Tutorial on Bayesian Inference (2019)
solutions & hints

3) Satellites: the next generation

We investigate how the contraints of an old satellite are updated when a new satellite carries out the same measuremen and increased precision. What we want, is to measure θ.
The sampling distribution of the future data point y from the upcoming satellite is

$$\mathcal{P}(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}\right). \tag{0.1}$$

The prior distribution is given by the constraints from the old satellite, and Gaussian according to the exercise, so

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left(-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\tau_0^2}\right). \tag{0.2}$$

The hyper-parameters are τ_0 and μ_0, the prior could therefore equally be written as $\pi(\theta|\tau_0, \mu_0)$, but we do not write out this dependence here, since τ_0 and μ_0 are simply features of the old satellite, and we therefore do not need to infer them (we know them already). μ_0 is not a parameter we have control over (it is the peak of the old measurement, so this position is set by Nature). In contrast, τ_0 is something the scientific generation before us had control over: it is the precision of the experiment, set by the experimental design.
To get the posterior for the joint constraints, we need to apply Bayes' theorem

$$\mathcal{P}(\theta|y) \propto \mathcal{P}(y|\theta)\pi(\theta). \tag{0.3}$$

The proportionality is sufficient, since we wish to infer θ, and the posterior can be normalized afterwards. Accepting the proportionality means that $\mathcal{P}(y)$ should not be of any interest.
Multiplying likelihood and prior together, we get

$$\mathcal{P}(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right). \tag{0.4}$$

To see that this is still a Gaussian, we need to write the exponent as a quadratic form. The trick to get there is always the same (and features in the script, e.g. also where I derive the Wiener filter): First expand the squares currently in the exponent. Then complete the square in θ. Lastly collect terms which do not depend on θ, they will go into the normalization (i.e. they are uninteresting). The result is

$$\mathcal{P}(\theta|y) = \mathcal{G}(\mu_1, \tau_1^2), \tag{0.5}$$

where

$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{y}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}. \tag{0.6}$$

Plotting is straightforward. If μ_0 is many σ away from y, then the prior and the sampling distribution will be discrepant. If $\tau_0 \ll \sigma$, then the prior will be much narrower than the measurement uncertainty, i.e. here we have the limit of a highly informative prior.

5) Straight line fitting with a twist

We have a single data 'point', i.e. a single point in a two-dimensional plot. Of this point, we have the noisy $\hat{x}$-value, and the noisy $\hat{y}$ value. We assume there is a straight line explaining the relationship between x, y, but we cannot observe x and y directly. Being Bayesian, everything we can't observe becomes a parameter. Parameters can either be nuisance parameters or parameters of interest. Since we want to measure the slope m, we have that m is the parameter of interest, and x and y are nuisance parameters.

1. The straigh line with slope m reads $y = mx$, hence the slope m relates y to x, not $\hat{y}$ to $\hat{x}$.

2. The sampling distribution $\mathcal{P}(\hat{x}|x)$ is (according to the exercise)

$$\mathcal{P}(\hat{x}|x) = \mathcal{G}(x, \sigma^2)$$

   and we have likewise for $\hat{y}$ that

$$\mathcal{P}(\hat{y}|y) = \mathcal{G}(y, \sigma^2).$$

3. Since $\hat{x}$ and $\hat{y}$ are drawn independently from two Gaussians (the key word is independence), we have that the joint probability is the product of the two individual probabilities

$$\mathcal{P}(\hat{x}, \hat{y}|x, y) = \mathcal{P}(\hat{x}|x)\mathcal{P}(\hat{y}|y).$$

4. The relationship $y = mx$ is noise-free. The distribution which expresses a precise relationship is Dirac's delta distribution $\delta_D(\cdot)$. We therefore have

$$\mathcal{P}(m|x, y) = \delta_D(m - y/x)$$

   and

$$\mathcal{P}(y|x, m) = \delta_D(y - mx).$$

5. The exercise wishes us to invert the conditional dependence $\mathcal{P}(m|\hat{x}, \hat{y})$. The inverted order is $\mathcal{P}(\hat{x}, \hat{y}|m)$. The question which confuses many people is 'how do I invert a conditional dependence if there are more than *two* variables in the game?'. The answer is that one imagines that an event like the joint draw of $\hat{x}, \hat{y}$ is a single new event: $(\hat{x}, \hat{y}) = \hat{A}$. This means that the conditional dependence can be inverted blockwise. We use Bayes' theorem and get

$$\mathcal{P}(\hat{x}, \hat{y}|m) = \frac{\mathcal{P}(m|\hat{x}, \hat{y})\pi(\hat{x}, \hat{y})}{\pi(m)}. \tag{0.7}$$

