9 Parameter estimation

Most practical data analysis does not consider a specific hypothesis in isolation, but rather is concerned with assessing what values of some quantity are consistent with the available information. In terms of Bayesian inference, this means not just calculating the posterior probability (Section 8) of the single hypothesis, effectively that a quantity has a particular value (or that is value is in some range), but calculating a posterior probability distribution (Section 6) over the range of possible values. A posterior distribution can be used to represent the full state of knowledge about any sort of quantity, although it is particularly relevant for parameters that take on numerical values; the task of calculating such a posterior distribution is usually termed parameter estimation. It is useful to describe the basic formalism in the context of a discrete parameter space (Section 9.1), although by far the most common application is to parameters that can take on a continuous range of values (Section 9.2). To avoid duplication only the continuous case is considered explicitly thereafter, although the correspondence between continuous and discrete distributions given in Section 6.2.4 means that any of the subsequent results can be applied to discrete (or categorical) quantities as well.

9.1 Discrete parameter space

Using the notation of Section 6, the general quantity or parameter of interest is taken to be Q and the set of values it could take to be $V = \{v_1, v_2, \ldots\}$. Conditional on data d and background knowledge $\mathcal{K}$, the probability that $Q = v_i \in V$ is given by Bayes's theorem as

$$\mathbb{P}(Q = v_i | d, \mathcal{K}) = \frac{\mathbb{P}(Q = v_i | \mathcal{K}) \, \mathbb{P}(d | Q = v_i, \mathcal{K})}{\mathbb{P}(d | \mathcal{K})}, \tag{99}$$

where, using the terminology of Equation 88, $\mathbb{P}(Q = v_i|\mathcal{K})$ is the prior probability that $Q = v_i$, $\mathbb{P}(d|Q = v_i, \mathcal{K})$ is the likelihood function, and $\mathbb{P}(d|\mathcal{K})$ is the marginal likelihood. Using the law of total probability, this last term can be expressed as

$$\mathbb{P}(d|\mathcal{K}) = \sum_{v_j \in V} \mathbb{P}(Q = v_j | \mathcal{K}) \, \mathbb{P}(d | Q = v_j, \mathcal{K}), \tag{100}$$

where the sum extends over all the elements in V. It is hence possible to re-write Equation 99 as

$$\mathbb{P}(Q = v_i | d, \mathcal{K}) = \frac{\mathbb{P}(Q = v_i | \mathcal{K}) \, \mathbb{P}(d | Q = v_i, \mathcal{K})}{\sum_{v_j \in V} \mathbb{P}(Q = v_j | \mathcal{K}) \, \mathbb{P}(d | Q = v_j, \mathcal{K})}. \tag{101}$$

As Equation 101 holds for each of the different possible values Q could take, this means that Bayes's theorem has been re-cast entirely in terms of distributions.

9.2 Continuous parameter space

Exploiting the correspondence between discrete and continuous distributions given in Section 6.2.4, and adopting the standard notation that a parameter combination is $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ and that the data-set is $\boldsymbol{d} = (d_1, d_2, \ldots)$ then allows the posterior distribution to be written in its standard compact form as

$$\mathbb{P}(\boldsymbol{\theta} | \boldsymbol{d}, \mathcal{K}) = \frac{\mathbb{P}(\boldsymbol{d} | \mathcal{K}) \, \mathbb{P}(\boldsymbol{d} | \boldsymbol{d}, \mathcal{K})}{\int \mathrm{d}\boldsymbol{\theta}' \, \mathbb{P}(\boldsymbol{\theta}' | \mathcal{K}) \, \mathbb{P}(\boldsymbol{d} | \boldsymbol{\theta}', \mathcal{K})}. \tag{102}$$

The form of the denominator is such that $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{K})$ is a correctly normalised over all the possible parameter values, because

$$\begin{aligned}\int \mathrm{d}\boldsymbol{\theta}\, \mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{K}) &= \int \mathrm{d}\boldsymbol{\theta}\, \frac{\mathbb{P}(\boldsymbol{\theta}|\mathcal{K})\, \mathbb{P}(d|\boldsymbol{\theta}, \mathcal{K})}{\int \mathrm{d}\boldsymbol{\theta}'\, \mathbb{P}(\boldsymbol{\theta}'|\mathcal{K})\, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}', \mathcal{K})} \\ &= \frac{\int \mathrm{d}\boldsymbol{\theta}\, \mathbb{P}(\boldsymbol{\theta}|\mathcal{K})\, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K})}{\int \mathrm{d}\boldsymbol{\theta}'\, \mathbb{P}(\boldsymbol{\theta}'|\mathcal{K})\, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}', \mathcal{K})} \\ &= 1, \qquad (103)\end{aligned}$$

as required.

