Bayesian Methods

Alan Heavens, ICIC, Imperial College London
ADA X, Heraklion, Crete. September 2023

September 19, 2023

Contents

1 Books and Syllabus

1.1 Some books for further reading

- D. Silvia & J. Skilling: Data Analysis: a Bayesian Tutorial (CUP). *Nice small book for the basics.*

- P. Saha: Principles of Data Analysis. (Capella Archive)

  `https://www.physik.uzh.ch/~psaha/pda/`

  Similarly, a good, clear, small volume. Free online version as well as a physical book.

- T. Loredo: Bayesian Inference in the Physical Sciences
  `http://www.astro.cornell.edu/staff/loredo/bayes/`
- D. Mackay: Information Theory, Inference and Learning Algorithms. (CUP) `http://www.inference.phy.cam`
  More on the information theory basis of the subject.
- A. Gelman et al: Bayesian Data Analysis (CRC Press) *Comprehensive.*

2 Inverse Problems

- Analysis problems are inverse problems: given some data, we want to infer something about the process that generated the data
- Generally harder than predicting the outcome, given a physical process
- The latter is called forward modelling, or a generative model
- Typical classes of problem:
- Parameter inference
- Model comparison

3 Bayesian Inference

What questions do we want to answer? Parameter Inference:

- I have a set of (x, y) pairs, with errors. If I assume $y = mx + c$, what are m and c?
- I have detected 5 X-ray photons from a source at known distance in the lab. What is the power output of the source and its uncertainty?
- Given LIGO gravitational wave data, what are the masses of the inspiralling objects?

What questions do we want to answer? Model Comparison:

- Do data support General Relativity or Newtonian gravity?
- Is the standard cosmological model (ΛCDM) more probably than (specified) alternatives?
- Do LHC data support the existence of the Higgs boson, or no Higgs boson?

4 The meaning of probability

- Probability describes the relative frequency of outcomes in infinitely long trials (Frequentist view)
- Probability (often) expresses a degree of belief (Bayesian view)

4.1 Probability rules and Bayes theorem

5 Probability rules

- $p(x) + p(\sim x) = 1$ (sum; $\sim$ means not)
- $p(x, y) = p(x|y)\, p(y)$ (product) (the | means 'given'; it is a 'conditional' distribution)
- $p(x) = \sum_k p(x, y_k)$ (**marginalisation** over all possible discrete y_k values)
- $p(x) = \int p(x, y)\, dy$ (marginalisation, continuous variables. $p(\geq 0)$ = probability density function (pdf), s.t. $p(x, y)dxdy$ = probability that x and y occur in an interval $dxdy$ around values x, y. Note that p can be greater than 1 - it is not a probability, but a probability density.)
- Since $p(x, y) = p(y, x) \Rightarrow$ Bayes theorem:
- Bayes Theorem:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

6 Parameter Inference

6.1 Notation

- **Data** d; **Model** M; Model **parameters** θ
- Rule 1: write down what you want to know
- Usually, it is the probability distribution for the parameters, given the data, and assuming a model: i.e. $p(\theta|\mathrm{d}, M)$[1]
- This is the **Posterior**
- To compute it, we use Bayes theorem:

$$p(\theta|d, M) = \frac{p(d|\theta, M)p(\theta|M)}{p(d|M)}$$

- where the **Likelihood** is $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$
- and the **Prior** is $\pi(\theta) = p(\theta|M)$

[1] Sometimes the probabilities are written as begin dependent on any prior information I, i.e. we want $p(\theta|d, M, I)$. We won't include the I explicitly unless we have to for clarity, since it makes the equations look more complicated.

- $p(\mathrm{d}|M)$ is the **(Bayesian) Evidence**, which is important for Model Comparison, but not for Parameter Inference, where its role is simply to normalise the posterior
- Dropping the M dependence for now (we will return to it when we discuss Model Comparison):

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\,\pi(\theta)}{p(d)}$$

In the context of parameter inference (i.e. for a given fixed model M), the Evidence serves only to make the posterior a properly normalised probability distribution as a function of the parameters θ. For continuous parameters (re-introducing M),

$$p(d|M) = \int p(d|\theta, M)\,\pi(\theta)\,d\theta \tag{6.1}$$

where the integral may be multidimensional (multiple parameters).

