10 Model comparison

The basic application of Bayes's theorem to inference (Section 8) shows how to evaluate the probability that a single hypothesis is true (or false); extending this to a set of competing hypotheses, or models, is termed model comparison. In general, the posterior (and prior) distribution over the possible models is a categorical distribution (Section 6.2.1). Taking this approach ensures that the formalism is applicable to completely distinct models that do not share any conceptual features, other than making predictions for the data at hand.

As defined above, model comparison could refer to the inference of any categorical or discrete distribution, although a common convention is to apply it primarily to potentially complicated models that have undetermined internal parameters, values for which are required to define and evaluate the likelihood. Examples of such model comparison problems are:

- Given a measured time series of temperature measurements a climate scientist wants to assess the long-term trends: one (null) model is that the temperature is constant, the value of which is unspecified; an alternative model is that there is a linear trend, which implies both an offset and trend which need to be specified; there might even be motivation for considering further models with more extreme quadratic or exponential trends.
- An astronomer takes a (noisy) image of the night sky in the hope of seeing a comet that has been predicted to appear: one (null) model is that there is no source detected; the obvious alternative model is that the comet is detected, in which case it must be at certain position and have a certain flux (*i.e.*, brightness); there is also the possibility of multiple sources being present.
- A gambler wants to assess if a flipped coin can be trusted: the (null) model is that it is fair; the (obvious) alternative is that it is biased, in which case it could be anything from slightly-weighted to double-headed (or double-tailed).
- In testing the reliability of a manufacturing process someone raises the possibility that the failures might be correlated: the (null) model is that each run is independent, but with unspecified failure probability; the alternative model is that something in the process (*e.g.*, a loose material) means that a failure is more likely if the previous item was flawed, in which case there must be separate failure probabilities after a success and after a failure.

10.1 Formalism

The starting point for model comparison is that there is are $N \geq 2$ models, $M_1, M_2, \ldots, M_N$, under consideration to explain some phenomenon. If there were only $N = 1$ model under consideration then the task becomes the inherently non-Bayesian problem of hypothesis-testing (although setting up this problem from a Bayesian perspective can yield some interesting results).

The assumed background knowledge $\mathcal{K}$ is taken to include the set of models (so that they do not have to enumerated explicitly), as well as defining associated prior probabilities $\mathbb{P}(M_i|\mathcal{K}) \geq 0$ (for $i \in \{1, 2, \ldots, N\}$). These must together define a (categorical) distribution, so that $\sum_{i=1}^{N} \mathbb{P}(M_i|\mathcal{K}) = 1$, which is equivalent to asserting that the correct model is included in this list. In practice this is not necessarily the case – and arguably is never the case (*cf.* the often repeated assertion, generally attributed to George Box, that "All models are wrong but some are useful."). This is quite distinct from the inference of a discrete (numerical) parameter or continuous parameter estimation, in which all the possibilities can be included in the calculation (even if an infinite set of values is possible *a priori*). Conversely, there is no way of describing or even enumerating all the possible physical models that might describe the laws of nature. There is always the possibility of an unknown model that nobody has yet thought of. In this sense, the formalism for model comparison described here applies to the reduced problem of assessing the *relative* merits of a set of models in the context of certain relevant information (*i.e.*, the background knowledge, and possibly some data), which includes the assertion that one of them is correct.

The posterior probability that model M_i is correct in light of some data $\boldsymbol{d}$ (along with the backgroun knowledge $\mathcal{K}$) is given by Bayes's theorem and the law of total probability as

$$\mathbb{P}(M_i|\boldsymbol{d}, \mathcal{K}) = \frac{\mathbb{P}(M_i|\mathcal{K})\,\mathbb{P}(\boldsymbol{d}|M_i)}{\sum_{j=1}^{N} \mathbb{P}(M_j|\mathcal{K})\,\mathbb{P}(\boldsymbol{d}|M_j)}, \tag{111}$$

where $\mathbb{P}(\boldsymbol{d}|M_i)$ is the likelihood under mdoel M_i, which is assumed to be sufficiently prescriptive that the background information $\mathcal{K}$ can be omitted.