One implication of the above result is that the denominator is irrelevant to the inference encoded in Equation 102, as it is just a normalisation integral (or sum, in the discrete case) that could be computed after undertaking an inference calculation. The important logical content of Equation 102 can be summarised as

$$\underbrace{\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{K})}_{\text{posterior distribution}} \propto \underbrace{\mathbb{P}(\boldsymbol{\theta}|\mathcal{K})}_{\text{prior distribution}} \times \underbrace{\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K})}_{\text{likelihood function}}, \qquad (104)$$

where the constant of proportionality is determined by the requirement that the posterior distribution be normalised. This equation once again emphasises the idea that Bayesian inference is a formalism for learning, but whereas in Section 8.2 the focus was on incorporating new information about a single hypothesis, it is now a way of updating what is known about the likely value(s) of a quantity or parameter. The two distributions that appear in the right-hand side of Equation 104 play distinct roles in the inference formalism, with both the prior distribution (Section 9.3) and the likelihood function (Section 9.4) being subject to different restrictions, although all these reduce to the requirement that the resultant posterior distribution (Section 9.5) must be normaliseable.

9.3 The prior distribution

Just as the prior probability (Section 8.2.1) encodes everything that the assumed background information implies about the likely truth of a particular hypothesis, the prior distribution in a parameter encodes everything that is already known about its plausible value(s).

9.3.1 Prior assignment

The most important distinction between the prior distribution and the posterior distribution is that, whereas the latter is calculated, the former must, in general, be assigned. If there are useful – and quantifiable – prior constraints on a quantity then the source of this assignment is at least clear; but an obvious problem arises when the prior knowledge appears to be minimal or even non-existent. The assignment of prior distributions in such situations is an entire sub-topic of Bayesian inference, and a number of overlapping approaches have been advocated, including:

- **Informative priors:** An informative prior distribution is not motivated by any general statistical or mathematical considerations, but by encoding some external information that is specific to the problem at hand. This might include the results of previous related measurements or some more theoretical understanding about what parameter values are reasaonble.

- **Uninformative priors:** An uninformative (or at least minimally informative) prior distribution is any that is intended to have as minimal influence on the form of the resultant posterior distribution. This hence includes many of the below classes.

- **Objective priors:** An objective prior distribution is such that, with only minimal information about the structure of the problem, is intended to be generic and widely applicable. In most cases objective priors are also uninformative priors.

- **Jeffreys priors:** A Jeffreys prior distribution is motivated by the desideratum that the resultant posterior distribution be invariant under reparameterisation. Any requirement based on the posterior distribution must have some dependence on the form of the likelihood function, which in turn relates to the measurement or data-gathering process; but this implies that the nature (and even existence) of a measurement somehow determines what was known before the measurement was made. So while Jeffreys priors are mathematically elegant, they violate the underlying logic of Bayesian inference.

- **Maximum entropy priors:** A maximum entropy prior distribution is defined as the probability distribution with lowest information content, subject to certain sufficient constraining information. A maximum entropy prior is hence the least informative prior and so allows whatever data is available to exert the maximum influence on the posterior distribution. For a discrete quantity with a finite set of possible values the maximum entropy prior assigns the same probability to each, examples of which include uniform priors assigned to each side of a coin, each face of a (cubic) die, or to each card in a pack. For a parameter that could take on one of a continuous range of values the maximum entropy distribution requires some constraining information: if maximum and minimum values are known then the result is a uniform distribution between these; if the mean and standard deviation are known then the result is a normal distribution (which is a particularly powerful result as it is completely independent of the central limit theorem).

- **Reference priors:** A reference prior on a parameter is defined with reference to a potentially repeatable measurement, the idea being that the information gain after taking an infinite number of such measurements is maximised.