6.1.1 The Likelihood and the Sampling Distribution

It is important to pause here to think about $\mathcal{L}$. We can view this distribution two ways. If we fix θ (as is rather implied by the expression), then we have the distribution of the data for given θ. This is a proper probability distribution that integrates to unity when integrated over all possible data d. Used this way it is properly called the **Sampling Distribution**.

In Bayesian inference, though, the data are fixed (that is what he have), and this term is treated as a function of θ. In this context, it is called the **Likelihood**, and is not a proper probability distribution, in the sense that integrating it over θ at fixed d does not give unity. Only the full posterior does this.

6.2 How to set up a problem

6.2.1 Analyse the problem:

Everything is focussed on getting at the posterior, $p(\theta|\mathrm{d}) \propto \mathcal{L}(\theta)\,\pi(\theta)$.

What are the data, d?
What is the model for the data?
What are the model parameters?
What is the likelihood function $\mathcal{L}(\theta)$?
What is the prior $\pi(\theta)$?

6.3 Marginalisation

This is a straightforward application of the marginalisation rules, e.g. marginalising over all n parameters except θ_1 and θ_2:

$$p(\theta_1, \theta_2|d) = \int p(\theta_1, \ldots, \theta_n|d)\,d\theta_3 \ldots d\theta_n \tag{6.2}$$

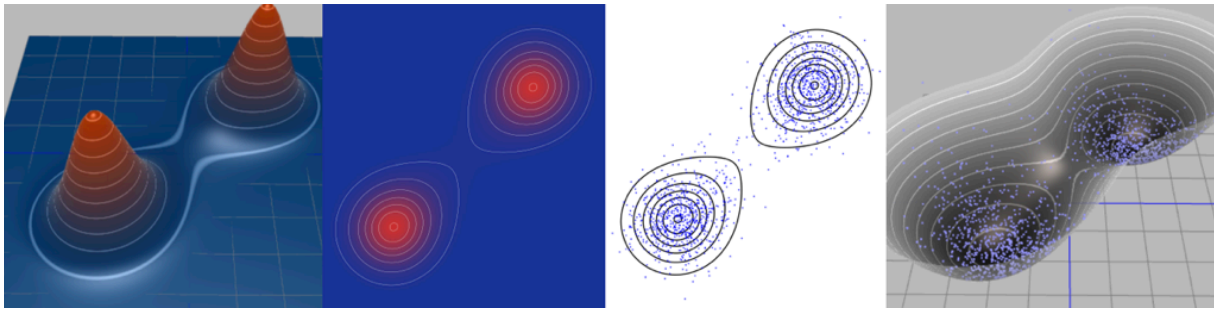

Figure 1: Credit: Alex Rogozhnikov

7 Sampling

The posterior is often not expressible analytically, so it usually needs to be computed numerically. For 1, 2 or 3 dimensions, evaluating it on a grid in parameter space is usually effective, but this becomes prohibitively expensive as the dimensionality increases, so another technique is needed. This is to use a completely different representation of the posterior $p(\theta)$: a large number of samples drawn from the distribution, with (expected) density that is proportional to $p(\theta)$. This is usually in an ordered list, called a 'chain', of values of the parameters θ. The samples may also have a weight associated with them, and are constructed such that the expected weighted number density is proportional to the posterior. Note that we don't need to calculate the constant of proportionality (which can be expensive to do), since in parameter inference problems, the relative probability of parameters is given by the ratio of p.

The reason why the list is ordered is that the algorithms for generating the chain typically produce correlated samples, so the ordering is important (one might, for example, want to 'thin' the chain by selecting only separated samples, thus reducing the correlations. If the samples are correlated, then the 'effective sample size' is smaller than the length of the chain.

