Bayesian Methods

Xander Byrne

Lent 2024

1 Introduction

When we do an experiment, we obtain some data $\mathbf{D}$. We typically do this experiment in order to constrain some parameters $\boldsymbol{\theta}$ of a model. By 'model', we mean that the probability of obtaining the data $\boldsymbol{\theta}$ for a given set of parameters is given by the *likelihood* $\mathcal{L} = P(\mathbf{D}|\boldsymbol{\theta})$. *Inference* is the process of estimating the true values of the parameters $\boldsymbol{\theta}$ based on the data $\mathbf{D}$. This can be done by inverting the probability distribution using Bayes' Theorem:

$$P(\boldsymbol{\theta}|\mathbf{D}) = \frac{P(\mathbf{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{D})} \equiv \frac{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z(\mathbf{D})}$$

where in the second equality we have done some relabelling:

- The *evidence* $Z(\mathbf{D})$ is essentially a normalisation constant on $P(\boldsymbol{\theta}|\mathbf{D})$, which naturally must integrate ($\mathrm{d}\boldsymbol{\theta}$) to 1. Note that it does not depend on $\boldsymbol{\theta}$.

- The *prior* $\pi(\boldsymbol{\theta})$ captures our ignorance of the parameters $\boldsymbol{\theta}$. If we know very little about the values of $\boldsymbol{\theta}$, the function $\pi(\boldsymbol{\theta})$ will be quite broad and $P(\boldsymbol{\theta}|\mathbf{D})$ will be very similar to $\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})$. If we already have a pretty good idea of the parameters then $\pi(\boldsymbol{\theta})$ will be very sharp; in the limit that we know that the true values of $\boldsymbol{\theta}$ are $\boldsymbol{\theta}^*$, we have $\pi(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = P(\boldsymbol{\theta}|\mathbf{D})$: whatever data we get won't change our mind.

- $P(\boldsymbol{\theta}|\mathbf{D})$ is called the *posterior*.

1.1 Choosing Priors

Controversially, in Bayesian probability we are more or less free to choose the prior function. Typically people use a uniform prior $\pi(\boldsymbol{\theta}) = \mathbb{1}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ to profess no knowledge of $\boldsymbol{\theta}$ at all except that it is within some region $\boldsymbol{\Theta}$ of parameter space. Alternatively, "scale parameters" which are > 0 choose priors such that $\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \pi(\alpha\boldsymbol{\theta})\mathrm{d}(\alpha\boldsymbol{\theta}) \Rightarrow \pi \propto 1/\boldsymbol{\theta}$.

It can be useful for algebraic reasons to choose a prior such that, for a given likelihood function, the posterior has the same functional form as the prior; such priors are called *conjugate priors* of that particular likelihood.

1.2 Information

If the data are very high quality, then the likelihood is a strongly peaked function of $\boldsymbol{\theta}$: the range of $\boldsymbol{\theta}$ which could produce the given $\mathbf{D}$ is very small. This 'quality' of a likelihood is quantified by the *information* $I(\boldsymbol{\theta})$. The information is given in terms of the *score* $S(\boldsymbol{\theta}) = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln \mathcal{L}$. Consider the expectation of S, taken over all possible datasets (we denote the space of all possible datasets as $\mathcal{D}$); the probability of a given dataset $\mathbf{D} \in \mathcal{D}$ is given by $\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}^*)$:

$$\mathbb{E}_{\mathcal{D}}[S] = \int_{\mathcal{D}} S(\boldsymbol{\theta})\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}^*)\mathrm{d}\mathbf{D} = \int_{\mathcal{D}} \boldsymbol{\nabla}_{\boldsymbol{\theta}}[\ln \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})]\ \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}^*)\,\mathrm{d}x = \int_{\mathcal{D}} \boldsymbol{\nabla}_{\boldsymbol{\theta}}[\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})]\frac{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}^*)}{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})}\mathrm{d}\mathbf{D}$$

Now the above, the expectation value of the score, is a function of $\boldsymbol{\theta}$: we can take loads of datasets and measure the score for each (using a particular choice of the parameters $\boldsymbol{\theta}$), and we will find that the expected value of the score is as above. If we evaluate the above quantity at the true value $\boldsymbol{\theta}^*$, we get:

$$\mathbb{E}_{\mathcal{D}}[S]\bigg|_{\boldsymbol{\theta}^*} = \int_{\mathcal{D}} \boldsymbol{\nabla}_{\boldsymbol{\theta}}[\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})]\bigg|_{\boldsymbol{\theta}^*} \mathrm{d}\mathbf{D} = \boldsymbol{\nabla}_{\boldsymbol{\theta}}\Bigg[\underbrace{\int_{\mathcal{D}} \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\,\mathrm{d}X}_{1}\Bigg]\Bigg|_{\boldsymbol{\theta}^*} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} 1\bigg|_{\boldsymbol{\theta}^*} = 0$$

The information, rather than the expectation of the score, is the *variance* of the score: $I(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}[S(\boldsymbol{\theta})^2] - \mathbb{E}_{\mathcal{D}}[S(\boldsymbol{\theta})]^2$. If we evaluate the information at the true value $\boldsymbol{\theta}^*$, then the second term vanishes as we have shown, and we are left with:

$$I(\boldsymbol{\theta}^*) = \mathbb{E}_{\mathcal{D}}\big[S(\boldsymbol{\theta}^*)^2\big] = \cdots = -\int_{\mathcal{D}} \frac{\partial^2 \ln(\mathbf{D}|\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^2}\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}^*)\mathrm{d}\mathbf{D}$$

which is related to the second derivative of $\mathcal{L}$ with respect to $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}^*$. This is a good quantifier of how informative $\mathcal{L}$ is: if the likelihood function is very strongly peaked around the true value, then the second derivative will be very negative and I will be high.

If we zoom in on the function $\ln \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})$ near $\boldsymbol{\theta}^*$, we get the Taylor expansion:

$$\ln \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}) \approx \ln \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \cdot S(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^2\ \frac{\partial^2 \ln \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\bigg|_{\boldsymbol{\theta}^*}$$

Averaging across an ensemble of datasets, and using the results above, we obtain:

$$\mathbb{E}_{\mathcal{D}}[\ln \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})] \approx \text{const.} - \frac{I(\boldsymbol{\theta}^*)}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^2$$

1.3 Summary Statistics

The **marginal distribution** is the posterior distribution for a subset of the parameters $\boldsymbol{\theta}$. This might be done because $\boldsymbol{\theta}$ may contain a lot of nuisance parameters. Given a posterior $P(\boldsymbol{\theta}|\mathbf{D})$, the probability distribution for some subset $\boldsymbol{\phi} \in \boldsymbol{\theta}$ is given by:

$$P(\boldsymbol{\phi}|\mathbf{D}) = \int_{\theta_i \notin \boldsymbol{\phi}} P(\boldsymbol{\theta}|\mathbf{D})\ \mathrm{d}\theta_i$$

Point estimates of $\boldsymbol{\theta}$ might be the mean $\mathbb{E}_{\mathcal{D}}[\boldsymbol{\theta}] = \int \mathrm{d}\boldsymbol{\theta} P(\boldsymbol{\theta}|\mathbf{D})$, the mode (AKA the maximum a posteriori (MAP) value), or perhaps the median of each of the 1D marginal distributions

Credible intervals, e.g. a 90% credible interval, are a Bayesian version of error bars. Lots of them are flawed, but we might choose the narrowest possible interval containing the right % of the posterior distribution; or the highest density region, where the boundary of the interval is an isoprobability surface; or in 1D we might choose an equal-tailed distribution.

2 Markov Chain Monte Carlo (MCMC) Methods

Often, the posterior $P(\boldsymbol{\theta}|\mathbf{D})$ is difficult to generate analytically, or even impossible. Maybe the likelihood $\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})$ is calculated by some simulation, with $\boldsymbol{\theta}$ as input parameters, rather than simply a function. However, it is usually possible to *sample* from the posterior distribution, generating a sequence of $\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} P(\boldsymbol{\theta}|\mathbf{D})$. In this way, we can approximate the posterior distribution by drawing a large sample (of size N).

Simple methods – such as the accept-reject method, the inverse CDF method, or directly fitting a histogram – all scale horribly to multiple dimensions. However, the family of Markov chain Monte Carlo methods are often not too bad. They also apply to sampling a general probability distribution $P^*(\mathbf{x})$, not just a posterior $P(\boldsymbol{\theta}|\mathbf{D})$. For conciseness and consistency, we will use the variable $\mathbf{x}$ to refer to the variable we are trying to sample, but within this course the $\mathbf{x}$ are thought of as the parameters of a model, distributed according to the posterior.

2.1 Basic Concepts

2.1.1 Markov Chains

A Markov chain is a sequence of points $\{\mathbf{x}_i\}$ in the sample space $\mathcal{X}$ (in Bayesian terms, our parameter space $\boldsymbol{\Theta}$), whereby the probability distribution of the next sample $\mathbf{x}_{i+1}$ depends only on the current value $\mathbf{x}_i$:

$$P(\mathbf{x}_{i+1}|\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_i) = P(\mathbf{x}_{i+1}|\mathbf{x}_i)$$

This is effectively a “transition probability”.