- **Conjugate priors:** A conjugate prior is chosen, on purely mathematical grounds, so that the posterior distribution is of the same functional "family" of distributions as the prior distribution. It is hence dependent on the likelihood, which again implies that it violates some of the underlying logical principles of Bayesian inference. However, in situations where the data are sufficiently informative that the prior distribution has minimal influence on the inference, a conjugate prior can be a useful way of obtaining a closed form posterior distribution.

9.3.2 Prior sensitivity

Whatever prior distribution is assigned to the parameter(s) of interest, it is unlikely to be compelling: other choices would have been reasonable; or, in more human terms, different people could have sensibly assigned different priors. The implication is that, even though the subsequent inference follows the formal Bayesian outlined above, different posterior distributions could result.

This does not imply, however, that the posterior distributions differ *significantly*: if the data are sufficiently informative then the posterior distribution might be similar for all plausible priors; and in the limit of infinitely good data, or large sample size in the case of multiple measurements, the posterior is independent of the prior (assuming certain regularity conditions). Some elegant asymptotic results are possible, but inevitably apply only for the idealised case of large samples of data which can be modelled perfectly.

Of more importance from a practical perspective is to test prior sensitivity by repeating an inference calculation with different (reasonable) prior choices. In the case of parameter estimation this means trying different prior distributions. If the posterior distribution, especially in the subset of parameters of interest, is comparable for the different priors then the inference is robust, and can be used with confidence. But if the results are very different the implication is that, between the background knowledge $\mathcal{K}$ and data $\boldsymbol{d}$, there simply is not enough information to make a substantive statement about the plausible values of the parameters, $\boldsymbol{\theta}$.

9.3.3 Improper priors

Expressing Bayes's theorem in terms of proportionality, as in Equation 104, implies that multiplying the prior distribution $\mathbb{P}(\boldsymbol{\theta}|\mathcal{K})$ by a quantity with no $\boldsymbol{\theta}$ dependence would not affect the final inference. Making the substitution $\mathbb{P}(\boldsymbol{\theta}|\mathcal{K}) \rightarrow \tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K}) = c\,\mathbb{P}(\boldsymbol{\theta}|\mathcal{K})$ for any positive number c would leave the posterior unchanged. This means that there is no practical need to ensure that the prior distribution is properly normalised in this context, and so it is possible to use an unnormalised prior distribution $\tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K})$ that still satisfies $\tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K}) \geq 0$ for all $\boldsymbol{\theta}$, but for which $\int \mathrm{d}\boldsymbol{\theta}\, \tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K}) \neq 1$.

This idea can be extended even further, and it is sometimes possible to operate with a prior distribution that is not just unnormalised, but *unnormaliseable*, with $\int \mathrm{d}\boldsymbol{\theta}\, \tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K}) = \infty$. (Examples of such distributions include an infinitely broad uniform distribution over $\mathbb{R}^N$, for $N \geq 1$, or a power-law of any index over $\mathbb{R}_+$.) This is an improper prior distribution and, while often a mathematically convenient option, requires great care be taken to ensure that the resultant posterior distribution is normaliseable (*i.e.*, proper). Whether this is the case depends on the nature of the data and model, specifically the integral of the product of the (improper) prior distribution and likelihood over the entire parameter space. There are three distinct regimes:

- The integral is zero, *i.e.*, $\int \mathrm{d}\boldsymbol{\theta}\, \tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K})\, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K}) = 0$. The implication is that the data could not possibly have been obtained for any of the allowed parameter values, and so any subsequent inference (and, in particular, the implied posterior distribution) is invalid.
- The integral is finite, *i.e.*, $0 < \int \mathrm{d}\boldsymbol{\theta}\, \tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K})\, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K}) < \infty$. The improper prior is, at least in the context of this likelihood, valid. The correctly normalised posterior distribution is given by
$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{K}) = \frac{\tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K})\, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K})}{\int \mathrm{d}\boldsymbol{\theta}'\, \tilde{\mathbb{P}}(\boldsymbol{\theta}'|\mathcal{K})\, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}', \mathcal{K}).} \tag{105}$$
- The integral is infinite, *i.e.*, $\int \mathrm{d}\boldsymbol{\theta}\, \tilde{\mathbb{P}}(\boldsymbol{\theta}|\mathcal{K})\, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K}) = \infty$. The posterior distribution cannot be normalised and is improper, and the inference is invalid. The combination of background information and data are, between them, insufficient to constrain the parameter values.