Even if the list of models under consideration is complete, there is often minimal prior information about which of the models would be preferred. In such situations of indifference it makes sense to assign a prior of $\mathbb{P}(M_i|\mathcal{K}) = 1/N$ for all models, in which case the above posterior simplifies to be

$$\mathbb{P}(M_i|\boldsymbol{d}, \mathcal{K}) = \frac{\mathbb{P}(\boldsymbol{d}|M_i)}{\sum_{j=1}^{N} \mathbb{P}(\boldsymbol{d}|M_j)}. \tag{112}$$

This means that the posterior odds ratio between any pair of models (say M_i and M_j) is just give by the Bayes factor (Section 8.3),

$$B_{i,j} = \frac{\mathbb{P}(\boldsymbol{d}|M_i)}{\mathbb{P}(\boldsymbol{d}|M_j)}, \tag{113}$$

which explains why Bayes factors are often quoted in a model comparison scenario. To the degree that there is no prior reason to prefer one model to another, the Bayes factor(s) between any pair of models is a complete model comparison summary.

10.1.1 The marginal likelihood

If the i'th model M_i has no undetermined parameters then $\mathbb{P}(\boldsymbol{d}|M_i)$ is just a straightforward and well-specified likelihood. If, however, model M_i has parameters $\boldsymbol{\theta}_i = (\theta_{i,1}, \theta_{i,2}, \ldots)$, whose values are not specified precisely by the model itself (or other external information) then the law of total probability implies that the likelihood under the model is

$$\mathbb{P}(\boldsymbol{d}|M_i) = \int \mathrm{d}\boldsymbol{\theta}_i\, \mathbb{P}(\boldsymbol{\theta}_i|M_i)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}_i, M_i), \tag{114}$$

where $\mathbb{P}(\boldsymbol{\theta}_i|M_i)$ is the parameter prior in the context of this model. One way of thinking about this expression is as the average of the likelihood, $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}_i, M_i)$, over the entire parameter space, weighted by the prior, $\mathbb{P}(\boldsymbol{\theta}_i|M_i)$. Another way to understand this construction is that, in the case that a model has unspecified parameters, sampling from $\mathbb{P}(\boldsymbol{d}|M_i)$ could be done using a two-step process: first draw $\boldsymbol{\theta}_i$ from the prior $\mathbb{P}(\boldsymbol{\theta}_i|M_i)$; and then draw data from $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}_i, M_i)$, now acting as a sampling distribution.

The quantity defined in Equation 161 is most commonly referred to as the “marginal likelihood”, although it is also known as the “model-averaged likelihood” or, particularly in physics and astronomy, as the “(Bayesian) evidence”. This quantity is also identical to the normalising denominator in Bayes’s theorem as applied to parameter estimation (Section 9). In the parameter estimation setting this integral can be ignored, as it is sufficient to be able to evaluate the posterior ratio of any two parameter values; this comes about as, effectively, the model is assumed to be true as part of the background information. Here, however, the marginal likelihood is an absolutely vital ingredient in the calculation because it *is* the relevant likelihood. That means that the parameter prior $\mathbb{P}(\boldsymbol{\theta}_i|M_i)$ must be a properly-normalised distribution; there is no option of adopting an improper prior for the parameters under the model.

A corollary is that a model is only fully specified if the prior distribution of its parameters are given. It is formally impossible – and in practice difficult (see Section 13.4) – to perform model comparison between models with unspecified parameter priors. Aside from algebraic checking, one way to assess whether a model is fully specified is to establish that it can be simulated from fully (*i.e.*, from the prior and then the likelihood) to make mock data.

A further implication is that the results of model comparison depend on the form of the parameter priors adopted. In many cases there is no compelling prior distribution (*e.g.*, there might be reason to believe a fitting curve is quadratic, but that does not in itself compel any prior choice for the three relevant polynomial coefficients). In model comparison, however, there is almost always significant prior sensitivity. In particular, in the limit of infinitely broad parameter priors the marginal likelihood of a model almost always tends to zero (the obvious exception being a case in which the parameter in question is irrelevant). In these sorts of situations one should at very least repeat the calculation using different generic priors to establish an approximate level of prior sensitivity. If it is shown that the model comparison results differ greatly for different apparently plausible parameter prior choies, then either i) the question is ill-posed (*i.e.*, under-specified) in the first place, and has no quantitative answer or ii) there is actually more information available which should be included to better specify the model(s).