The samples effectively replace the continuous density p by a (weighted) sum of Dirac delta functions:

$$p(\theta) \simeq \frac{\sum_{s=1}^{S} w_s \delta(\theta - \theta_s)}{\sum_{s=1}^{S} w_s}. \tag{7.3}$$

This is clearly crude for p itself, but for integrated quantities, it makes sense. e.g. an estimate of the expectation value is

$$\hat{\mu} = \langle \theta \rangle = \int p(\theta)\theta \, d\theta \simeq \int \frac{\sum_{s=1}^{S} w_s \delta(\theta - \theta_s)}{\sum_{s=1}^{S} w_s} \theta \, d\theta = \frac{\sum_{s=1}^{S} w_s \theta_s}{\sum_{s=1}^{S} w_s}. \tag{7.4}$$

8 Sampling methods

There are several generic methods for generating samples. We will concentrate on three of the most common ones, highlighting when each of them can usefully be applied. They are:

- Metropolis-Hastings
- Gibbs Sampling
- Hamiltonian (or Hybrid) Monte Carlo (HMC)

First, though, some general remarks.

8.1 Markov Chain Monte Carlo (MCMC)

These are all examples of MCMC (Markov Chain Monte Carlo), where random steps are taken in parameter space, according to a proposal distribution. The goal is always to give a chain of samples of the target distribution (usually the posterior or the likelihood), with an expected number density proportional to the posterior. The target distribution need not be normalised, but it needs to be everywhere positive, and normalisable (i.e. the integral is finite).

8.1.1 Markov processes

Markov processes are sequential processes for which the new element depends only on the previous element, and not on any previous ones. In MCMC, the next point in the chain depends only on the parameters (and the target value) of the previous point.

The general algorithm is as follows:

- Choose a starting point θ_0. No general rule here, but (see later) there are advantages in having a 'dispersed' starting point, which is not near the peak of the target distribution (for convergence tests). A random point drawn from a prior distribution is common.
- Subsequent points θ_{s+1} are generated from θ_s by generating a trial point through some random process, and which is either accepted or rejected (depending on the algorithm)[2]
- If the trial point is accepted, it becomes the next point in the chain. If it is rejected, the previous sample is repeated in the chain (or equivalently, its weight is increased from 1 to 2 (and can go higher if subsequent trials are also rejected).
- The chain is stopped at some point. There is no magic answer as to when to stop, but the main point it that you need to test for convergence.

8.1.2 Detailed balance

If the sampling procedure satisfies detailed balance, the expected number density to be proportional to the target distribution $p(\theta)$, which is what we desire.[3] We don't want the target distribution ρ to evolve as the chain develops, so it is a stationary distribution. In Bayesian inference problems, the target is sometimes the posterior, sometimes the likelihood, and it can be something different again.

Let us assume there is a discrete set of parameters (the argument generalises to continuous parameters), labelled by an index (it can still be a label in a multi-dimensional parameter space). As we move from one sample to the next in the chain, there is a probability that the state shifts from i to j given by P_{ij}. The MCMC chain satisfies detailed balance if

$$\rho_i P_{ij} = \rho_j P_{ji}. \tag{8.5}$$

One can think of the left hand side as being the flux of probability flowing from i to j, and the r.h.s. from j to i. If they balance, the chain is stationary.

[2] Some algorithms, such as Gibbs, may always accept, dependent on some factors.

[3] For weighted samples, with weights w_s, we want the density of points to be proportional to p/w_s.

Detailed balance is a stronger condition than that required to give a stationary distribution (which can be achieved via a more complicated route).

Proof: if we have samples drawn from a density distribution ρ_i, then after a transition, the probability distribution changes to an expected value ρ_j given by

$$\sum_{\text{all } i} \rho_i P_{ij} \tag{8.6}$$

including transitions to j from all other states i. If detailed balance is satisfied, this is $\sum_i \rho_j P_{ji} = \rho_j \sum_i P_{ji} = \rho_j$, since, in the last step, the state j must end up in some i, so the sum of probabilities is 1. So the expected density stays as ρ and does not change.

8.2 Metropolis-Hastings algorithm

This is perhaps the most common form of MCMC, and is suitable for relatively low-dimensional problems (perhaps up to 5 or 10). We define a proposal distribution to generate a new proposed sample, which is either accepted or rejected.

$$q_{ij} \equiv q(\theta_j|\theta_i) \tag{8.7}$$

= probability of a proposed sample at θ_j from a previous state θ_i. Typically this is a function of $\theta_j - \theta_i$, but it doesn't have to be, and a common choice is a gaussian centred on the previous sample in the chain.