Many Markov chains are *time-homogeneous*[1]: the transition probability doesn’t change with “time” (that is, i):

$$P(\mathbf{x}_{i+1}|\mathbf{x}_i) = \rho(\mathbf{x}', \mathbf{x})$$

for some ρ, which may not be symmetric. For example, we have $P(\mathbf{x}_1|\mathbf{x}_0) = \rho(\mathbf{x}_1, \mathbf{x}_0)$. As for $P(\mathbf{x}_2|\mathbf{x}_0)$, this will be the sum over the possibilities of arriving at $\mathbf{x}_2$ via all the possible intermediate value of $\mathbf{x}_1$:

$$P(\mathbf{x}_2|\mathbf{x}_0) = \int_{\mathcal{X}} \mathrm{d}\mathbf{x}_1 \; P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1|\mathbf{x}_0) = \int_{\mathcal{X}} \mathrm{d}\mathbf{x}_1 \; \rho(\mathbf{x}_2, \mathbf{x}_1)\rho(\mathbf{x}_1, \mathbf{x}_0)$$

and so on.

2.1.2 Stationary Distribution and the Detailed Balance Condition

Different MCMC methods consist essentially of different choices of $\rho(\mathbf{x}', \mathbf{x})$. To be a suitable transition function, we require that the *limit distribution* $\lambda(\mathbf{x}_i) \equiv \lim_{i\to\infty} P(\mathbf{x}_i|\mathbf{x}_0)$ tends towards the target distribution $P^*(\mathbf{x})$.

If the samples in our MCMC chain reach the target distribution, we’d like them to stay there. In other words, if $\mathbf{x}_i \sim P^*(\mathbf{x})$, we want $\mathbf{x}_{i+1} \sim P^*(\mathbf{x})$ also. Now $\mathbf{x}_{i+1}|\mathbf{x}_i \sim \rho(\mathbf{x}_{i+1}, \mathbf{x}_i)$, so this condition requires:

$$P(\mathbf{x}_{i+1}) \equiv \int_{\mathcal{X}} \mathrm{d}\mathbf{x}_i \; P(\mathbf{x}_{i+1}|\mathbf{x}_i)P(\mathbf{x}_i) = \int_{\mathcal{X}} \mathrm{d}\mathbf{x}_i \; \rho(\mathbf{x}_{i+1}, \mathbf{x}_i)P^*(\mathbf{x}_i)$$

[1] As far as I can tell, this isn’t a requirement for MCMC, but is simply a convenient choice.

to be equal to the target distribution $P^*(\mathbf{x}_{i+1})$. This requires ρ to be such that:

$$P^*(\mathbf{x}') = \int \mathrm{d}\mathbf{x}\ P^*(\mathbf{x})\rho(\mathbf{x}', \mathbf{x}) \tag{$\mathcal{S}$}$$

Technically, this is equivalent to imposing ρ such that the chain reaches a *stationary distribution* that is equal to the target distribution[2] In the limit of large i, this will also be the limit distribution.

The above condition can be quite difficult to show for many MCMC methods. Instead, we can derive a stricter condition, which is sufficient, but not necessary, for having a stationary distribution. This is the *detailed balance condition*:

$$P^*(\mathbf{x})\rho(\mathbf{x}', \mathbf{x}) = P^*(\mathbf{x}')\rho(\mathbf{x}, \mathbf{x}') \tag{$\mathcal{DBC}$}$$

That this is sufficient for the chain to reach a stationary distribution of the target distribution can easily be shown by integrating with respect to $\mathbf{x}'$:

$$\int \mathrm{d}\mathbf{x}'\ P^*(\mathbf{x})\rho(\mathbf{x}', \mathbf{x}) = \int \mathrm{d}\mathbf{x}'\ P^*(\mathbf{x}')\rho(\mathbf{x}, \mathbf{x}') \qquad \Rightarrow \qquad P^*(\mathbf{x}) = \int \mathrm{d}\mathbf{x}'\ P^*(\mathbf{x}')\rho(\mathbf{x}, \mathbf{x}')$$

from which $\mathcal{S}$ can be derived simply by exchanging the variables in the various functions. Note that the converse doesn't hold.

The significance of $\mathcal{DBC}$ is that if the chain reaches the point $\mathbf{x}$, it is equally likely to move to the point $\mathbf{x}'$ as if it reached the point $\mathbf{x}'$ and move to the point $\mathbf{x}$. This can be easily shown by calculating the probabilities $P(\mathbf{x} \in \mathcal{A}\ \cap\ \mathbf{x}' \in \mathcal{B})$ and vice versa, where $\mathcal{A}, \mathcal{B} \subset \mathcal{X}$; using detailed balance one can show the two to be equal.

2.2 Metropolis-Hastings

The Metropolis-Hastings algorithm makes use of a *proposal distribution* $Q(\mathbf{y}|\mathbf{x}_i)$ to suggest candidates for $\mathbf{x}_{i+1}$; for example:

$$Q(\mathbf{y}|\mathbf{x}_i) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}\|\mathbf{y} - \mathbf{x}_i\|^2\right)$$

At a given point in the chain $\mathbf{x}_i$, a point $\mathbf{y} \sim Q(\mathbf{y}|\mathbf{x}_i)$ is drawn (Q should be chosen as a distribution that is easy to sample!). Whether $\mathbf{y}$ is then accepted as the sample in the Markov chain $\mathbf{x}_{i+1}$ depends on the quantity:

$$a(\mathbf{y}, \mathbf{x}_i) = \frac{P^*(\mathbf{y})Q(\mathbf{x}_i|\mathbf{y})}{P^*(\mathbf{x}_i)Q(\mathbf{y}|\mathbf{x}_i)}$$

Note that if Q is symmetric in its two arguments (as in the normal example given), then a reduces to $P^*(\mathbf{y})/P^*(\mathbf{x}_i)$; this special case is described as the *Metropolis algorithm*. Also, note that because we take the ratio of two target distributions, it doesn't need to be normalised. If a is larger than a number $u \sim \mathcal{U}(0, 1)$, then $\mathbf{x}_{i+1} = \mathbf{y}$. Otherwise, the sample is rejected; we leave $\mathbf{x}_{i+1} = \mathbf{x}_i$.

[2] By swapping out $P^*(\mathbf{x})$ for another distribution, you can make ρ produce a chain that has an arbitrary stationary distribution.

We now show that this algorithm corresponds to a $\rho(\mathbf{x}_{i+1}, \mathbf{x}_i)$ that has the right stationary distribution, by showing that it satisfies $\mathcal{DBC}$. We first derive $\rho(\mathbf{x}', \mathbf{x})$:

$$\begin{aligned}\rho(\mathbf{x}', \mathbf{x}) \equiv P(\mathbf{x}'|\mathbf{x}) &= \int \mathrm{d}\mathbf{y}\ P(\mathbf{x}'|\mathbf{x}, \mathbf{y})P(\mathbf{y}|\mathbf{x}) = \int \mathrm{d}\mathbf{y}\ P(\mathbf{x}'|\mathbf{x}, \mathbf{y})Q(\mathbf{y}|\mathbf{x}) \\ &= \int \mathrm{d}\mathbf{y}\ \Big[P(\mathbf{x}'|\mathbf{x}, \mathbf{y}, A(\mathbf{y}, \mathbf{x}) > u) \cdot P(A(\mathbf{y}, \mathbf{x}) > u|\mathbf{x}) \\ &\qquad\qquad + P(\mathbf{x}'|\mathbf{x}, \mathbf{y}, A(\mathbf{y}, \mathbf{x}) < u) \cdot P(A(\mathbf{y}, \mathbf{x}) < u|\mathbf{x})\Big] \cdot Q(\mathbf{y}|\mathbf{x})\end{aligned}$$

where $A(\mathbf{y}, \mathbf{x}) \equiv \min(1, a(\mathbf{y}, \mathbf{x}))$. Now all four of the probabilities shown are known:

$$\begin{aligned} P(\mathbf{x}'|\mathbf{x}, \mathbf{y}, A > u) &= \delta(\mathbf{x}' - \mathbf{y}) \qquad & P(A > u|\mathbf{x}) &= A \\ P(\mathbf{x}'|\mathbf{x}, \mathbf{y}, A < u) &= \delta(\mathbf{x}' - \mathbf{x}) \qquad & P(A < u|\mathbf{x}) &= 1 - A \end{aligned}$$

giving us:

$$\begin{aligned}\rho(\mathbf{x}', \mathbf{x}) &= \int \mathrm{d}\mathbf{y}\ \Big[\delta(\mathbf{x}' - \mathbf{y})A(\mathbf{y}, \mathbf{x}) + \delta(\mathbf{x}' - \mathbf{x})(1 - A(\mathbf{y}, \mathbf{x}))\Big]Q(\mathbf{y}|\mathbf{x}) \\ &= A(\mathbf{x}', \mathbf{x})Q(\mathbf{x}'|\mathbf{x}) + \delta(\mathbf{x}' - \mathbf{x}) \int \mathrm{d}\mathbf{y}\ (1 - A(\mathbf{y}, \mathbf{x}))Q(\mathbf{y}|\mathbf{x})\end{aligned}$$