10.2 Occam’s razor

The model comparison formalism described above is mathematically unambiguous, but it is perhaps not immediately obvious from this what represents a “good” model for certain data and what is a “bad” model. The fact that it is the marginal likelihood which links a model to the data in Equation 158 implies it is only important how well a model predicts the observed data, which can be seen as an implication of the likelihood principle. A corrollary of this is that there is no direct dependence on a number of parameters the model has.

There are several distinct reasons why a model could have a low marginal likelihood:

- The model cannot explain the data for *any* of its possible parameter values;
- The prior density is only high in region(s) of parameter space for which the likelihood is low (*i.e.*, fit to the data is bad);
- The parameter prior is so broad that it includes large regions of parameter space for which the likelihood is low; the requirement that the prior be unit-normalised then implies that the prior density is low everywhere, and in particular where the likelihood is high.

This last point is effectively a natural implementation of “Occam’s razor”, the idea that an explanation (for anything!) should be regarded as inferior if it is unnecessarily complicated. This is one way of expressing the common wisdom that, given two models that explain the data equally “well”, the model that is in some sense “simpler” is preferable. Unfortunately, these concepts of complexity and simplicity are potentially ambiguous; Bayes’s theorem avoids any such semantic problems by naturally implementing such principles mathematically.

This is clearest in the case that two models have the same parameter(s) (*i.e.*, $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}$) but different parameter priors [*i.e.*, $\mathbb{P}(\boldsymbol{\theta}|M_1) \neq \mathbb{P}(\boldsymbol{\theta}|M_2)$]. If the prior in the second model is broader than that in the first then, provided both priors cover the region(s) of parameter space with high likelihood, the first model will (correctly) be preferred, as it predicts the data better.

An obvious way in which one model would be more complicated than another is if it has a higher number of free parameters. A generic example is polynomial fitting to noisy data: as the order of the polynomial approaches the number of data points the fit becomes better but at the expense of over-fitting; if there are more coefficients than data points then the situation is degenerate, as there is potentially an infinite variety of models that fit the data well. The Bayesian approach to such a problem naturally penalises the higher-order models as the parameter space is so much larger that the density of the normalised posterior is lower, although it is actually the lack of predicitivity of the more complicated model that drives its decreased marginal likelihood.

10.2.1 Irrelevant extra parameters

An extreme case of the above situation is that in which a model is extended by adding an extra parameter that is completely irrelevant to the problem. An example might be in fitting climate data, where one model has a single parameter, a coefficient relating temperature change to carbon dioxide levels, and a second model has an additional parameter, the the rate at which buses pass a certain stop in London. The presence of an extra parameter in the second model implies it is more complicated; and, given that it makes the same predictions as the first model and so cannot fit the data any better, an informal application of Occam’s razor suggests the second model should be regarded as worse.

To assess this situation quantitatively requires evaluating the marginal likelihood for the two models, M_1 and M_2, which in turn requires priors for the parameters, $\boldsymbol{\theta}_1 = \theta_1$ and $\boldsymbol{\theta}_2 = (\theta_1, \theta_2)$. A corollary of the second parameter’s irrelevance is that the prior distribution in the two parameters is separable and so $\mathbb{P}(\theta_1, \theta_2|M_2) = \mathbb{P}(\theta_1|M_2)\,\mathbb{P}(\theta_2|M_2)$. It is also reasonable to assume that the prior on the first parameter is the same in two models, so that $\mathbb{P}(\theta_1|M_1) = \mathbb{P}(\theta_1|M_2)$. Under these assumptions the marginal likelihood for the second model can be calculated as

$$\begin{aligned}
\mathbb{P}(\boldsymbol{d}|M_2) &= \int \mathrm{d}\theta_1 \int \mathrm{d}\theta_2\, \mathbb{P}(\theta_1, \theta_2|M_2)\, \mathbb{P}(\boldsymbol{d}|\theta_1, \theta_2, M_2) \\
&= \left[\int \mathrm{d}\theta_1\, \mathbb{P}(\theta_1|M_1)\mathbb{P}(\boldsymbol{d}|\theta_1)\right] \left[\int \mathrm{d}\theta_2\, \mathbb{P}(\theta_2|M_2)\right] \\
&= \int \mathrm{d}\theta_1\, \mathbb{P}(\theta_1|M_1)\, \mathbb{P}(\boldsymbol{d}|\theta_1) \\
&= \mathbb{P}(\boldsymbol{d}|M_1), \qquad (115)
\end{aligned}$$

which is, perhaps unexpectedly, the marginal likelihood for the first model. The addition of a completely irrelevant parameter to a model makes no difference to any model-level inference results.

