reasonably self-explanatory. Equation 6.19 is evaluated by the function `signal`, and `logupost` gives the (natural) logarithm of the unnormalized posterior (equation 6.23). By default, functions in R return whatever is evaluated in their last line, so you don't have to use the `return` command for this. The posterior is sampled on a uniform grid of size `Nsamp` $\times$ `Nsamp`. Each dimension is sampled in the same way as in `coin1.R` in section 5.1.1. Note that some data points can have a very low likelihood and therefore a very low posterior density. As explained below (section 6.3.3) these – in particular `exp(z)` – can become numerically identical to zero. If, as you experiment with changing values (see the suggestions after the code), you produce a posterior which is very narrow, and you use a sampling which is too coarse, then you may find that all grid evaluations give zero posterior density. This would cause a divide by zero when you try to normalize the posteriors. It would be straightforward to catch these errors, but this is not implemented in this code.

R file: `signal_background_estimation.R`

```
##### Infer posterior PDF over amplitude and background parameters

# Define function to return true signal at position x (generative model)
signal <- function(x, a, b, x0, w, t) {
  t*(a*exp(-(x-x0)^2/(2*w^2)) + b)
}

# Define function to return (natural) log posterior over (a,b).
# Prior on a and b: P(a,b) = const if a>0 and b>0, = 0 otherwise.
# Likelihood for one point is Poisson with mean d(x), so total
# likelihood is their product. Unnormalized posterior is product of these.
# d and x are equal length vectors (or scalars). The rest are scalars.
logupost <- function(d, x, a, b, x0, w, t) {
  if(a<0 || b <0) {return(-Inf)} # the effect of the prior
  sum(dpois(d, lambda=signal(x, a, b, x0, w, t), log=TRUE))
}

# Set model parameters (true and fixed)
x0    <- 0 # centre of peak
w     <- 1 # sd of peak
atrue <- 2 # amplitude
btrue <- 1 # background
t     <- 5 # scale factor (exposure time -> sets SNR)

# Simulate some data (by drawing from the likelihood)
```

```
set.seed(205)
xdat  <- seq(from=-7*w, to=7*w, by=0.5*w)
strue <- signal(xdat, atrue, btrue, x0, w, t)
ddat  <- rpois(length(strue), strue)

# Define sampling grid to compute posterior (will be normalized
# over this range too). uniGrid spans the range 0-1 with Nsamp
# points. This is then scaled to cover the ranges alim and blim.
alim  <- c(0.0, 4.0)
blim  <- c(0.5, 1.5)
Nsamp <- 1e2
uniGrid <- seq(from=1/(2*Nsamp), to=1-1/(2*Nsamp), by=1/Nsamp)
delta_a <- diff(alim)/Nsamp
delta_b <- diff(blim)/Nsamp
a <- alim[1] + diff(alim)*uniGrid
b <- blim[1] + diff(blim)*uniGrid

# Compute log unnormalized posterior, z = ln P^*(a,b|D), on a regular grid
z <- matrix(data=NA, nrow=length(a), ncol=length(b))
for(j in 1:length(a)) {
  for(k in 1:length(b)) {
    z[j,k] <- logupost(ddat, xdat, a[j], b[k], x0, w, t)
  }
}
z <- z - max(z) # set maximum to zero

# Compute normalized marginalized posteriors, P(a|D) and P(b|D)
# by summing over other parameter. Normalize by gridding.
p_a_D <- apply(exp(z), 1, sum)
p_a_D <- p_a_D/(delta_a*sum(p_a_D))
p_b_D <- apply(exp(z), 2, sum)
p_b_D <- p_b_D/(delta_b*sum(p_b_D))

# Compute mean, standard deviation, covariance, correlation, of a and b
mean_a <- delta_a * sum(a * p_a_D)