   The element of confusion (which is to be discussed in detail, and asked for) is the appearance of $\pi(\hat{x}, \hat{y})$.

6. Finally, we solve the BHM. By 'solving a BHM' it is meant that the sought posterior is expressed as a function of known distributions. It can then be calculated.

   We are searching for $\mathcal{P}(m|\hat{x}, \hat{y})$. To get there, we investigate the joint distribution of all our variables. This is $\mathcal{P}(\hat{x}, \hat{y}, x, y, m)$.[1]

   To compute $\mathcal{P}(\hat{x}, \hat{y}, x, y, m)$, different students will come up with different solutions. The two most frequently encountered ones are (a) the physical forward model, and (b) a telescopic expansion.

[1] In the lectures, I split this distribution rather rapidly into the components that I want. This speed comes with experience. You can also systematically try multiple possibilities, the result will be the same, but the calculations then become somewhat lengthy. At each step, you have to argue logically; it is always a case-by-case situation, which is why I explain the whys-and-hows of each step at the blackboard.

Via the physical forward model:
The forward model goes 'First we need a prior on our parameter to be measured, this is $\pi(m)$. Then we need a prior on the x-value since it is latent, this is $\pi(x)$. Now we know that y is automatically given once we have drawn from $\pi(m)$ and $\pi(x)$, hence we have $\delta_D(y-mx)$. Finally, the observables scatter around x and y, hence we have $\mathcal{P}(\hat{x}|x)$ and $\mathcal{P}(\hat{y}|y)$.'

Putting the above together in right-to-left order, we arrive at

$$\mathcal{P}(\hat{x},\hat{y},x,y,m) = \mathcal{P}(\hat{x}|x)\mathcal{P}(\hat{y}|y)\delta_D(y-mx)\pi(x)\pi(m). \tag{0.8}$$

In reality, such a 'I multiply all distributions together by arguing in a physicist's manner' often leads to confusion or insecurity. The slower approach via a telescopic expansion typically clarifies this, as it lays bare each single assumption.

Via a telescopic expansion:
A more cautious way to facilitate the joint distribution runs via a telescopic expansion. Here, it is often a great ease if the variables are ordered at least approximately in the order they appear in the physical forward model (I have done that in the exercise already, to make it easier. In general however, the ordering has to found.)

We begin by systematically splitting off the last terms. This is slow, but it always works. Let us see what happens if we split off the very last term:

$$\mathcal{P}(\hat{x},\hat{y},x,y,m) = \mathcal{P}(\hat{x},\hat{y},x,y|m)\pi(m). \tag{0.9}$$

This did not help us too much. All we learned is that we will need a prior $\pi(m)$.

We hence split off more than the last variable:

$$\begin{aligned}\mathcal{P}(\hat{x},\hat{y},x,y,m) &= \mathcal{P}(\hat{x},\hat{y},y|x,m)\pi(x,m)\\ &= \mathcal{P}(\hat{x},\hat{y},y|x,m)\pi(x|\cancel{m})\pi(m)\\ &= \mathcal{P}(\hat{x},\hat{y},y|x,m)\pi(x)\pi(m)\end{aligned} \tag{0.10}$$

This did help already: We first split off x and m, then wrote out their joint distribution as a conditional and a prior, and then facilitated the conditional: x does not depend on m. When solving a BHM, the first two equalities here are always the same: it simply splits off variables and then expands their joint distributions, mathematically this is always correct. It is the last equality that is problem-specific: here, one needs to think. Does the problem-setup allow a facilitation? Does the general maths allow for a conditional dependence which does not exist in our specific problem? (This was the case here, where I consequently cancelled the m in the second line.)