This result is seemingly in contradiction with Occam's razor. It seems that the second model, with its irrelevant parameter, ought to be "worse" (in some sense) than the first model; but the key point is that M_2 is no less predictive than M_1. What actually penalises a model is if, *a priori*, the observed data are surprising given the model, whereas here the two models make identical predictions for the data. Put another way, the addition of extra parameters to a model only penalises it if they result in the model making (typically) bad predictions.

So how can this result be reconciled with Occam's razor? The answer is that models with unmotivated extra parameters should be assigned lower prior probabilities than those with parameters that have some connection to the phenomenon being investigated. The equations of Bayesian inference can only ever reveal probabilities calculated conditional on certain background information; it is up to statisticians and researchers to assign that information in a sensible way.

10.3 Comparison of nested models

It is often the case that the two models under comparison are related, and specifically that the first model, M_1, is a simpler version of the second, M_2. In this case M_1 and M_2 are referred to as nested models because the parameter space of M_1, defined by $\boldsymbol{\theta}_1$, is "nested" inside the parameter space of M_2 covered by $\boldsymbol{\theta}_2$. Within the context of M_2 it is possible to recover M_1 can by restricting $\boldsymbol{\theta}_2$ to certain values (often zero). Examples of nested models are:

- A common situation in physics and astronomy is that is two models are invoked to explain an energy spectrum (which is defined as the strength of emission as a function of wavelength, frequency or energy): M_1 is a smooth continuum-only function of wavelendth; M_2 is there is both the continuum emission and an additional sharp emission feature/line. If the strength or flux of the line is set to zero then $M_2 \to M_1$.

- In a regression or curve-fitting problem there could be a comparison between polynomials of different order: M_1 is a linear model; M_2 is quadratic. If the coefficient of the highest order term is set to zero then $M_2 \to M_1$.

- One flexible option in modelling a probability distribution is to adopt a mixture model in which the density is represented as a wieghted sum of (often Gaussian) kernels: M_1 might be a single kernel; M_2 could be two kernels centred at different locations. If the weight of the second kernel is set to zero then $M_2 \to M_1$.

While such situations are reasonably common, the nesting of models has no fundamental significance in Bayesian inference, in so far as the general formalism of model comparison still applies. But if the models are neseted then some of the sensitivity of the model comparison results to the form of the parameter priors can be decreased. Specifically, if the parameter priors on the common parameters are assumed to be the same then there is the potential for cancellation when calculating the Bayes factor between two models, and hence increased robustness to prior choice. In this situation there is also some calculational simplification, with the integration burden reduced.

In order to explore this behaviour quantitatively, a precise definition of nesting needs to be adopted. In order for one model M_1 (with parameters $\boldsymbol{\theta}_1$) to be considered nested within another, M_2 (with parameters $\boldsymbol{\theta}_2$), the following criteria must be satisfied:

1. The parameters which describe model M_2 must be separable into two disjoint sub-sets: the parameters $\boldsymbol{\theta}_1 = \boldsymbol{\phi} = (\phi_1, \phi_2, \ldots)$ which are common to both M_1 and M_2; and the remaining parameters $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots)$ which are unique to M_2. Hence the full list of parameters for the simpler model is $\boldsymbol{\theta}_1 = (\boldsymbol{\phi}) = (\phi_1, \phi_2, \ldots)$ and for the more complicated model is $\boldsymbol{\theta}_2 = (\boldsymbol{\phi}, \boldsymbol{\psi}) = (\phi_1, \phi_2, \ldots, \psi_1, \psi_2, \ldots)$.

2. There must be a combination of the second model's parameters, $\boldsymbol{\psi} = \boldsymbol{\psi}_1$, such that the second model becomes identical to the first model for all values of $\boldsymbol{\phi}$.

3. The prior distribution of the second model's parameters must be separable into independent priors on $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, *i.e.*, $\mathbb{P}(\boldsymbol{\theta}_2|M_2) = \mathbb{P}(\boldsymbol{\phi}, \boldsymbol{\psi}|M_2) = \mathbb{P}(\boldsymbol{\phi}|M_2)\,\mathbb{P}(\boldsymbol{\psi}|M_2)$, with $\mathbb{P}(\boldsymbol{\psi}_1|M_2) > 0$.