mean_b <- delta_b * sum(b * p_b_D)
sd_a   <- sqrt( delta_a * sum((a-mean_a)^2 * p_a_D) )
sd_b   <- sqrt( delta_b * sum((b-mean_b)^2 * p_b_D) )
# To calculate the covariance I need to normalize P(a,b|D) = exp(z).
# I do it here by brute force with two loops (there are better ways in R).
# The normalization constant is Z = delta_a*delta_b*sum(exp(z)).
# This is independent of (a,b) so can be calculated outside of the loops.
# The factor delta_a*delta_b will just cancel in the expression for
# cov_ab, so I omit it entirely.
cov_ab <- 0
for(j in 1:length(a)) {
  for(k in 1:length(b)) {
    cov_ab <- cov_ab + (a[j]-mean_a)*(b[k]-mean_b)*exp(z[j,k])
  }
}
cov_ab <- cov_ab / sum(exp(z))
rho_ab <- cov_ab / (sd_a * sd_b)
cat("  a = ", mean_a, "+/-", sd_a, "\n")
cat("  b = ", mean_b, "+/-", sd_b, "\n")
cat("rho = ", rho_ab, "\n")
```

```
# Compute normalized conditional posteriors, P(a|b,D) and P(b|a,D)
# using true values of conditioned parameters. Vectorize(func, par)
# makes a vectorized function out of func in the parameter par.
p_a_bD <- exp(Vectorize(logupost, "a")(ddat, xdat, a, btrue, x0, w, t))
p_a_bD <- p_a_bD/(delta_a*sum(p_a_bD))
p_b_aD <- exp(Vectorize(logupost, "b")(ddat, xdat, atrue, b, x0, w, t))
p_b_aD <- p_b_aD/(delta_b*sum(p_b_aD))

# Make plots

pdf("signal_background_estimation.pdf", 7, 7)
# Plot true model and data
par(mfrow=c(2,2), mgp=c(2,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
xplot <- seq(from=min(xdat), to=max(xdat), by=0.05*w)
splot <- signal(xplot, atrue, btrue, x0, w, t)
plot(xplot, splot, ylim=range(c(splot, ddat)), xlab="x", ylab="s or d",
     type="l", col="grey", lwd=2)
points(xdat, ddat)
# Plot unnormalized 2D posterior as contours.
# Note that they are labelled by posterior density relative to peak,
# NOT by how much probabilty they enclose.
contour(a, b, exp(z), nlevels=5, labcex=0.5, lwd=2, xlab="amplitude, a",
        ylab="background, b")
abline(v=2,h=1,col="grey")
# Plot the 1D marginalized posteriors
plot(b, p_b_D, xlab="background, b", yaxs="i",
     ylim=1.05*c(0,max(p_b_D, p_b_aD)), ylab="P(b | D)  and  P(b | a,D)",
     type="l", lwd=2)
lines(b, p_b_aD, lwd=2, lty=2)
abline(v=btrue, col="grey")
plot(a, p_a_D, xlab="amplitude, a", yaxs="i",
     ylim=1.05*c(0,max(p_a_D, p_a_bD)), ylab="P(a | D)  and  P(a | b,D)",
     type="l", lwd=2)
lines(a, p_a_bD, lwd=2, lty=2)
abline(v=atrue, col="grey")

dev.off()
```

6.3.2 Suggested experiments

It is instructive to experiment with changing both the simulated data and the values of the parameters in the above example. Here are some suggestions.

(1) Change the exposure time used to generate the data. Recall from the telescope example on page 19 that increasing the number of photons gathered will increase the signal-to-noise ratio. Try values of t of 0.5, 1, 2, 10, 50, 100. You will need to adjust the range over which the posterior is calculated (`alim`, `blim`) in order to sample it adequately (i.e. to cover its full range and to sample it at high enough resolution). If you set the limits to less than 0 you will see that the posterior is correctly truncated by the prior. A larger t corresponds to more photons. You should find that this gives a more accurate and precise determination of the parameters (and a smaller t a lower accuracy and precision).