We can now show that this satisfies $\mathcal{DBC}$. Note from the definition of a, that $a(\mathbf{x}, \mathbf{x}') = 1/a(\mathbf{x}', \mathbf{x})$. There are therefore two possibilities: either $A(\mathbf{x}, \mathbf{x}') < 1$ and $A(\mathbf{x}', \mathbf{x}) = 1$, or vice versa. Consider the first case: the LHS and RHS of the DBC are then given by:

$$\begin{aligned}\text{LHS} &= P^*(\mathbf{x})\rho(\mathbf{x}', \mathbf{x}) = P^*(\mathbf{x})\left[Q(\mathbf{x}'|\mathbf{x}) + \delta(\mathbf{x}' - \mathbf{x}) \int \mathrm{d}\mathbf{y}\ (1 - A(\mathbf{y}|\mathbf{x}))Q(\mathbf{y}|\mathbf{x})\right] \\ \text{RHS} &= P^*(\mathbf{x}')\rho(\mathbf{x}, \mathbf{x}') = P^*(\mathbf{x}')\left[\frac{P^*(\mathbf{x})Q(\mathbf{x}'|\mathbf{x})}{P^*(\mathbf{x}')Q(\mathbf{x}|\mathbf{x}')}Q(\mathbf{x}|\mathbf{x}') + \delta(\mathbf{x} - \mathbf{x}') \int \mathrm{d}\mathbf{y}\ (1 - A(\mathbf{y}|\mathbf{x}))Q(\mathbf{y}|\mathbf{x})\right]\end{aligned}$$

It can be seen that the delta terms are equal to each other, most easily when one considers that they will each be 0 unless $\mathbf{x}' = \mathbf{x}$. Similarly, when one simplifies the fraction in the first term of the RHS, it can be seen that this is equal to the LHS. Thus detailed balance is satisfied; thus this algorithm produces a transition probability $\rho(\mathbf{x}', \mathbf{x})$ which leads to the Markov chain's stationary distribution being the target distribution $P^*(\mathbf{x})$.

2.2.1 Choice of Proposal Distribution $Q(\mathbf{y}|\mathbf{x})$

The efficiency of the algorithm depends on the choice of proposal distribution $Q(\mathbf{y}|\mathbf{x})$; generally there is no good way to choose this first time. If the proposal distribution is very sharp compared to the target distribution, then $\mathbf{x}_{i+1}$ will be very close to $\mathbf{x}_i$, which will mean that the two are not really independent, and also that if the starting position is unfortunately far away from regions with a high value of the target distribution then it will take a while for the samples to reach this distribution. To counteract these, we *thin* the chain and remove a *burn-in*. By thinning the chain, we mean we only take every T samples as being iid, and thus representative samples of the target distribution. A good value of T can be estimated by calculating the *autocorrelation length* of the chain (the distance over which the samples are intercorrelated, rather than being dictated by the target distribution), and setting T to a value of a few times the autocorrelation length. To decide how large the burn-in should be, one can employ *trace plots*, plotting the value of each parameter against distance along the chain. It is typically clear roughly after how many samples the chain has converged.

Choosing a large proposal distribution is also problematic, as this will often lead to $\mathbf{y}$ being a long way away from $\mathbf{x}_i$, and hence $P^*(\mathbf{y})$ being very small, rejection being more likely, and it being less likely that the chain will move and explore the parameter space properly.

2.3 Gibbs Sampling

Gibbs sampling samples from the 1D conditional distributions for each parameter, such as $P(x^0|x^1, x^2, \dots)$; there are as many such distributions as there are parameters. We will denote these as,

$$P(x^k|\mathbf{x}^{-k}) = P(x^k|x^0, \dots, x^{k-1}, x^{k+1}, \dots) \equiv \frac{P^*(\mathbf{x})}{\int P^*(\mathbf{x})\,\mathrm{d}x^k}$$

These distributions are effectively slices through the full distribution. The Gibbs sampling procedure is as follows. At each step, the following sample is produced by randomly selecting a particular component k, and changing the kth component of $\mathbf{x}$ to a value sampled from the conditional distribution. The components of $\mathbf{x}_{i+1}$ are thus given by:

$$x'^j = \begin{cases} y \sim P(x^k|\mathbf{x}^{-k}) & j = k \\ x^j & j \neq k \end{cases}$$

Which component k to change at each step is typically chosen such that each component is equally likely to be selected (with probability $1/\dim \mathbf{x}$), but a variant allows weightings w_k, where $\sum_j w_j = 1$. Note again that the normalised distribution is not required, as the conditional distribution involves a quotient. However, unlike Metropolis-Hastings, Gibbs moves at every step: $\mathbf{x}_{i+1} \neq \mathbf{x}_i$.

We now show that the Gibbs sampling method satisfies $\mathcal{DBC}$. We can find $\rho(\mathbf{x}', \mathbf{x})$ relatively easily:

$$\rho(\mathbf{x}', \mathbf{x}) \equiv P(\mathbf{x}'|\mathbf{x}) = w_k \delta(\mathbf{x}'^{-k} - \mathbf{x}^{-k}) P(x'^k|\mathbf{x}^{-k})$$

Consider now the LHS of $\mathcal{DBC}$:

$$\text{LHS} = P^*(\mathbf{x})\rho(\mathbf{x}', \mathbf{x}) = w_k P^*(\mathbf{x})\delta(\mathbf{x}'^{-k} - \mathbf{x}^{-k}) P(x'^k|\mathbf{x}^{-k})$$

and note that the RHS is the same under $\mathbf{x}' \leftrightarrow \mathbf{x}$. Now both these will be 0, except where $\mathbf{x}' = \mathbf{x}$, at which point they will clearly both be equal. Thus $\mathcal{DBC}$ is satisfied and the Gibbs rule leads to a satisfactory stationary distribution.

A variant on the Gibbs sampling is the *Gibbs sweep*, which is aimed at reducing the significant redundancy present in consecutive samples. In this algorithm, rather than picking a component to vary each time, we cycle through the components:

$$\begin{aligned} x'^0 &\leftarrow y \sim P(x^0|x^1, x^2, \dots) \\ x'^1 &\leftarrow y \sim P(x^1|x'^0, x^2, \dots) \end{aligned}$$

Note that after each component has been sampled, for subsequent samples we need to condition the distribution on the *new* components, otherwise it turns out that this will not converge to the target distribution. In fact, it turns out that the Gibbs sweep doesn't actually satisfy $\mathcal{DBC}$, though it does actually converge to the target distribution (recall that $\mathcal{DBC}$ is sufficient, but not necessary). This is shown below for the two-dimensional case, where we go from $\mathbf{x} = (x^0, x^1)$ to $\mathbf{x}' = (x'^0, x^1)$ via $\mathbf{y} = (x'^0, x^1)$. We first calculate the transition probability:

$$\rho(\mathbf{x}', \mathbf{x}) \equiv P(\mathbf{y}|\mathbf{x})P(\mathbf{x}'|\mathbf{y}) = \frac{P^*(x'^0, x^1)}{\int P^*(x, x^1)\,\mathrm{d}x} \frac{P^*(x'^0, x'^1)}{\int P^*(x'^0, x)\,\mathrm{d}x}$$

Substituting into the RHS of $\mathcal{S}$,

$$\begin{aligned}
\text{RHS} &= \int \mathrm{d}\mathbf{x}\ P^*(\mathbf{x}) \frac{P^*(x'^0, x^1)}{\int P^*(x, x^1)\,\mathrm{d}x} \frac{P^*(x'^0, x'^1)}{\int P^*(x'^0, x)\,\mathrm{d}x} \\
&= \frac{P^*(x'^0, x'^1)}{\int P^*(x'^0, x)\,\mathrm{d}x} \int \mathrm{d}x^0 \int \mathrm{d}x^1\ P^*(x^0, x^1) \frac{P^*(x'^0, x^1)}{\int P^*(x, x^1)\,\mathrm{d}x} \\
&= \frac{P^*(x'^0, x'^1)}{\int P^*(x'^0, x)\,\mathrm{d}x} \int \mathrm{d}x^1\ \frac{P^*(x'^0, x^1)}{\int P^*(x, x^1)\,\mathrm{d}x} \int \mathrm{d}x^0 P^*(x^0, x^1) \\
&= P^*(x'^0, x'^1) = P^*(\mathbf{x}')
\end{aligned}$$

where two pairs of integrals have cancelled from the penultimate line, giving the LHS of $\mathcal{S}$ and showing that the Gibbs sweep has a stationary distribution equal to the target.

Another variant of Gibbs is the *blocked Gibbs* sampler, which groups together "blocks" of components, and samples from the subspaces of these blocks one at a time, conditional on the components of the other blocks.