4. The prior distribution of $\boldsymbol{\phi}$ in the two models must be the same, *i.e.*, $\mathbb{P}(\boldsymbol{\phi}|M_1) = \mathbb{P}(\boldsymbol{\phi}|M_2)$.

It is a useful conceit to add the second model's parameters to those of the first model, fixing their values at $\boldsymbol{\psi} = \boldsymbol{\psi}_1$. In this case the parameter prior on the first model becomes

$$\mathbb{P}(\boldsymbol{\phi}, \boldsymbol{\psi}|M_1) = \mathbb{P}(\boldsymbol{\phi}|M_1)\,\delta_{\mathrm{D}}(\boldsymbol{\psi} - \boldsymbol{\psi}_1). \tag{116}$$

With both models thus cast in terms of the full parameter set $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\psi})$, the parameter likelihoods, $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}, M_1)$ and $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}, \boldsymbol{\psi}, M_2)$, also have the same form and can both be written as $\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}, \boldsymbol{\psi})$, without explicit reference to which model has been adopted. (If $\boldsymbol{\psi} = \boldsymbol{\psi}_1$ then the likelihood is as appropriate for M_1 or M_2; otherwise it is only valid for M_2.)

The Savage–Dickey density ratio

The implications of some data $\boldsymbol{d}$ for two nested models are, from above, summarised by the Bayes factor,

$$B_{1,2} = \frac{\mathbb{P}(\boldsymbol{d}|M_1)}{\mathbb{P}(\boldsymbol{d}|M_2)} = \frac{\int \mathbb{P}(\boldsymbol{\phi}'|M_1)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}', M_1)\,\mathrm{d}\boldsymbol{\phi}'}{\int \mathbb{P}(\boldsymbol{\phi}'', \boldsymbol{\psi}''|M_2)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}'', \boldsymbol{\psi}'', M_2)\,\mathrm{d}\boldsymbol{\phi}''\,\mathrm{d}\boldsymbol{\psi}''}, \tag{117}$$

where care has been taken to avoid confusing the distinct dummy integration variables in the numerator and the denominator. Applying the above criteria for nested models, this can be rewritten as

$$B_{1,2} = \frac{\int \mathbb{P}(\boldsymbol{\phi}'|M_1)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}', \boldsymbol{\psi}_1)\,\mathrm{d}\boldsymbol{\phi}'}{\int \mathbb{P}(\boldsymbol{\phi}''|M_1)\,\mathbb{P}(\boldsymbol{\psi}''|M_2)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}'', \boldsymbol{\psi}'')\,\mathrm{d}\boldsymbol{\phi}''\,\mathrm{d}\boldsymbol{\psi}''}. \tag{118}$$

Multiplying both numerator and denominator by $\mathbb{P}(\boldsymbol{\psi}_1|M_2)$ then gives

$$\begin{aligned} B_{1,2} &= \frac{1}{\mathbb{P}(\boldsymbol{\psi}_1|M_2)} \frac{\int \mathbb{P}(\boldsymbol{\phi}'|M_1)\,\mathbb{P}(\boldsymbol{\psi}_1|M_2)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}', \boldsymbol{\psi}_1)\,\mathrm{d}\boldsymbol{\phi}'}{\int \mathbb{P}(\boldsymbol{\phi}''|M_1)\,\mathbb{P}(\boldsymbol{\psi}''|M_2)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}'', \boldsymbol{\psi}'')\,\mathrm{d}\boldsymbol{\phi}''\,\mathrm{d}\boldsymbol{\psi}''} \\ &= \frac{1}{\mathbb{P}(\boldsymbol{\psi}_1|M_2)} \int \frac{\mathbb{P}(\boldsymbol{\phi}'|M_1)\,\mathbb{P}(\boldsymbol{\psi}_1|M_2)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}', \boldsymbol{\psi}_1)}{\int \mathbb{P}(\boldsymbol{\phi}''|M_1)\,\mathbb{P}(\boldsymbol{\psi}''|M_2)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}'', \boldsymbol{\psi}'')\,\mathrm{d}\boldsymbol{\phi}''\,\mathrm{d}\boldsymbol{\psi}''}\,\mathrm{d}\boldsymbol{\phi}', \end{aligned} \tag{119}$$