The simplest and most common improper prior is an infinitely-broad uniform distribution, of the form $\tilde{\mathbb{P}}(\theta|\mathcal{K}) = 1$ (or any other positive number) in the single parameter case. This is typically an attempt to impose as little prior information on the problem as possible, but is also useful because it provides a more rigorous method of obtaining valid inferences under an improper prior via a limiting process:

1. Define a proper prior that is a uniform distribution between $-a$ and a,

$$\mathbb{P}(\theta|\mathcal{K}) = \mathrm{U}(\theta; -a, a) = \frac{\Theta(\theta - a)\,\Theta(a - \theta)}{2\,a}, \tag{106}$$

where $a > 0$.

2. Calculate the posterior distribution, conditional on a, as

$$\mathbb{P}(\theta|a, \mathcal{K}) = \frac{\Theta(\theta - a)\,\Theta(a - \theta)/(2\,a)\,\mathbb{P}(\boldsymbol{d}|\theta, \mathcal{K})}{\int_{-a}^{a} \mathrm{d}\theta'\,\mathbb{P}(\boldsymbol{d}|\theta', \mathcal{K})/(2\,a)} = \frac{\Theta(\theta - a)\,\Theta(a - \theta)\,\mathbb{P}(\boldsymbol{d}|\theta, \mathcal{K})}{\int_{-a}^{a} \mathrm{d}\theta'\,\mathbb{P}(\boldsymbol{d}|\theta', \mathcal{K})}, \tag{107}$$

which is valid for any likelihood $\mathbb{P}(d|\theta, \mathcal{K})$ provided only that it is non-zero for at least some $-a \leq \theta \leq a$.

3. If the limit of this distribution as $a \to \infty$ exists (and is finite) then the appropriate posterior distribution under the improper uniform prior is

$$\mathbb{P}(\theta|\mathcal{K}) = \lim_{a \to \infty} \frac{\Theta(\theta - a)\,\Theta(a - \theta)\,\mathbb{P}(\boldsymbol{d}|\theta, \mathcal{K})}{\int_{-a}^{a} \mathrm{d}\theta'\,\mathbb{P}(\boldsymbol{d}|\theta', \mathcal{K}),}. \tag{108}$$

(If this limit does not exist or is infinite then the improper prior cannot be used.)

There is nothing special about the use of a uniform prior distribution here – a normal distribution would also work in much the same way – although it is appealing that, other than the integration limits, the prior does not appear in the final expression. The important point here is that the limiting procedure follows the inference calculation – the limit of a ratio is not, in general, equal to the ratio of limits.

9.4 The likelihood function

Treated as a probability distribution, the likelihood function $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K})$ is equivalent to the sampling distribution over all the possible data that could have been produced given the parameter values $\boldsymbol{\theta}$ and the background information $\mathcal{K}$. As such, this should be a normalised distribution in $\boldsymbol{d}$, which is equivalent to saying that the simulation of such data-sets given $\boldsymbol{\theta}$ and $\mathcal{K}$ is fully defined.

The likelihood function is not, however, a distribution in $\boldsymbol{\theta}$, so the integral $\int \mathrm{d}\boldsymbol{\theta}\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K})$ is not, in general, finite. The fact that this integral sometimes *is* finite, most obviously when the data places strong constraints on the parameters, can be misleading: it is then tempting to devise inference procedures that depend only the likelihood; but such an approach cannot form a part of any generally-applicable inference formalism.

The central role of the likelihood function is not unique to Bayesian inference, and in particular maximum likelihood methods are (as the name suggests) also based on this link between parameters and data. Operationally, any such likelihood-based methods require a specific form for the likelihood function, but this must be obtained on a case-by-case basis. Determining the stochastic link between the parameters and the data is often the main challenge of setting up a parameter estimation problems.