2.4 Hamiltonian Monte Carlo

Both Metropolis-Hastings and Gibbs suffer from "random walk" behaviour, where successive samples are quite close together, and it takes a long time to explore the sample space. Hamiltonian Monte Carlo (HMC) allows for larger steps, and makes sure those steps are not in a terrible direction. It is based on principles from Hamiltonian dynamics, and there is terminology associated with this. For example, we usually work in terms of the "potential energy" $V(\mathbf{x}) = -\log P^*(\mathbf{x}) + \text{const.}$. Unlike the previous methods, we also need the gradient $\boldsymbol{\nabla} V$, though this is sometimes accessible by autodifferentiation.

HMC proceeds by first doubling the dimensionality of the space, by tacking on the end of the samples $\mathbf{x}$ a vector $\mathbf{p}$ of the same dimensionality. For $\mathbf{p}$, we invent a new distribution $Q(\mathbf{p})$, which is essentially arbitrary other than $Q(-\mathbf{p}) = Q(\mathbf{p})$; it should also be easy to sample from: we often choose $Q \propto \exp\left(-\frac{1}{2}\mathbf{p}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{p}\right)$, and hence $-\log Q = \frac{1}{2}\mathbf{p}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{p}$. We then define $K(\mathbf{p}) = -\log Q(\mathbf{p}) + \text{const.}$ as the "kinetic energy".

As might be anticipated, we then define the Hamiltonian $\mathcal{H} = K + V = -\log R(\mathbf{x}, \mathbf{p})$, for some quantity $R = P^*(\mathbf{x})Q(\mathbf{p})$ which we can show is the joint distribution of $\mathbf{x}$ and $\mathbf{p}$:

$$\begin{aligned}
\int R(\mathbf{x}, \mathbf{p})\ \mathrm{d}\mathbf{p} &= e^{-V(\mathbf{x})} \int e^{-K(\mathbf{p})}\ \mathrm{d}\mathbf{p} \\
&= P^*(\mathbf{x}) \int Q(\mathbf{p})\mathrm{d}\mathbf{p} \\
&= P^*(\mathbf{x})
\end{aligned}
\qquad
\begin{aligned}
\int R(\mathbf{x}, \mathbf{p})\ \mathrm{d}\mathbf{x} &= e^{-K(\mathbf{p})} \int e^{-V(\mathbf{x})}\ \mathrm{d}\mathbf{x} \\
&= Q(\mathbf{p}) \int P^*(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= Q(\mathbf{p})
\end{aligned}$$

Sampling then proceeds in this double-dimensional space using the equations of Hamiltonian dynamics:

$$\frac{\mathrm{d}x^k}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial p^k}; \qquad \frac{\mathrm{d}p^k}{\mathrm{d}t} = -\frac{\partial \mathcal{H}}{\partial x^k}$$

One chooses a timestep s, and integrates the above equations, using the leapfrog algorithm[3]:

$$\begin{aligned}
\mathbf{p} &\leftarrow \mathbf{p} - \frac{1}{2} s \boldsymbol{\nabla}_{\mathbf{x}} E \\
\mathbf{x} &\leftarrow \mathbf{x} + s \boldsymbol{\Sigma} \mathbf{p} \\
\mathbf{p} &\leftarrow \mathbf{p} - \frac{1}{2} s \boldsymbol{\nabla}_{\mathbf{x}} E
\end{aligned}$$

For each sample, this is repeated L times, where L is some hyperparameter. This will of course not integrate the equations exactly, but there will be an error of order s^2 difference between $\mathcal{H}_{\text{fin}}$ and $\mathcal{H}_{\text{init}}$.

As with Metropolis-Hastings, we then either accept or reject the new values of $\mathbf{x}$ and $\mathbf{p}$, based on a quantity a compared to a random variable $u \sim \mathcal{U}(0, 1)$. This time, a is defined as

$$a = \exp(\mathcal{H}_{\text{init}} - \mathcal{H}_{\text{fin}})$$

If $a > u$, then the new $\mathbf{x}$ (and $\mathbf{p}$) are accepted as $\mathbf{x}_{i+1}$; if not then $\mathbf{x}_{i+1} = \mathbf{x}_i$. Most samples should be selected if s and L are small, as then $\mathcal{H}_{\text{init}} \approx \mathcal{H}_{\text{fin}}$.

A common variation on Hamiltonian Monte Carlo is the *no U-turn sampler* (NUTS). This essentially allows for real-time choice of L. NUTS evolves the Hamiltonian forwards and backwards in time until a U-turn occurs, when we should stop because that would be wasted time.

2.5 Slice Sampling

As a final small one, slice sampling is similar to accept-reject sampling. It's very simple. To go from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$, first sample $\mathbf{y} \sim \mathcal{U}(0, P^*(\mathbf{x}_i))$ uniformly from 0 to $P^*(\mathbf{x}_i)$. Then, find all the regions of sample space where $P^*(\mathbf{x}) > \mathbf{y}$, and sample uniformly from those: $\mathbf{x}_{i+1} \sim \mathcal{U}(\{\mathbf{x} : P^*(\mathbf{x}) > \mathbf{y}\})$. That's it! $P^*(\mathbf{x})$ doesn't even need to be normalised.

The hard part is generally finding the region(s) $\{\mathbf{x} : P^*(\mathbf{x}) > \mathbf{y}\}$. This can be done using a root-finding algorithm, or perhaps a stepping-out procedure. Even these are horrible to implement in > 1 dimensions, so slice sampling is only usually used in 1D, or *within* Gibbs sampling to sample from the 1D conditionals.

[3] Other algorithms are available, but it does need to have particular ("symplectic") properties such as time-reversibility and volume preserving, which is not true of e.g. Euler or Runge-Kutta integration.

3 Bayesian Model Comparison

Suppose there are two models, A and B, which each attempt to describe the data $\mathbf{D}$, with parameters $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$. Now these models may take entirely different forms, so maybe $\dim \boldsymbol{\theta}_A \neq \dim \boldsymbol{\theta}_B$. One might think that two models can be compared using the maximum likelihood ratio:

$$\mathrm{MLR} = \frac{\max_{\boldsymbol{\theta}_A} \mathcal{L}(\mathbf{D}|A, \boldsymbol{\theta}_A)}{\max_{\boldsymbol{\theta}_B} \mathcal{L}(\mathbf{D}|B, \boldsymbol{\theta}_B)}$$

but because the different models may have different numbers of parameters, the one with more parameters will *naturally* be a better fit, regardless of how "good" a model it is.

The Bayesian way to compare models is to calculate the *posterior odds ratio*:

$$\begin{aligned}\mathcal{O}_{AB} &= \frac{P(A|\mathbf{D})}{P(B|\mathbf{D})} = \frac{P(\mathbf{D}|A)}{P(\mathbf{D}|B)}\frac{\pi(A)}{\pi(B)} = \frac{\int P(\mathbf{D}, \boldsymbol{\theta}_A|A)\ \mathrm{d}\boldsymbol{\theta}_A}{\int P(\mathbf{D}, \boldsymbol{\theta}_B|B)\ \mathrm{d}\boldsymbol{\theta}_B} \cdot \frac{\pi(A)}{\pi(B)} \\ &= \frac{\int P(\mathbf{D}|A, \boldsymbol{\theta}_A)\ P(\boldsymbol{\theta}_A|A)\ \mathrm{d}\boldsymbol{\theta}_A}{\int P(\mathbf{D}|B, \boldsymbol{\theta}_B)\ P(\boldsymbol{\theta}_B|B)\ \mathrm{d}\boldsymbol{\theta}_B} \cdot \frac{\pi(A)}{\pi(B)} \\ &= \frac{\int \mathcal{L}(\mathbf{D}|A, \boldsymbol{\theta}_A)\ \pi(\boldsymbol{\theta}_A|A)\ \mathrm{d}\boldsymbol{\theta}_A}{\int \mathcal{L}(\mathbf{D}|B, \boldsymbol{\theta}_B)\ \pi(\boldsymbol{\theta}_B|B)\ \mathrm{d}\boldsymbol{\theta}_B} \cdot \frac{\pi(A)}{\pi(B)} = \frac{Z_A(\mathbf{D})}{Z_B(\mathbf{D})} \cdot \frac{\pi(A)}{\pi(B)}\end{aligned}$$

which is thus the product of the *evidence ratio* (AKA *Bayes factor*) and the *prior odds ratio.*

This is nice, but in a Bayesian framework the prior odds ratio is not defined, but simply whatever you think it is. If one has no reason to prefer one over another, it is set to 1. We start with a quantitative prior belief of which model is better, and use the evidences for the respective models to update that belief. Bayes!!

Furthermore, the priors $\pi(\boldsymbol{\theta}_M|M)$ are also free choices. However, note that bad models with many parameters will have smaller evidences[4], giving rise to a built-in "Occam penalty".

3.1 Computing the Evidence

The evidence is the crucial part of model comparison, so we outline some ways to calculate it here. Our target is the integral $\int \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}$.

3.1.1 Analytic Evaluation

The evidence integral is rarely analytic, but sometimes is if one uses a conjugate prior.