where the fact that there are different integration variables in the numerator and denominator has been exploited. By comparison, the normalised parameter posterior for $(\boldsymbol{\phi}, \boldsymbol{\psi})$ under the second model is

$$\mathbb{P}(\boldsymbol{\phi}, \boldsymbol{\psi}|\boldsymbol{d}, M_2) = \frac{\mathbb{P}(\boldsymbol{\phi}|M_1)\,\mathbb{P}(\boldsymbol{\psi}|M_2)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}, \boldsymbol{\psi}_1)}{\int \mathbb{P}(\boldsymbol{\phi}''|M_1)\,\mathbb{P}(\boldsymbol{\psi}''|M_2)\,\mathbb{P}(\boldsymbol{d}|\boldsymbol{\phi}'', \boldsymbol{\psi}'')\,\mathrm{d}\boldsymbol{\phi}''\,\mathrm{d}\boldsymbol{\psi}''} \tag{120}$$

for all values of $\boldsymbol{\psi}$ (and, in particular, for $\boldsymbol{\psi} = \boldsymbol{\psi}_1$). Hence

$$B_{1,2} = \frac{\int \mathbb{P}(\boldsymbol{\phi}', \boldsymbol{\psi}_1 | \boldsymbol{d}, M_2) \, \mathrm{d}\boldsymbol{\phi}'}{\mathbb{P}(\boldsymbol{\psi}_1 | M_2)}. \tag{121}$$

But integrating a posterior distribution over some sub-set of the parameters is simply marginalisation, and in this case the integral in the numerator reduces to the marginalised posterior $\mathbb{P}(\boldsymbol{\psi}_1 | \boldsymbol{d}, M_2)$. This leads to the final result that

$$B_{1,2} = \frac{\mathbb{P}(\boldsymbol{\psi}_1 | \boldsymbol{d}, M_2)}{\mathbb{P}(\boldsymbol{\psi}_1 | M_2)} = \left. \frac{\mathbb{P}(\boldsymbol{\psi} | \boldsymbol{d}, M_2)}{\mathbb{P}(\boldsymbol{\psi} | M_2)} \right|_{\boldsymbol{\psi} = \boldsymbol{\psi}_1}. \tag{122}$$

This is termed the Savage–Dickey density ratio (SDDR) as it was first published by Dickey (1971) who, in turn, attributed the result to L. J. Savage. Using the SDDR makes it possible to perform rigorous model comparison between nested models without the need for evaluating the potentially challenging multi-dimensional integrals over the two models' parameters (although the partial marginalisation over $\boldsymbol{\phi}$ is still required).

It may seem surprising that such an apparently complicated ratio of integrals has such a simple expression, but the intepretation of this final result is at least clear. The simpler nested model is preferred if, within the context of the second model, the new data result in an increased probability that $\boldsymbol{\psi} = \boldsymbol{\psi}_1$; the more complicated model is preferred if the data imply that $\boldsymbol{\psi} = \boldsymbol{\psi}_1$ is disfavoured relative to its prior probability under M_2.

10.4 Model comparison in the absence of well-motivated parameter priors

The above formalism for Bayesian model comparison follows unambiguously from Bayes's theorem, but relies on a unit-normalised prior distribution $\mathbb{P}(\boldsymbol{\theta} | M)$ being defined for the parameter(s) $\boldsymbol{\theta}$ of any model M. Further, the model comparison results are strongly dependent on the form of $\mathbb{P}(\boldsymbol{\theta} | M)$, so simply adopting arbitrary priors will produce essentially meaningless results. Inverting this argument then leads to the conclusion that models with unspecified parameters that do *not* have well-motivated priors cannot be included in a self-consistent Bayesian model comparison formalism.

And yet model comparison has effectively been used in many cases where such priors are not available. In particular, it has often proved possible to reject a model because it cannot explain the data for any of its permissable parameter values. One way of looking at this is that even if the best possible parameter prior (a delta function at the best fit model) cannot explain the data, then a very low upper bound can be placed on the marginal likelihood. While this has something of a flavour of an hypothesis test, the key point is that it is possible in practice to assess the utility of a model (at least in the case that it performs poorly) even without priors for its unspecified parameters.

One way to approach such situations is to use some of the available data to obtain constraints on the model's parameters, and then to use this distribution can as the prior needed to produce a defined marginal likelihood for the remainder of the data. The implementation of this idea for separable data is described in Section 13.4.1.