9.4.1 Sufficient statistics

Bayesian inference is, especially when compared to classical statistics, highly prescriptive – there is no choice about what sort of estimator to use, whether it is more important to minimize variance or bias, what confidence level to adopt, *etc.* While there is sometimes ambiguity in what prior distribution(s) to use, the structure of the link between the prior and data and the final inferences is fixed by Bayes's theorem.

One aspect of this, known as the likelihood principle, is that data can only enter the inferential scheme through the likelihood: it is only the probability of obtaining the observed data that is relevant, not the values of any statistics or estimators that might be computed from the data. That said, there is still a potential role for statistics, defined here as any quantity (or quantities) $\boldsymbol{s}$ that can be calculated from the data according to some function $\boldsymbol{s}(\boldsymbol{d})$. Common examples of statistics, most obviously defined for the case that the data consist of a number of repeated or similar measurements, include the mean, median, minimum and maximum of the sample.

It is certainly possible operationally to replace the data with some statistic calculated from it, and hence obtain a posterior distribution of the form

$$\tilde{\mathbb{P}}(\boldsymbol{\theta}|\boldsymbol{s},\mathcal{K}) \propto \mathbb{P}(\boldsymbol{\theta}|\mathcal{K})\,\mathbb{P}(\boldsymbol{s}|\boldsymbol{\theta},\mathcal{K}), \tag{109}$$

where it is assumed that the sampling distribution $\mathbb{P}(\boldsymbol{s}|\boldsymbol{\theta},\mathcal{K})$ can be calculated from knowledge of $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta},\mathcal{K})$ and the mapping $\boldsymbol{s}(\boldsymbol{d})$. In general the resultant posterior distribution will differ from $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d},\mathcal{K})$, but will presumably encode at least some of the relevant information. If, however, the ratio $\mathbb{P}(\boldsymbol{s}|\boldsymbol{\theta},\mathcal{K})/\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta},\mathcal{K})$ is independent of the value of $\boldsymbol{\theta}$ then all the relevant information about $\boldsymbol{\theta}$ is encoded in $\boldsymbol{s}(\boldsymbol{d})$, which is hence termed a sufficient statistic.

It is also possible to use insufficient statistics (*i.e.*, any statistic which is not sufficient) to perform sub-optimal inference, provided only that the sampling distribution of the statistics, $\mathbb{P}(\boldsymbol{s}|\boldsymbol{\theta},\mathcal{K})$ is known. The reason this is sub-optimal is that at least some of the information contained in the data-set must have been lost. This sacrifice could easily be justified if, in simplifying the problem (most obviously by reducing a large amount of data to a small number of statistics), it is made more tractable. This is particularly true for inference in a data-rich scenario: if the "true" posterior for the full data would be far narrower than needed for the problem at hand, using even significantly sub-optimal statistics could be a very reasonable practical choice. Obtaining a good answer quickly is generally preferable to a perfect but late solution.

Another, similar, justification for a sub-optimal approach would be if the exact sampling distribution is unknown. It might be that a statistic is far less sensitive to the uncertainty in the sampling distribution than the data – a classic example is the median of a set of measurements. Seen in this way, it allows the sub-field of robust estimation to be put in a Bayesian context.

9.5 The posterior distribution

The full outcome of a parameter inference calculation is the full joint posterior distribution in all the parameters. This is sometimes numerically challenging to calculate, but also very powerful as it whatever uncertainty there is in the final result is fully encoded. It might be a reasonable mathematical approximation to report the distribution in the form of a few important numbers (*e.g.*, the mean and the variance) or a set of samples drawn from the posterior (Section 7). In all cases these should be seen simply as mathematical approximations to the full distribution; they are only meaningful to the degree that they allow the full distribution to be reconstructed.

An implication of the above results is that the normalisation of the posterior is generally unimportant when doing parameter estimation. It is only the dependence of the posterior on the parameter values that matters; multiplying it by a constant factor would not change any subsequent inferences. Alternatively, the prior could be calculated with any arbitary (positive) choice of normalisation and then multiplied through by its inverse to obtain a correctly normalised prior. Indeed, this is what the denominator in Equation 103 effectively does.