3.1.2 Laplace's Approximation

Laplace's approximation approximates the integrand (the unnormalised posterior) as a Gaussian. This approximation would eventually become very accurate in the limit of infinite data.

The Gaussian is approximated by finding the maximum-posterior value of the parameters, $\hat{\boldsymbol{\theta}}$ (perhaps numerically), and Taylor-expanding $\ln P(\boldsymbol{\theta}|\mathbf{D})$ about this point:

$$\ln P(\boldsymbol{\theta}|\mathbf{D}) \approx \ln P(\hat{\boldsymbol{\theta}}|\mathbf{D}) - \frac{1}{2}\Big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\Big)^{\mathsf{T}} \mathbf{C} \Big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\Big), \qquad C_{ij} = -\left.\frac{\partial^2 \ln P(\boldsymbol{\theta}|\mathbf{D})}{\partial\theta_i \partial\theta_j}\right|_{\hat{\boldsymbol{\theta}}}$$

[4] Although the maximum-likelihood will be better in a higher-dimensional parameter space, the evidence involves integrating over the *entire* parameter space, which for an overfit model will contain vast regions where the likelihood is tiny. A fewer-parameter model that fits the data just ok will peak lower, but broader, giving a larger evidence integral.

$$\Rightarrow P(\boldsymbol{\theta}|\mathbf{D}) \approx P(\hat{\boldsymbol{\theta}}|\mathbf{D}) \exp\left[-\frac{1}{2}\Big(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\Big)^{\mathsf{T}}\mathbf{C}\Big(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\Big)\right] \qquad \Rightarrow Z \approx P(\hat{\boldsymbol{\theta}}|\mathbf{D})\sqrt{\frac{(2\pi)^{\dim\boldsymbol{\theta}}}{\|\mathbf{C}\|}}$$

where we use the standard integral of a multidimensional Gaussian. The components of the (negative) Hessian could be calculated numerically or automatically.

The Gaussian approximation becomes bad if there are multiple peaks, but if they are well-separated we can simply repeat the above integral at each peak and add them together.

Under a change of parameters $\boldsymbol{\theta} \to \boldsymbol{\phi}(\boldsymbol{\theta})$, the evidence should be invariant. However, the Laplace approximation is not. This is slightly awkward, but it can be used to our advantage: by changing parameters, we might find a set of parameters in which the posterior is more Gaussian, so in terms of these coordinates the approximation will be more accurate.

3.1.3 Thermodynamic Integration

The first step of thermodynamic integration is to *anneal* the likelihood function by taking it to the power $\beta \in [0, 1]$. We then have $\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \beta) \equiv \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})^{\beta}$. This annealing has the effect of smoothening out the likelihood function as $\beta \to 0$[5]. The analogy here is with thermodynamics, where $\beta = 1/kT$. Thus high temperature corresponds to low β, where the likelihood melts out.

Bayes' theorem including this new parameter β becomes:

$$P(\boldsymbol{\theta}|\mathbf{D}, \beta) = \frac{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})^{\beta}\pi(\boldsymbol{\theta})}{Z(\beta)}; \qquad Z(\beta) \equiv \int \mathrm{d}\boldsymbol{\theta}\ \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})^{\beta}\pi(\boldsymbol{\theta})$$

Note that $\beta = 1$ recovers the original case, whereas $Z(0) = \int \mathrm{d}\boldsymbol{\theta}\ \pi(\boldsymbol{\theta}) = 1$.

Consider now the derivative $\mathrm{d}\ln Z(\beta)/\mathrm{d}\beta$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\beta}\ln Z(\beta) &= \frac{1}{Z(\beta)}\frac{\mathrm{d}}{\mathrm{d}\beta}\int \mathrm{d}\boldsymbol{\theta}\ \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})^{\beta}\pi(\boldsymbol{\theta}) \\ &= \frac{1}{Z(\beta)}\int \mathrm{d}\boldsymbol{\theta}\ \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})^{\beta}\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \int \mathrm{d}\boldsymbol{\theta}\ P(\boldsymbol{\theta}|\mathbf{D}, \beta)\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}) \equiv \mathbb{E}_{\boldsymbol{\theta}}\Big[\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\ \Big|\ \beta\Big]\end{aligned}$$

This is the slightly weird expectation value of $\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})$, over all possible values of the parameters, weighted by their posterior probabilities. For a given value of β, this can be estimated using an MCMC chain, where at each point in the chain we evaluate $\ln\mathcal{L}(\mathbf{D}|\mathbf{x}_i)$, and once the chain converges we take the average.

This gives us values of $\mathrm{d}\ln Z(\beta)/\mathrm{d}\beta$ at various values of β. We can then numerically integrate this, for example using the trapezium rule, giving:

$$\ln Z = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}}\Big[\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\ \Big|\ \beta\Big]\ \mathrm{d}\beta \approx \frac{1}{2}\sum_{j=1}^{n}\Big(\mathbb{E}_{\boldsymbol{\theta}}\Big[\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\ \Big|\ \beta_i\Big] + \mathbb{E}_{\boldsymbol{\theta}}\Big[\ln\mathcal{L}(\mathbf{D}|\boldsymbol{\theta})\ \Big|\ \beta_{i+1}\Big]\Big)\Delta\beta_i$$

where in the first step we use $Z(\beta = 0) = 1 \Rightarrow \ln Z(\beta = 0) = 0$, and $Z(\beta = 1) = Z$.

This method is practically quite difficult to implement, due to the large number of MCMC chains that need to be run. Some time can be saved in the burn-ins by using the end of one chain (with $\beta = \beta_i$) as the start of the next chain ($\beta = \beta_{i+1}$), as the two will have similar stationary distributions.

[5] This would also make MCMC walkers less likely to get stuck in local minima, which is nice. One might initially start with low β, allowing the walkers to explore the parameter space, before increasing it again to 1 and allowing the walkers to properly characterise the space.

3.2 Computing Evidence Ratios

Recall that for Bayesian model comparison, we are actually only interested in the evidence *ratio* between two models. There are a few tricks which avoid calculating the actual evidences for each model, and skip straight to calculating the evidence ratio.

3.2.1 Savage-Dickey Method

The Savage-Dickey method can be used with two models where one is *nested* inside another, that is, the parameter space of one is a subspace of that of the other. Let the parameters of model B be $\boldsymbol{\theta}_A = (\epsilon, \boldsymbol{\phi})$, and model A have the same parameters $\boldsymbol{\phi}$, but set $\epsilon = 0$. We also require the priors on the shared parameters to be "consistent", that is, $\pi(\boldsymbol{\phi}|A) = \pi(\boldsymbol{\phi}|\epsilon = 0, B)$. In this case, the evidence for the simpler model A is given by:

$$\begin{aligned} Z_A \equiv P(\mathbf{D}|A) &= \int \mathrm{d}\boldsymbol{\phi}\ \mathcal{L}(\mathbf{D}|A, \boldsymbol{\phi})\pi(\boldsymbol{\phi}|A) \\ &= \int \mathrm{d}\boldsymbol{\phi}\ \mathcal{L}(\mathbf{D}|B, \boldsymbol{\phi}, \epsilon = 0)\pi(\boldsymbol{\phi}|\epsilon = 0, B) \\ &= P(\mathbf{D}|\epsilon = 0, B) \\ &= \frac{P(\epsilon = 0|\mathbf{D}, B)\ P(\mathbf{D}|B)}{P(\epsilon = 0|B)} = \frac{P(\epsilon = 0|\mathbf{D}, B)}{P(\epsilon = 0|B)} Z_B \end{aligned}$$

where to get to the final line we use Bayes' theorem. In this way, we can get the evidence ratio:

$$\frac{Z_A}{Z_B} = \frac{P(\epsilon = 0|\mathbf{D}, B)}{P(\epsilon = 0|B)}$$

which depends only on the second model. More specifically, this is the ratio of the posterior probability of ϵ being 0 to the prior probability of ϵ being 0, which is intuitive. If the prior isn't too bad, the denominator will be fine. The numerator, however, requires estimation using MCMC.