(2) Vary the sampling resolution of x used to generate the data, but keeping the sampling range the same (so the number of data points will change). The samples are defined as follows.
```
xdat <- seq(from=-7*w, to=7*w, by=0.5*w)
```
In the above we have a resolution of $0.5w$, so try changing this to $0.1, 0.25, 1, 2, 3$ times w, for example. A better sampling of the spectral line will result in a more accurate and precise determination of its amplitude.

(3) Vary the sampling range (keep it centred on x_0), but keep the sampling resolution the same (so the number of data points will change). I suggest values of $\pm 3w$, $\pm 10w$, $\pm 20w$, $\pm 50w$. If a larger fraction of the grid evaluations are dominated by the background, is the accuracy and precision with which we determine the amplitude reduced? What about the background? Is the covariance affected?

(4) Vary the sampling range (keep it centred on x_0) between $0.5w$ and $2w$, but now with a sampling resolution of $0.1w$. In these cases we are barely sampling across the whole line. What happens to the posterior PDFs? If we set the range to $\pm 0.5w$ but have high resolution, say $0.01w$, do we improve things?

(5) Change the ratio a/b used to simulate the data (keeping both positive in accordance with the prior). The smaller this is, the less prominent the line. Is the amplitude then less accurately determined?

6.3.3 A note on computation with finite precision

When you experiment with the above example you will begin to realise (if you haven't already) that solving continuous mathematical problems on discrete computers involves dangers. Functions cannot be sampled with infinite resolution, nor can numbers be represented with infinite precision. In particular, numbers cannot be arbitrarily small on computers. This becomes a problem when sampling likelihoods or posteriors. As the likelihood for N independent data points is the product of N numbers that are often less than one, the likelihood can become arbitrarily small for larger data sets. This is the case even at the true parameter values (see section 5.2). While the absolute (small) value of the likelihood does not matter in principle, it does matter in practice, because with finite precision the likelihood (and therefore posterior) may be truncated numerically to zero for all sampled parameters. This will lead to a zero normalization constant, resulting in divide by zero and possibly other errors. You can run into this problem quicker than you realise, so you should be careful to trap such errors in code which manipulates likelihoods.

One way to address this is to work with log likelihoods, because this provides us with a much larger dynamical range: the number 10^{-1000} may be truncated by the computer representation to be numerically identical to zero, but the number -1000 will not be. In some situations we must calculate actual likelihoods, however. This is the case when we calculate the marginal likelihood (equation 3.16) which we will need to do in chapter 11. We can mitigate numerical problems to some extent with the following method.

As we will see in section 8.3, the marginal likelihood may be approximated by the average of the likelihoods computed at a set of parameter values drawn from the prior. Let $\{\log L_l\}$ be this set of N log likelihoods. Suppose that some or all of these values are so

negative (because the likelihoods are so small) that $10^{\log L_l}$ is truncated by the computer representation to be identical to zero. The estimated marginal likelihood

$$E = \frac{1}{N}\sum_{l=1}^{N} 10^{\log L_l} \tag{6.29}$$

would then be incorrect (underestimated), and in the worst case zero. To overcome this we add a constant h to every $\log L_l$ term. Provided the range of the set $\{L_l\}$ is less than the computer's dynamic range,[3] we can choose h to ensure that $10^{(h+\log L_l)}$ is not truncated for any l. When we work with these we will compute the marginalized likelihood

$$\begin{aligned} E' &= \frac{1}{N}\sum_{l=1}^{N} 10^{(h+\log L_l)} \\ &= 10^h E \\ \log E' &= h + \log E \end{aligned} \tag{6.30}$$

from which we can calculate the quantity we actually want, E. A good approach is to set the largest values of the exponent to be zero, as we may not care about just the smallest of the set of likelihoods being numerically indistinguishable from zero. In that case an appropriate choice for h is $-\max(\log L_l)$.

[3] This is about 630 orders of magnitude on my machine when using double precision.