10.4.1 Separable data

A two-step approach to model comparison is most obviously suited to separable data, $\boldsymbol{d} = (\boldsymbol{d}_1, \boldsymbol{d}_2)$, which can be split into a fitting/training sample, $\boldsymbol{d}_1$, and a testing sample, $\boldsymbol{d}_2$. The most obvious form of separable data is the case of repeated (and possibly i.i.d.) measurements of some quantity, but a more general definition is any case for which the likelihood can be factorized as

$$\mathbb{P}(\boldsymbol{d}|\boldsymbol{\theta}, M) = \mathbb{P}(\boldsymbol{d}_1, \boldsymbol{d}_2|\boldsymbol{\theta}, M) = \mathbb{P}(\boldsymbol{d}_1|\boldsymbol{\theta}, M)\,\mathbb{P}(\boldsymbol{d}_2|\boldsymbol{\theta}, M), \tag{123}$$

where $\boldsymbol{\theta}$ are the parameters of the model, M, under consideration. There is no specified scheme for separating the data as yet; the only requirement at this stage is that it can be split.

The first step to calculating the marginal likelihood under model M is to use the fitting sample to obtain a unit-normalised distribution for the model's parameters. This is given by

$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}_1, M) = \frac{\tilde{\mathbb{P}}(\boldsymbol{\theta}|M)\,\mathbb{P}(\boldsymbol{d}_1|\boldsymbol{\theta}, M)}{\int \mathrm{d}\boldsymbol{\theta}'\,\tilde{\mathbb{P}}(\boldsymbol{\theta}'|M)\,\mathbb{P}(\boldsymbol{d}_1|\boldsymbol{\theta}', M)} \tag{124}$$

where $\tilde{\mathbb{P}}(\boldsymbol{\theta}|M)$ is some uninformative, generally improper, prior on the parameters. The requirement that it is possible to use an improper prior at this stage means that sufficient data has to be included in $\boldsymbol{d}_1$ to ensure that $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}_1, M)$ is a valid density, as described further in Section 9.3.3.

Clearly $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}_1, M)$ depends to some degree on the form of $\tilde{\mathbb{P}}(\boldsymbol{\theta}|M)$, which is precisely the distribution for which no well-motivated form is available; however, the key idea of the technique is not that the results are rigorously independent of the assumed prior, but that there is a way of obtaining a reasonable answer despite the prior being unmotivated and/or unnormalised. In most cases a number of plausible priors could be adopted, and the usual exploration of prior sensitivity is recommended.

Having obtained $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{d}_1, M)$, the next step is to interpret this distribution as the required prior distribution in $\boldsymbol{\theta}$, which can then be used as the necessary ingredient for calculating the marginal likelihood. The marginal likelihood for the as-yet unused data is hence

$$\begin{aligned}
\mathbb{P}(\boldsymbol{d}_2|\boldsymbol{d}_1, M) &= \int \mathrm{d}\boldsymbol{\theta}'\,\mathbb{P}(\boldsymbol{\theta}'|\boldsymbol{d}_1, M)\,\mathbb{P}(\boldsymbol{d}_2|\boldsymbol{\theta}', M) \\
&= \int \mathrm{d}\boldsymbol{\theta}' \left[\frac{\mathbb{P}(\boldsymbol{\theta}'|M)\,\mathbb{P}(\boldsymbol{d}_1|\boldsymbol{\theta}', M)}{\int \mathrm{d}\boldsymbol{\theta}''\,\mathbb{P}(\boldsymbol{\theta}''|M)\,\mathbb{P}(\boldsymbol{d}_1|\boldsymbol{\theta}'', M)}\right] \mathbb{P}(\boldsymbol{d}_2|\boldsymbol{\theta}', M) \\
&= \frac{\int \mathrm{d}\boldsymbol{\theta}'\,\mathbb{P}(\boldsymbol{\theta}'|M)\,\mathbb{P}(\boldsymbol{d}_1|\boldsymbol{\theta}', M)\ \mathbb{P}(\boldsymbol{d}_2|\boldsymbol{\theta}', M)}{\int \mathrm{d}\boldsymbol{\theta}''\,\mathbb{P}(\boldsymbol{\theta}''|M)\,\mathbb{P}(\boldsymbol{d}_1|\boldsymbol{\theta}'', M)} \\
&= \frac{\int \mathrm{d}\boldsymbol{\theta}'\,\mathbb{P}(\boldsymbol{\theta}'|M)\,\mathbb{P}(\boldsymbol{d}_1, \boldsymbol{d}_2|\boldsymbol{\theta}', M)}{\int \mathrm{d}\boldsymbol{\theta}''\,\mathbb{P}(\boldsymbol{\theta}''|M)\,\mathbb{P}(\boldsymbol{d}_1|\boldsymbol{\theta}'', M)} \\
&= \frac{\mathbb{P}(\boldsymbol{d}_1, \boldsymbol{d}_2|M)}{\mathbb{P}(\boldsymbol{d}_1|M)}.
\end{aligned} \tag{125}$$