3.2.2 Augmented Model Method

Suppose models A and B have parameters $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$. We then define an *augmented* model C, with parameters $\boldsymbol{\theta}_A$, $\boldsymbol{\theta}_B$, and ϵ, which have priors and likelihood:

$$\pi(\boldsymbol{\theta}_A, \boldsymbol{\theta}_B, \epsilon|C) = \mathbb{1}_0^1(\epsilon)\ \pi(\boldsymbol{\theta}_A|A)\ \pi(\boldsymbol{\theta}_B|B)$$

$$\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}_A, \boldsymbol{\theta}_B, \epsilon; C) = \begin{cases} \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}_A; A) & \epsilon < \frac{1}{2} \\ \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}_B; B) & \epsilon > \frac{1}{2} \end{cases}$$

We then use MCMC to sample from the posterior of the augmented model; we show below that the fractions of the samples on either side of $\epsilon = 1/2$ give the evidence ratio. Consider the fraction of the samples which will be on the side $\epsilon < 1/2$. Assuming the MCMC chain has converged, this will be equal to

$$P(\epsilon < 1/2|\mathbf{D}, C) = \int_0^{1/2} \mathrm{d}\epsilon \iint \mathrm{d}\boldsymbol{\theta}_A\, \mathrm{d}\boldsymbol{\theta}_B\, P(\boldsymbol{\theta}_A, \boldsymbol{\theta}_B, \epsilon|\mathbf{D}; C)$$

$$
\begin{aligned}
&= \int_0^{1/2} \mathrm{d}\epsilon \iint \mathrm{d}\boldsymbol{\theta}_A \, \mathrm{d}\boldsymbol{\theta}_B \, \frac{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}_A, \boldsymbol{\theta}_B, \epsilon; C)\pi(\boldsymbol{\theta}_A, \boldsymbol{\theta}_B, \epsilon|C)}{Z_C} \\
&= \frac{1}{Z_C} \int_0^{1/2} \mathrm{d}\epsilon \iint \mathrm{d}\boldsymbol{\theta}_A \, \mathrm{d}\boldsymbol{\theta}_B \, \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}_A; A)\pi(\boldsymbol{\theta}_A|A)\pi(\boldsymbol{\theta}_B|B) \\
&= \frac{1}{2Z_C} \int \mathrm{d}\boldsymbol{\theta}_A \, \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}_A; A)\pi(\boldsymbol{\theta}_A|A) \\
&= \frac{Z_A}{2Z_C}
\end{aligned}
$$

Similarly, the fraction of samples with $\epsilon > 1/2$ will be $Z_B/2Z_C$. The evidence ratio Z_A/Z_B will therefore be the fraction of samples with $\epsilon > 1/2$ divided by the fraction with $\epsilon < 1/2$. This method can be extended to compare any number of models, though the high dimensionality may lead to difficulties with the MCMC.

Sometimes, all of the samples may end up on one side of $\epsilon = 1/2$. This is a bit of a problem, but at least one can then put an upper/lower bound on the evidence ratio.

3.2.3 Importance Sampling

Suppose that models A and B have the same sets of parameters and priors $\pi(\boldsymbol{\theta}|A) = \pi(\boldsymbol{\theta}|B)$. In this case, we can calculate the evidence ratio as follows:

$$
\begin{aligned}
Z_B \equiv P(\mathbf{D}|B) &= \int \mathrm{d}\boldsymbol{\theta} \, \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; B)\pi(\mathbf{D}|B) \\
&= \int \mathrm{d}\boldsymbol{\theta} \, \frac{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; B)}{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; A)} \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; A)\pi(\mathbf{D}|A) \\
&= \int \mathrm{d}\boldsymbol{\theta} \, \frac{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; B)}{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; A)} P(\boldsymbol{\theta}|\mathbf{D}; A) Z_A \\
\Rightarrow \frac{Z_B}{Z_A} &= \int \mathrm{d}\boldsymbol{\theta} \, \frac{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; B)}{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; A)} P(\boldsymbol{\theta}|\mathbf{D}; A) \\
&= \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; B)}{\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}; B)} \right]
\end{aligned}
$$

where $\boldsymbol{\theta} \sim P(\boldsymbol{\theta}|\mathbf{D}; A)$ is distributed according to the posterior assuming model A. This can be estimated using MCMC under model A.

In practice, the above methods aren't used very often to calculate the evidence. Really, everybody uses...

3.3 Nested Sampling

Nested sampling directly calculates the evidence Z, not a ratio between two models. It is particularly good in high-dimensional parameter spaces, and for non-Gaussian or multimodal posteriors. It also produces stochastic samples, like an MCMC method. We will thus revert to describing the parameters $\boldsymbol{\theta}$ by a sample vector $\mathbf{x}$ in the sample space $\mathcal{X}$.

3.3.1 Theory

Let L correspond to a particular value of the likelihood. Consider the subspace $\mathcal{Y}(L) \subseteq \mathcal{X}$ where all the samples $\mathbf{x} \in \mathcal{X}$ have a likelihood greater than L:

$$\mathcal{Y} = \{\mathbf{x} \in \mathcal{X} : \mathcal{L}(\mathbf{D}|\mathbf{x}) > L\}$$

So that $\mathcal{Y}(L = 0) = \mathcal{X}$; if we let $\mathcal{L}^* = \max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})$ be the maximum likelihood value over the parameter space, then $\mathcal{Y}(L = \mathcal{L}^*) = \varnothing$. $\mathcal{Y}$ may be several disjoint regions of the sample space if the likelihood surface has an mountainous topography.

We now define $\xi(L)$ to be the prior probability that $\mathcal{L} > L$, in other words the integral of the prior over $\mathcal{Y}(L)$:

$$\xi(L) \equiv \int_{\mathcal{Y}(L)} \pi(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\{\mathbf{x}:\mathcal{L}(\mathbf{D}|\mathbf{x})>L\}} \pi(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathcal{L}>L} \pi(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

such that $\xi(0) = 1$ and $\xi(\mathcal{L}^*) = 0$. Now from the definition of $\xi(L)$, it must be monotonically decreasing, and hence invertible to $L(\xi)$ on the range $L \in [0, \mathcal{L}^*]$.

In many dimensions, only a very small fraction of the parameter space has a likelihood which is anywhere near $\mathcal{L}^*$. Consider the region where $\xi(L)$ is close to 0, i.e where L is close to $\mathcal{L}^*$. This is an elite, select region, where only the parameters with the very highest likelihoods are included. Even if we increase ξ only slightly, to do so would be to include a lot more of the parameter space with more modest values of the likelihood; L thus needs to decrease by quite a lot to bring about a small increase in ξ from 0. Thus the graph of $L(\xi)$ against ξ is peaks very strongly near $\xi \gtrsim 0$, before intersecting the axis at $L(\xi = 0) = \mathcal{L}^*$.

How does any of this relate to the evidence? Consider the iso-likelihood surfaces in $\mathcal{X}$. These will also be iso-ξ surfaces, since $\xi(L)$ is invertible. Consider then the infinitesimal region $\mathrm{d}\mathcal{Y}$ in between two iso-surfaces, where L takes values in between L and $L + \mathrm{d}L$, and ξ is between $\xi(L)$ and $\xi + \mathrm{d}\xi = \xi(L + \mathrm{d}L)$. In $\mathrm{d}\mathcal{Y}$, the total prior probability is $\mathrm{d}\xi = \int_{\mathcal{Y}} \pi(\mathbf{x})\,\mathrm{d}\mathbf{x}$, recalling that ξ is the integral of the prior over some region. Also, the likelihood in this region is approximately L, by definition. thus the contribution of this region to the evidence, which recall is $Z = \int \mathcal{L}\pi\,\mathrm{d}\mathbf{x} = \int \mathcal{L}\mathrm{d}\xi$, is therefore $\mathrm{d}Z = L\,\mathrm{d}\xi$, because in $\mathrm{d}\mathcal{Y}$ the likelihood is approximately L. Therefore, to calculate the total evidence Z, we simply integrate $L(\xi)\,\mathrm{d}\xi$ over all possible values of ξ, i.e. from 0 to 1. We thus have:

$$Z = \int_0^1 L(\xi)\,\mathrm{d}\xi$$

We have thus converted the high-dimensional integral to a one-dimensional integral, at the cost of having to find and invert $\xi(L)$. This is tackled by the nested sampling algorithm, outlined below.

3.3.2 Algorithm

A large number $N_{\text{live}} \sim 10^3$-$10^4$ of "live points" are drawn from the prior distribution: $\mathbf{x}_j \sim \pi$, for $j = 1, 2, ..., N_{\text{live}}$. The likelihood is then evaluated at all of these points $L_j = \mathcal{L}(\mathbf{D}|\mathbf{x}_j)$, and whichever point $\mathbf{x}^\dagger$ has the lowest value of the likelihood $L^\dagger = \min_j L_j$ is "killed". That point is then replaced by a new point, which is drawn from the prior *subject to* having a larger likelihood than the point which was just killed:

$$\mathbf{x}^\dagger \leftarrow \mathbf{x} \sim \pi\Big(\mathbf{x}\Big|\mathcal{L}(\mathbf{D}|\mathbf{x}) > L^\dagger\Big)$$

In this way, the live points gradually crowd in towards the maximum (or maxima) of the likelihood function.

At each step i, the dead likelihood $L_i^\dagger$ is recorded, forming an increasing sequence: $L_i^\dagger > L_{i-1}^\dagger$. We also record the associated values of $\xi_i^\dagger$, which form a decreasing sequence. This is done probabilistically. Because the surviving points have been sampled according to the prior, the distribution of *their* values of ξ_j will be uniformly distributed from 0 to $\xi_i^\dagger$. Therefore, after every death, the next largest ξ_j, that is, $\max\{\xi_j\} = \xi_{i+1}^\dagger$ (the point which will die at the next step) will only be a fraction $t \in [0, 1]$ of $\xi_i^\dagger$, for some t which is a random variable following the distribution of the largest of N_{live} random variables $\sim \mathcal{U}(0, 1)$. If N_{live} is very large, then t will be only slightly below 1; in more detail, how is t distributed? Let $t = \max\{\tau_i\}$, where each $\tau \sim \mathcal{U}(0, 1)$. The probability that t is less than some value T is the probability that all of the τ_i are less than T, that is, $P(t < T) = T^{N_{\text{live}}}$. Differentiating, we find $P(t) = N_{\text{live}} t^{N_{\text{live}}-1}$. From this, we can find that $\mathbb{E}[\ln t] = -1/N_{\text{live}}$, and thus approximate that at each iteration $\xi_i^\dagger$ decreases by a factor of $\exp(-1/N_{\text{live}})$. Thus $\xi_i^\dagger \approx \exp(-i/N_{\text{live}})$.

