Posterior Estimates and Summary Statistics

Having obtained the Bayesian posterior probability distribution $P(\theta|x)$ what can we do with it?

0.1 Marginal distributions

It is often the case that only a subset some of the model parameters are of interest. The other parameters, those whose values we are not interested in, are known as "nuisance" parameters. For example, when analysing a signal in an unknown noise background: there are parameters describing the signal and nuisance parameters describing the noise.

We need to account for the nuisance parameters, but we would also like to remove them from our final result. This can be achieved by integrating out, or *marginalising*, these parameters from the posterior distribution.

We can construct the *1-dimensional marginal posteriors* for each parameter. For example, for the first parameter θ_1 we define

$$P(\theta_1|x) = \int \mathrm{d}\theta_2 \int \mathrm{d}\theta_3 \ldots \int \mathrm{d}\theta_n P(\theta_\mu|x)\,. \tag{1}$$

We can construct the *2-dimensional marginal posteriors* for each pair of parameters. For example, for the first two parameters θ_1 and θ_2 we define

$$P(\theta_1, \theta_2|x) = \int \mathrm{d}\theta_3 \int \mathrm{d}\theta_4 \ldots \int \mathrm{d}\theta_n P(\theta_\mu|x)\,. \tag{2}$$

It is common to arrange and visualise all possible 1 and 2 dimensional marginal posteriors in a *corner plot*, see Fig. 1

In problems with more parameters, say $m + n$ parameters $\{\theta_1, \theta_2, \ldots, \theta_n, \phi_1, \phi_2, \ldots, \phi_m\}$, you can marginalise over any number m of the parameters in the posterior as you want to leave a *n-dimensional marginal posterior*;

$$P(\theta_1, \theta_2, \ldots, \theta_n|x) = \int \mathrm{d}^m\phi\; P(\theta_1, \theta_2, \ldots, \theta_n, \phi_1, \phi_2, \ldots, \phi_m|x)\,. \tag{3}$$

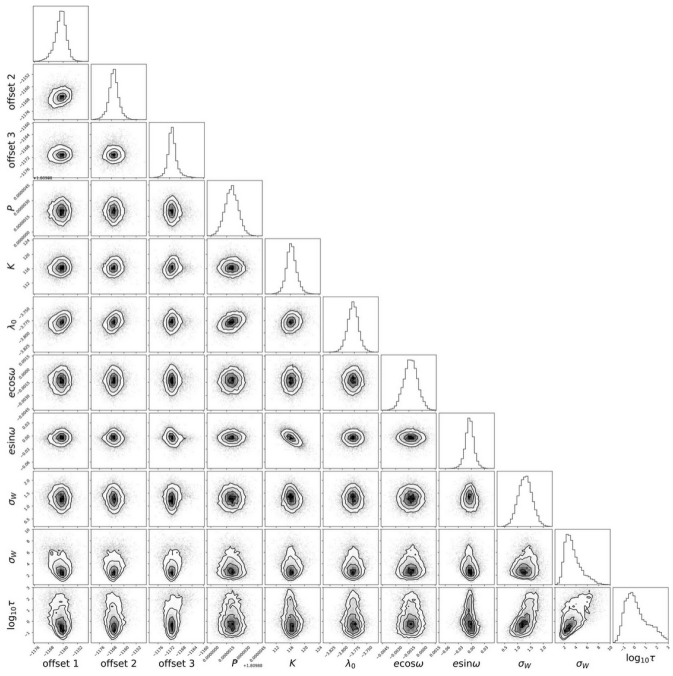

Figure 1: An example corner plot, Ehrenreich *et al.* (2020).

Note, if we marginalise (integrate) out *all* of the parameters in the posterior distribution we are just left with a constant,

$$P(x) = \int \mathrm{d}^n\theta_\mu \; P(\theta_\mu|x)\,. \tag{4}$$

We identify this as the Bayesian evidence, Z. For this reason you will sometimes find the *evidence* referred to by the alternative name *marginal likelihood*.

0.2 Point Estimates

The posterior is the complete description of our state of knowledge (after the experiment) about the model parameters, θ_μ. However, it is often desirable to be able to condense the information contained in the posterior into just a few numbers, or summary statistics.

If the posterior is dominated by a single peak in the parameter space, then it might make

sense to report a single value for our best guess of the model parameter values. A number of these *point estimates* are possible:

1. The posterior mean; e.g. for the first parameter

$$\langle \theta_1 \rangle = \int \mathrm{d}\theta_1 \; \theta_1 P(\theta_1|x) = \int \mathrm{d}^n\theta_\mu \; \theta_1 P(\theta_\mu|x). \tag{5}$$

   This can be calculated from either the full n-dimensional posterior or the 1-dimensional marginal posterior distribution.