The algorithm specifies that the point is accepted with probability

$$\alpha = \min\left[1, \frac{\rho(\theta_j)q(\theta_i|\theta_j)}{\rho(\theta_i)q(\theta_j|\theta_i)}\right] = \min\left[1, \frac{\rho_j q_{ji}}{\rho_i q_{ij}}\right] \tag{8.8}$$

Let us see if this satisfies detailed balance. For concreteness, let us assume (in a compact notation) that

$$\rho_j q_{ji} \leq \rho_i q_{ij} \tag{8.9}$$

The probability of an accepted transition from i to j is

$$P_{ij} = q_{ij}\min\left[1, \frac{\rho_j q_{ji}}{\rho_i q_{ij}}\right] = \frac{\rho_j q_{ji}}{\rho_i} \tag{8.10}$$

where the first term is the probability that the transition is proposed, and the second is the probability that it is accepted. To check detailed balance, we need to compute the reverse probability, which is

$$P_{ji} = q_{ji}\min\left[1, \frac{\rho_i q_{ij}}{\rho_j q_{ji}}\right] = q_{ji} \tag{8.11}$$

since the proposed sample is accepted with probability 1 in this case. Hence the detailed balance relation is satisfied, $\rho_i P_{ij} = \rho_j P_{ji}$, with Metropolis-Hastings. Note that if q is symmetric, (i.e. $q_{ij} = q_{ji}$), the acceptance probability is simplified, and the algorithm is called Metropolis.

Remember! If the proposed sample is rejected, the previous sample is **repeated** in the chain (or equivalently, its weight is increased from 1 to 2 (and to 3 if the next proposed sample is also rejected, and so on).

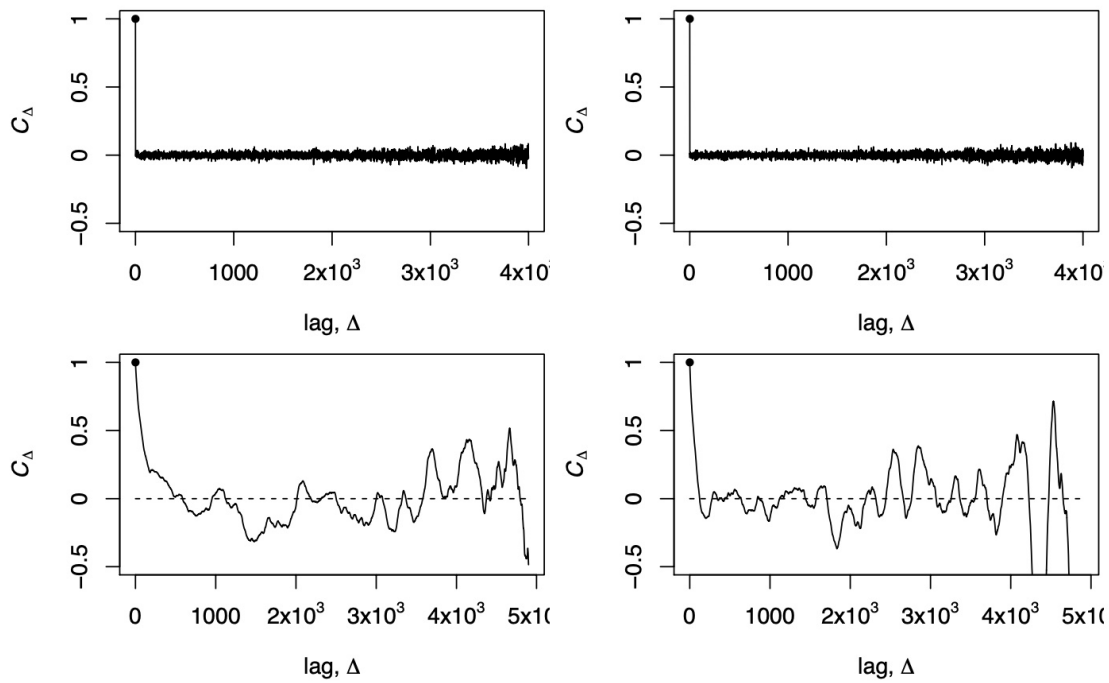

Figure 2: Correlation coefficient of samples for uncorrelated samples (top) and badly-correlated samples (bottom). From D. Mortlock.