In this way, the dead points give us lots of values of L_i (the $L_i^\dagger$) and corresponding values of ξ_i (the $\xi_i^\dagger$), as ξ_i works its way down from 1 to 0. We can use these values to estimate the evidence integral, $Z = \int_0^1 L\,\mathrm{d}\xi$, using the trapezium rule. In fact, we typically integrate "as we go", so we start with $(\xi_0, L_0) = (1, 0)$ and work our way *down* in ξ. We therefore have:

$$Z \approx \sum_{i=0}^{M-1} \frac{1}{2}(L_i + L_{i+1})(\xi_{i+1} - \xi_i) = \sum_{i=1}^{M} w_i L_i; \qquad w_i \equiv \frac{1}{2}(\xi_{i-1} - \xi_{i+1})$$

where we do some summation magic to convert to a simpler weighted sum over the L_i, with the weights defined as shown[6].

When should the sequence be terminated? One usually estimates the evidence contribution from the remaining live points at likelihoods L_j to be $\Delta Z \approx \xi_i \cdot \max L_j$. When this drops below some small tolerance value, we say that the integral has approximately converged.

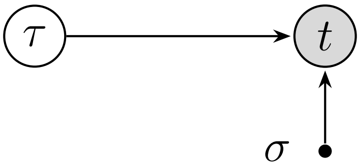
Apparently, the killed points $\mathbf{x}_i^\dagger$ form a set of stochastic samples from the posterior distribution, albeit weighted by $w_i L_i$.

[6] Note that for the summation to work, and not have any leftover terms, we need to set $\xi_{M+1} = \xi_M$. We're only running it for M goes though, so ξ_{M+1} doesn't really exist.

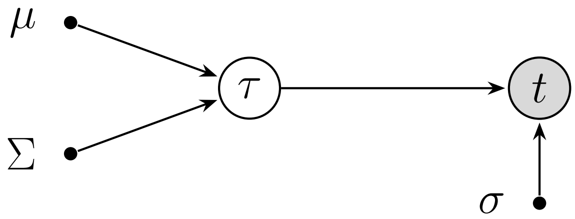
4 Hierarchical Bayesian Models (HBMs)

Bayes' Theorem often deals with conditional probabilities $P(X|Y)$. These probabilities can be chained together if Y itself depends on other quantities, a and b, say. Such multi-layer models constitute HBMs.

These are often displayed as *probabilistic graphical models* (PGMs): diagrams showing the conditional flow of a HBM. In these diagrams, unknown quantities are shown in circles, with an arrow towards the dependent quantity. Circles representing data are shaded; unknown random variables (whether parameters, or latent/nuisance parameters) are empty Known parameters, if displayed at all, are represented as points. For example, the likelihood function $\mathcal{L}(t|\tau, \sigma = 1)$ would be represented by the following PGM:



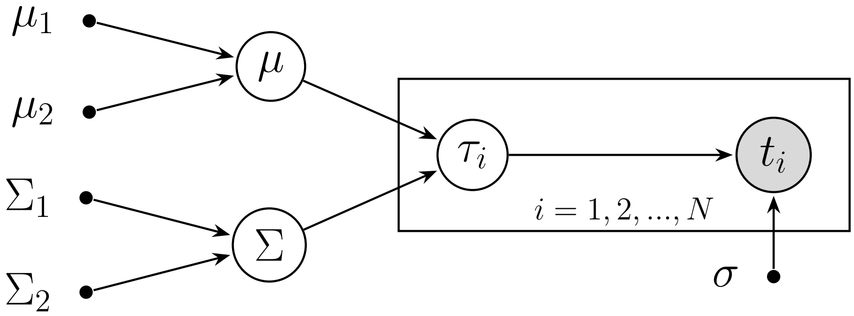
Suppose now that we have a prior on τ: it depends on the (fixed) parameters μ and Σ. The diagram would then become:



The point of these diagrams is that they allow us to follow the Bayesian flow when we are writing down quantities like the evidence, which the above diagram would help us to write as:

$$Z = \int \mathcal{L}(t|\tau, \sigma)\ \pi(\tau|\mu, \Sigma)\,\mathrm{d}\tau$$

Suppose now that μ and Σ are now random variables, with *hyperpriors* that depend on μ_1, μ_2, and Σ_1, Σ_2. Suppose also that there are a large number N of different random variables τ_i, on each of which the quantity t_i depends. The diagram would then be shown as:



This is a truly hierarchical Bayesian model, as it assumes that each of the τ_i depends on the same parameters μ and Σ one level below.

5 Gaussian Processes

For a real function $f(\mathbf{x})$, and some data $(\mathbf{x}_i, f_i)$, Gaussian processes (GPs) allow one to both interpolate and extrapolate beyond the $\mathbf{x}_i$.

Consider a space $\mathcal{S}$, which could be $\mathbb{R}^d$, but could be any space, even a torus or something. A GP on $\mathcal{S}$ is a collection of random variables such that for any collection of points $\{\mathbf{x}_i \in \mathcal{S}\}$ in this space, the vector

$$\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)) \in \mathbb{R}^N$$

is a random variable distributed as a multivariate Gaussian. This is denoted $f \sim \mathcal{GP}(\mu, k)$, for mean and *kernel functions* $\mu : \mathcal{S} \to \mathbb{R}$ and $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. When designing a Gaussian process, we take a set of points $\mathbf{x}$, choose a kernel function, and calculate the mean and covariance of the distribution: the i-component of the mean of $\mathbf{f}$ in this distribution is given by $\mu(\mathbf{x}_i)$; the i, j-component of the covariance matrix is $k(\mathbf{x}_i, \mathbf{x}_j)$ (the function k therefore must be symmetric). Then, by drawing random samples $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from the multivariate distrbution, we obtain points of a function which has effectively been drawn from the Gaussian process.

The mean function is essentially arbitrary, but the kernel must be positive-definite, for the same reason that a multivariate Gaussian distribution must have a positive-definite covariance matrix. This significantly limits the types of functions eligible to be kernel functions, as to qualify they have to produce positive-definite covariance matrices for any set of N points in the space. A few common ones are discussed below.

5.1 Examples of Kernel Functions

5.1.1 Linear Kernel

The linear kernel is $k_L(x, x') = xx'$, and is positive semi-definite. To show this, note that the covariance matrix produced by the linear kernel is $\boldsymbol{\Sigma} = \mathbf{x}\mathbf{x}^\intercal$. Multiplying this matrix by the vector $\mathbf{x}$ gives $\boldsymbol{\Sigma}\mathbf{x} = \mathbf{x}\mathbf{x}^\intercal\mathbf{x} = |\mathbf{x}|^2\mathbf{x}$, so $\mathbf{x}$ is an eigenvalue of $\boldsymbol{\Sigma}$, with eigenvalue $|\mathbf{x}|^2 > 0$. All other eigenvectors of this matrix $\mathbf{y}_i$ will be perpendicular to $\mathbf{x}$, and because $\boldsymbol{\Sigma}\mathbf{y}_i = \mathbf{x}\mathbf{x}^\intercal\mathbf{y} = \mathbf{x}(\mathbf{0})$, their eigenvalues are all 0. Thus $\boldsymbol{\Sigma}$'s eigenvalues are $|\mathbf{x}|^2$, and 0, so $\boldsymbol{\Sigma}$ is positive semi-definite.

Functions from the (zero-mean) Gaussian process with a linear kernel are always lines of constant gradient. To show this, let $\mathbf{x}$ sample the x-axis (perhaps $x_k = k\delta$ for some small value of δ). The samples from the resulting multivariate Gaussian are $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{x}\mathbf{x}^\intercal)$. Consider now $\mathbf{z} = a\mathbf{x}$, where $a \sim \mathcal{N}(0, 1)$; we now show that the $\mathbf{y}$s thus drawn follow the same distribution. The mean of the distribution of z_i in this latter case is $\mathbb{E}[z_i] = \mathbb{E}[a]x_i = 0$; the covariance of z_i and y_j is $\mathrm{cov}[y_i, y_j] = \mathrm{cov}[ax_i, ax_j] = \mathbb{E}[a^2]x_ix_j = x_ix_j = \Sigma_{ij}$. Thus $\mathbf{z}$ and $\mathbf{y}$ follow the same distribution, so $\mathbf{y} = a\mathbf{x}$, where $a \sim \mathcal{N}(0, 1)$. All of the draws $\mathbf{y}$ from the multivariate distribution of a GP with a linear kernel will therefore be multiples of the vectors $\mathbf{x}$ that are input to the GP, and therefore a plot of $\mathbf{y}$ against $\mathbf{x}$ will be linear.