This final result is, in one sense, a trivial application of the chain rule, as $\mathbb{P}(\boldsymbol{d}|M) = \mathbb{P}(\boldsymbol{d}_1, \boldsymbol{d}_2|M) = \mathbb{P}(\boldsymbol{d}_1|M)\,\mathbb{P}(\boldsymbol{d}_2|\boldsymbol{d}_1, M)$; but the route to calculating it is, while mathematically more complicated, more compelling conceptually. The process described mathematically above, of using some data to refine a hypothesis about the world and then gradually adopting that hypothesis if more data confirms it, is what humans do all the time.

This formalism provides a way of calculating a marginal likelihood for a model with unspecified parameters for which no strongly-motivated prior distribution is available, and is coherent in the sense that no part of the data is used multiple times. But, unlike almost all other Bayesian methods, there is an element of choice in how to split the data into the fitting and testing components. Further, there are several different types of ambiguity: one is the scheme to be used (*e.g.*, what fraction of the data to use for the parameter fitting step); another is how to divide the data (randomly, by the order it was obtained, *etc.*). The first issue could decided upon by calculation, but the second is more fundamental: two people applying the same algorithm to the same data would end up with different marginal likelihood values simply because they used a different random number generator. This is a decidedly unsatisfactory situation. While a number of proposed modifications to this basic algorithm have been suggested, most result in considerable complication of the algorithm or incoherent use of the data.

10.5 Incomplete set of models

In the above model comparison formalism it is implicitly assumed that one of the models in $\{M_1, M_2, \ldots, M_N\}$ is "correct". This includes the case that the model is a perfect mathematical description of the data-generation mechanism (*e.g.*, if it is known that a function is a polynomial and the only question is over its order) or if the model satisfies the lessr requirement of being able to explain all aspects of the data-set being analysed. This situation is sometimes referred to as "M-closed".

If, however, none of the models can explain the observations (*i.e.*, because $\{M_1, M_2, \ldots, M_N\}$ does not include the mechanism used to generate the data) there is the risk of obtaining meaningless model comparison results. If this is true then the situation is said to be "M-open", and is a considerably more challenging situation.

The difficulty is that if at least one of the N models under consideration is such that $\mathbb{P}(\boldsymbol{d}|M) > 0$ then the result of the model comparison formalism will be a formally valid (categorical) posterior distribution over $M_1, M_2, \ldots, M_N$. This will be true even if none of the models can explain the data well, and this formalism does not, in itself, include any mechanism for identifying this problem.

There is a similar potential problem in the case of parameter estimation if the assumed model is incorrect, but in that case is potential recourse to the higher level of inference that is model comparison. But that is not the case here, as any higher level would effectively just be another model (which presumably should have been included in $\{M_1, M_2, \ldots, M_N\}$).

Instead, what is needed is model-checking or hypothesis testing, which is a related, but fundamentally distinct topic.

Learning objectives

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

- Apply Bayesian inference to discrete/separate models, in particular evaluating marginal likelihoods for models with unspecified parameters.
- Formulate and solve model comparison problems in the presence of unspecified parameters, either by adopting well-motivated prior distributions or exploring the plausible options.
- Use the Savage–Dickey density ratio (SDDR) to perform efficient comparison of nested models.
- Quickly evaluate model preferences in the the limiting cases of infinitely broad priors and/or completely irrelevant (extra) parameters.
- Exploit the structure of separable data-sets by using some of the data to obtain the parameter posteriors which can then be used as the prior distributions necessary for model comparison on the remainder of the data.