9.5.1 The marginal posterior distribution

If some of the parameters are not of inherent interest, but have been included because they are necessary to describe the data-generation process, it might be the marginal posterior distribution in a subset of parameters that should be calculated (*cf.* Section 6.3.3). This is most clear in the case of "nuisance" parameters (*e.g.*, a calibration constant or noise level in a physical measurement), but can also be the case when some nominally meaningful parameters are not of interest in a particular context. Splitting the parameter vector as $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\phi})$, where $\boldsymbol{\psi}$ are the parameters of interest and $\boldsymbol{\phi}$ are the nuisance parameters, the marginal posterior distribution in the former is

$$\mathbb{P}(\boldsymbol{\psi}|\boldsymbol{d}, \mathcal{K}) = \int \mathrm{d}\boldsymbol{\phi} \, \mathbb{P}(\boldsymbol{\psi}, \boldsymbol{\phi}|\boldsymbol{d}, \mathcal{K}), \tag{110}$$

where $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{K}) = \mathbb{P}(\boldsymbol{\psi}, \boldsymbol{\phi}|\boldsymbol{d}, \mathcal{K})$ is the full posterior distribution.

Importantly, the marginal posterior distribution in the parameters of interest, $\boldsymbol{\psi}$, is not, in general, equal to the posterior distribution that would have been obtained by setting the other parameters $\boldsymbol{\phi}$ to some estimated value $\hat{\boldsymbol{\phi}}$ and then calculating $\mathbb{P}(\boldsymbol{\psi}|\boldsymbol{d}, \hat{\boldsymbol{\phi}}, \mathcal{K})$. Except in somewhat contrived cases the result of this procedure is to produce a posterior distribution in $\boldsymbol{\psi}$ that is more sharply peaked than that given in Equation 110 above. It might seem that tighter parameter constraints are a good thing, but not if they come about as a result of assuming knowledge that one does not really have. This emphasises that the formalism described above is a method for incorporating all sources of uncertainty in a parameter estimation calculation: all the relevant information is included; but no uncertainties are ignored.

9.6 Model mis-specification

All inferential results obtained using the above formalism are conditioninal on the assumed background knowledge $\mathcal{K}$, which implicitly specifies the model for how the data $\boldsymbol{d}$ is linked to the parameters $\boldsymbol{\theta}$. An obvious possibility that some aspects of $\mathcal{K}$, and specifically the form of $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K}$, are incorrect. The most important resultant difficulty is that parameter estimation does not test in any way whether this is the case: provided only that there are some parameter values $\boldsymbol{\theta}$ such that $\mathbb{P}(\boldsymbol{\theta}|\mathcal{K}) \, \mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K}) > 0$ the result will be an apparently reasonable posterior distribution. But this is essentially meaningless if even the best-fitting parameter values [*i.e.*, those which give the highest values of $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{K})$] are inconsistent with the data. Taken in isolation this is effectively a warning about unthinkingly doing parameter estimation without some level of model checking; but it also motivates combining parameter inference with higher-level inference of the models themselves.

9.7 Parameter estimation "recipe"

The above is largely a set of statements about the rules and limitations of Bayesian parameter inference, but they are only useful as guidance for a practical formalism. So an effective distillation of these points is a "recipe" for how to approach a parameter estimation calculation. This is not to suggest that this is the *only* way to approach such problems, but this is a potentially useful starting point if it is not immediately obvious how to proceed.

1. Identify the parameters that are to be constrained.
2. List whatever background/prior information about these parameters can potentially be assumed.
3. Identify the data which will (or might) provide information about the parmeters.
4. Determine the stochastic aspects of the data model (often the observation/measurement process) and hence the write down or derive the likelihood.
5. Calculate the unnormalised posterior distribution.
6. Normalise the posterior distribution.
7. Marginalise out irrelevant nuisance parameters.
8. Calculate summary statistics for the posterior distribution.

Learning objectives

After studying this section of the module you should be able to:

- Formulate parameter estimation problems using Bayes's theorem to find the posterior distribution in the parameters of interest.
- Incorporate unknown "nuisance" parameters in parameter estimation problems, using marginalisation to then focus on the parameters of interest.
- Work with improper priors and unnormalised posteriors as a "short hand", but remaining aware that they are really the result of taking limits.
- Use posterior summary statistics (point estimates, credible intervals, *etc.*) judiciously to report the results of Bayesian parameter inference.