5.1.2 Brownian Motion Kernel

The Brownian motion kernel is $k_{\mathrm{BM}}(x, x') = \min(x, x')$, and so called because it produces samples with continuous-undifferentiable-looking forms. To show that this kernel is positive semi-definite, we rewrite the kernel as:

$$k_{\mathrm{BM}}(x, x') = \int_0^\infty \mathrm{d}t\, [1 - H(x - t)][1 - H(x' - t)]$$

where $H(t;x)$ is the Heaviside step function: the integrand will only be positive where both $x < t$ and $x' < t$, i.e. only where $\min(x, x') < t$; the integral is then 1 between 0 and $\min(x, x')$, giving $\min(x, x')$ as a result; this integral is thus equivalent to the Brownian motion kernel. We then pre- and post-multiply the resultant covariance matrix with a general vector $\mathbf{z}$, and show that the resulting scalar is always non-negative:

$$\mathbf{z}^\intercal \boldsymbol{\Sigma} \mathbf{z} = \int_0^\infty \mathrm{d}t \, [1 - H(x_i - t)][1 - H(x_j - t)] z_i, z_j = \int_0^\infty \mathrm{d}t \, [z_i(1 - H(x_i - t))]^2 \geq 0$$

Thus the Brownian motion kernel is positive semi-definite.

5.1.3 Squared Exponential Kernel

The squared exponential kernel is the most popular kernel function, with realisations from the resulting distributions being nice and smooth. The kernel is given by:

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{1}{2}\left|\frac{\mathbf{x} - \mathbf{x}'}{\ell}\right|^2\right]$$

where ℓ is a length scale hyperparameter, controlling the scale of the wiggles in the draws from the distribution.

k_{SE} is an example of a *stationary* kernel, which are those that depend only on $\mathbf{X} = \mathbf{x} - \mathbf{x}'$. For such functions, it is generally easier to tell whether they are valid kernel functions by considering their Fourier transform $\tilde{k}(\boldsymbol{\omega})$:

$$\mathbf{z}^\intercal \boldsymbol{\Sigma} \mathbf{z} = \Sigma_{ij} z_i z_j = \int \mathrm{d}\boldsymbol{\omega} \, \tilde{k}(\boldsymbol{\omega}) \exp\big(-i\boldsymbol{\omega} \cdot (\mathbf{x}_i - \mathbf{x}_j')\big) z_i z_j = \int \mathrm{d}\boldsymbol{\omega} \, \tilde{k}(\boldsymbol{\omega}) |z_i \exp(-i\boldsymbol{\omega} \cdot \mathbf{x}_i)|^2$$

Thus $\mathbf{z}^\intercal \boldsymbol{\Sigma} \mathbf{z} \geq 0$, and the kernel function is positive semi-definite, if the Fourier transform $\tilde{k}(\boldsymbol{\omega}) \geq 0$. For the case of k_{SE}, the Fourier transform of a Gaussian is another Gaussian, which is positive, and so the squared-exponential kernel is positive-definite.

5.1.4 Periodic Kernel

If we want to generate periodic functions, then we can use a periodic kernel. This has the form:

$$k_P(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{2\sin^2\left(\pi|\mathbf{x} - \mathbf{x}'|/T\right)}{\ell^2}\right]$$

with hyperparameters ℓ and also T, which governs the period of the oscillations in the resulting functions.

5.1.5 Constructing New Kernel Functions

It is generally difficult to tell whether a given function constitutes a valid kernel function. However, valid kernels can be combined in various ways to form new kernels which are also valid:

- Adding or multiplying valid kernels gives a valid kernel.
- For a valid kernel $k(\mathbf{x}, \mathbf{x}')$, the function $\alpha(\mathbf{x})\alpha(\mathbf{x}')k(\mathbf{x}, \mathbf{x}')$ is also a valid kernel.
- Given a kernel function $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, and another function $U : \mathcal{S}' \to \mathcal{S}$, and vectors $\mathbf{x}, \mathbf{x}' \in \mathcal{S}'$ the "warped" function $k(U(\mathbf{x}), U(\mathbf{x}'))$ is a valid kernel function.

5.2 Gaussian Process Regression

Suppose there are some values of $\mathbf{x}$ ($\mathbf{x}_i^\dagger$, say) at which we *know* the values of $f(\mathbf{x}_i^\dagger)$, and some values of $\mathbf{x}$ ($\mathbf{x}_i^*$, say) at which we want to interpolate/extrapolate the function. We can do this by concatenating the points $\mathbf{x} = (\mathbf{x}^\dagger, \mathbf{x}^*)$, constructing a Gaussian process based on these points, and then *conditioning* the resulting multivariate distribution on the *known* values of f. We can then obtain estimates and errors on $f(\mathbf{x}_i^*)$.

The engine of this is the fact that a multivariate Gaussian distribution conditioned on certain variables taking any particular values is itself a (multivariate) Gaussian distribution. Imagine slicing a 2D Gaussian perpendicular to one of the axes: the cross-section will too be a Gaussian, no matter where the slice is.

In particular, it turns out that if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean vector is broken up into $\boldsymbol{\mu} = \left(\boldsymbol{\mu}^*, \boldsymbol{\mu}^\dagger\right)$ and similarly the symmetric covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^* & \left[\boldsymbol{\Sigma}^{*\dagger}\right]^\intercal \\ \boldsymbol{\Sigma}^{*\dagger} & \boldsymbol{\Sigma}^\dagger \end{pmatrix},$$

then the conditional distribution $\mathbf{x}^* | \mathbf{x}^\dagger \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$, where

$$\begin{aligned} \boldsymbol{\mu}' &= \boldsymbol{\mu}^* + \left[\boldsymbol{\Sigma}^{*\dagger}\right]^\intercal \left[\boldsymbol{\Sigma}^\dagger\right]^{-1} \left(\mathbf{x}^\dagger - \boldsymbol{\mu}^\dagger\right) \\ \boldsymbol{\Sigma}' &= \boldsymbol{\Sigma}^* - \left[\boldsymbol{\Sigma}^{*\dagger}\right]^\intercal \left[\boldsymbol{\Sigma}^\dagger\right]^{-1} \boldsymbol{\Sigma}^{*\dagger} \end{aligned}$$

Suppose we want to interpolate the value of a 1-dimensional function f at a single point, x^*, knowing the values of f at a selection of points $\mathbf{x}^\dagger = (x_i^\dagger)$. Taking the function to have a Gaussian process prior with zero mean function and kernel function k, we therefore have that the vector containing the function values at x^* and the $x_i^\dagger$ is:

$$\left(f(x^*), f(\mathbf{x}^\dagger)\right) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma^{**} & \left[\boldsymbol{\Sigma}^{*\dagger}\right]^\intercal \\ \boldsymbol{\Sigma}^{*\dagger} & \boldsymbol{\Sigma}^\dagger \end{pmatrix}\right);$$

$$\Sigma^{**} = k(x^*, x^*); \qquad \boldsymbol{\Sigma}_i^{*\dagger} = k(x^*, x_i^\dagger); \qquad \boldsymbol{\Sigma}_{ij}^\dagger = k(x_i^\dagger, x_j^\dagger)$$

With this, and the above formulae for the conditional multivariate distribution, $f(x^*)|f(\mathbf{x}^\dagger)$ follows a univariate normal distribution with mean and variance given by:

$$\begin{aligned} \mathbb{E}[f(x^*)] &= \left[\boldsymbol{\Sigma}^{*\dagger}\right]^\intercal \left[\boldsymbol{\Sigma}^\dagger\right]^{-1} \mathbf{x}^\dagger \\ \mathbb{V}[f(x^*)] &= k(x^*, x^*) - \left[\boldsymbol{\Sigma}^{*\dagger}\right]^\intercal \left[\boldsymbol{\Sigma}^\dagger\right]^{-1} \boldsymbol{\Sigma}^{*\dagger} \end{aligned}$$

Note that, as with everything in Bayesian analysis, this depends on the choice of k: that is, our priors on the form of f.

5.3 Pros and Cons of Gaussian Processes

Gaussian processes have many advantages. Crucially, it enables inclusion of prior information about the function, such as periodicity and smoothness. Functions can be arbitrarily complicated without overfitting. And errors on the interpolated function values can be calculated. Finally, $\mathcal{S}$ doesn't have to be the real line: the above can be equally applied to spaces with bizarre topologies.

The main disadvantage is computational scaling. For N known data points $\mathbf{x}^\dagger$, we must calculate the kernel matrix $\boldsymbol{\Sigma}^\dagger$, and then its inverse, which takes $\mathcal{O}(N^3)$. Sometimes direct inversion of the kernel matrix is avoided by calculating its *Cholesky decomposition*, in terms of the product of a lower-triangular matrix and its transpose.