2. The posterior mode, known as the *maximum a posteriori* (MAP) estimate $\theta_\mu^{\mathrm{MAP}}$ satisfies

$$\frac{\partial P(\theta_\mu|x)}{\partial \theta_\mu}\Big|_{\theta_\mu=\theta_\mu^{\mathrm{MAP}}} = 0. \tag{6}$$

   This must be calculated from the full n-dimensional posterior; note, that the marginal posterior distributions are in general NOT peaked at the MAP parameters.

3. The median of each 1-dimensional marginal posterior; e.g. for the first parameter this satisfies

$$\int_{-\infty}^{\bar{\theta}_1} \mathrm{d}\theta_1 \; P(\theta_1|x) = \frac{1}{2}. \tag{7}$$

   This must be calculated from the 1-dimensional marginal posterior distributions.

0.3 Credible Intervals

It is worth repeating, the posterior is the complete description of our state of knowledge (after the experiment) about the model parameters, θ_μ, including the uncertainties on these parameters. It is often desirable to be able to condense this information into some kind of "error bar".

A *credible interval* is an range of values within which the model parameters lies with a particular probability level. These are used in a similar to confidence intervals in frequentist inference. In more than 1 dimension we use the alternative name *credible region*.

Even when we have chosen a level (e.g. the 50% or 90% credible intervals are commonly used) the credible interval is not unique (see Fig. 2).

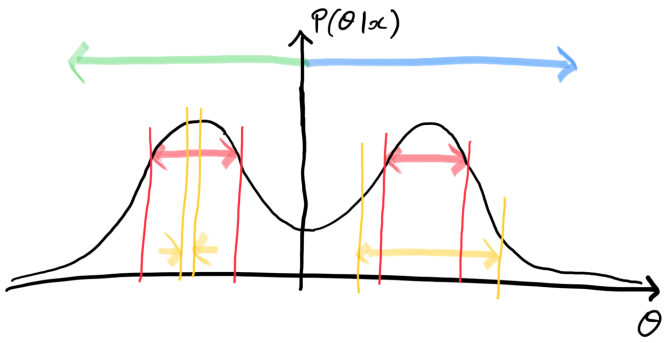

Figure 2: The credible interval is not unique. For this distribution the 4 intervals shown in blue, green, red and yellow are all possible 50% credible intervals.

Several systematic ways of constructing credible intervals that contain a fraction α of the posterior probability are possible:

1. In 1 dimension, choose the *narrowest interval* (or collection of intervals) I that contains a probability mass α.

   This I is *not* invariant under reparametrisation. This flaw can be fixed by changing our notion of width from "narrowest" (i.e. the smallest $\theta_{\rm max} - \theta_{\rm min}$) to use the prior probability measure; i.e. the interval that contains the smallest prior probability mass.

   In multiple dimensions, this can be generalised to the region with the smallest volume that contains a probability mass α.

2. In 1 dimension, choose the *highest density interval* which is the interval (or collection of intervals) $I(f_\alpha) = \{\theta : P(\theta|x) > f_\alpha\}$, where f_α is the largest constant such that $\text{Prob}\big(\theta \in I(f_\alpha)\big) \geq \alpha$.

   In multiple dimensions, this can be generalised using *iso-probability contours* in the sample space to bound the *highest density credible region*; i.e. the region $R(f_\alpha)$ defined in the same way as $I(f_\alpha)$.

3. In 1 dimension, choose the *equal-tailed interval* $\theta_{\rm min} < \theta < \theta_{\rm max}$ with $\text{Prob}(\theta < \theta_{\rm min}) = \text{Prob}(\theta > \theta_{\rm max}) = \alpha/2$.

   In multiple dimensions, this can be applied to each parameter individually using the 1-dimensional marginalised posterior distributions.

0.4 Computing summary statistics

If we are given a the posterior as a function of the model parameters, i.e. $f(\theta) \propto P(\theta|x)$, the summary statistics discussed in Secs. 0.0.2 and 0.0.3 can all be calculated. But they can be difficult to calculate, especially in multiple dimensions.

Finding the MAP point estimate parameters require us to numerically optimise the function $f(\theta)$ which is usually relatively easy. The mean and median point estimate parameters and require us to integrate the function $f(\theta)$ which is usually extremely hard in multiple dimensions. We won't discuss how this is done because, in practice, these summary statistics are normally not calculated from $f(\theta)$ but from *stochastic samples* instead.