Next, we split off three terms

$$\begin{aligned}\mathcal{P}(\hat{x},\hat{y},x,y,m) &= \mathcal{P}(\hat{x},\hat{y}|y,x,\cancel{m})\pi(y,x,m)\\ &= \mathcal{P}(\hat{x},\hat{y}|y,x)\underbrace{\mathcal{P}(y|x,m)\pi(x,m)}_{\pi(y,x,m)}\\ &= \mathcal{P}(\hat{x},\hat{y}|y,x)\delta_D(y-mx)\pi(x)\pi(m).\end{aligned} \tag{0.11}$$

This was a very successful split: in the second line, we rewrote the joint distribution of y,x,m, and in the third line used that the then appearing conditional distributions are known from the previous steps of the exercise.

Finally, the term to be facilitated is $\mathcal{P}(\hat{x},\hat{y}|y,x)$, but we know it is

$$\mathcal{P}(\hat{x},\hat{y}|y,x) = \mathcal{P}(\hat{x}|x)\mathcal{P}(\hat{y}|y). \tag{0.12}$$

We therefore arrive again at

$$\mathcal{P}(\hat{x},\hat{y},x,y,m) = \mathcal{P}(\hat{x}|x)\mathcal{P}(\hat{y}|y)\delta_D(y-mx)\pi(x)\pi(m), \tag{0.13}$$

which is the same result as Eq. (0.8).

Now we got what we need to compute what we wanted. We want the posterior of m **given the data**[2], so

$$\mathcal{P}(m,x,y|\hat{x},\hat{y}) = \frac{\mathcal{P}(\hat{x},\hat{y},x,y,m)}{\pi(\hat{x},\hat{y})}. \tag{0.14}$$

Of this we know the top. The variables x and y are however nuisance parameters (or latent variables), and we therefore still need to integrate them out:

$$\begin{aligned}\mathcal{P}(m|\hat{x},\hat{y}) &= \int\int \mathcal{P}(m,x,y|\hat{x},\hat{y})\,\mathrm{d}x\,\mathrm{d}y \\ &= \int\int \frac{\mathcal{G}(\hat{x}|x)\mathcal{G}(\hat{y}|y)\delta_D(y-mx)\pi(x)\pi(m)}{\pi(\hat{x},\hat{y})}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int \frac{\mathcal{G}(\hat{x}|x)\mathcal{G}(\hat{y}|mx)\pi(x)\pi(m)}{\pi(\hat{x},\hat{y})}\mathrm{d}x\end{aligned} \tag{0.15}$$

where we have integrated over the delta-function in the third line, which removed one integral.

7. Now we can plug in the actual distributions. We chose flat uniform priors for m and x (replace the π by a 1), and use $\sigma = 1$, which is not a specialization. If we do not have a missing-data problem, then we can omit the prior $\pi(\hat{x},\hat{y})$. It would simply produce an m-independent scaling, which can be absorbed in the normalization. With the specified Gaussians we then have

$$\mathcal{P}(m|\hat{x},\hat{y}) \propto \int \exp\left(-\frac{1}{2}(x-\hat{x})^2\right)\exp\left(-\frac{1}{2}(mx-\hat{y})^2\right)\mathrm{d}x. \tag{0.16}$$

The integral over x has an analytical solution, namely

$$\mathcal{P}(m|\hat{x},\hat{y}) \propto \frac{1}{\sqrt{1+m^2}}\exp\left(-\frac{1}{2}\frac{(\hat{y}-m\hat{x})^2}{1+m^2}\right). \tag{0.17}$$

Plotting this as a function of m leads to a slightly left-skewed posterior.

[2] Sometimes the mistake is made to compute the posterior of the interesting parameters, *given all the rest.* The rest is not only the data, but also the nuisance parameters. The nuisance parameters then end up on the wrong side of the vertical bar. For example, here one would then compute $\mathcal{P}(m|\hat{x},\hat{y},x,y)$, but x,y are random (since latent), so the correct solution is $\mathcal{P}(m,x,y|\hat{x},\hat{y})$, where all nuisance params are on the left, and only data on the right.