In lectures, we will discuss what issues to consider in choosing a proposal distribution. As a rule of thumb, an acceptance rate of ~ 0.3 is usually optimal.

8.3 Marginalisation from samples

This is trivial to do. Each sample has values for all of the parameters. If we want the distribution of θ_1 say, then we simply ignore the values of $\theta_i, i > 1$ in the chain, and plot the distribution of θ_1. A potentially conceptually hard multidimensional integral is solved very easily.

8.4 Correlated samples

Some sampling algorithms will produce correlated samples from the posterior (in fact this is normal). If nearby samples in the chain are correlated, the effective number of independent samples is smaller than the total number of samples. We can quantify this with the autocorrelation function, estimated by

$$\hat{C}_\Delta \equiv \frac{1}{S-\Delta} \sum_{s=1}^{S-\Delta} \frac{(\theta_s - \hat{\mu})(\theta_{s+\Delta} - \hat{\mu})}{\hat{\Sigma}} \tag{8.12}$$

where $\hat{\mu}$ is the estimate of the mean parameter (in practice, just the weighted average), and $\hat{\Sigma}$ is the estimated variance. We compute this for every parameter in the problem. Note that $\hat{C}_0 = 1$, and ideally we'd like $\hat{C}_\Delta$ to be zero otherwise. Fig. 2 shows some examples.

8.4.1 Effective sample size

The effective number of independent samples will be smaller than S if the chain is correlated. One definition of the effective sample size is

$$S_{\text{eff}} \equiv \frac{1}{1 + 2\sum_{\Delta=1}^{\Delta_0 - 1} \hat{C}_\Delta}, \tag{8.13}$$

and Δ_0 is the point where $\hat{C}_\Delta$ crosses zero for the first time.

8.5 Gibbs sampling

This is a powerful technique that is useful if the conditional distributions are known.

Algorithm:

- $\theta_1^{s+1} \sim p(\theta_1 | \theta_2^s, \theta_3^s, \ldots, \theta_m^s)$
- $\theta_2^{s+1} \sim p(\theta_2 | \theta_1^{s+1}, \theta_3^s, \ldots, \theta_m^s)$
- etc ...

Repeat, randomizing (or reversing) the order.

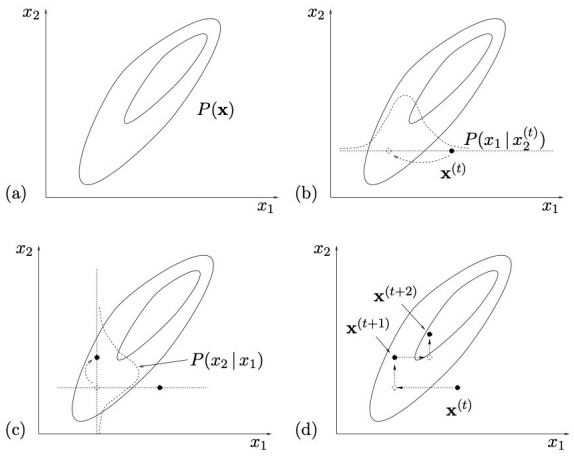

Figure 3: Illustration of Gibbs sampling (from Mackay 2003)

Sometimes this can be applied to very high-dimensional problems (millions). All samples are accepted, if the conditional distributions can be analytically sampled. (Otherwise, rejection sampling can often be employed). Can be slow if parameters are highly-correlated. Often useful for Bayesian Hierarchical Models (see later).

8.6 Hamiltonian Monte Carlo

This is an extremely powerful technique that can be applied to very high-dimensional problems as well. The snag is that it requires derivatives of the target function with respect to the model parameters. Nowadays, automatic differentiation techniques can help.

It is a neat idea, that treats the target distribution as a potential, and the samples are created by solving Hamilton's equations for particles orbiting in the potential. The particles are given a momentum, and move around. After some time, a new proposed sample is generated. The advantage is that, by taking advantage of knowing something about the shape of the target distribution (through the derivatives) it can move a long way across the target distribution ('good mixing') whilst still accepting most of the proposed samples.

HMC defines a potential

$$U(\boldsymbol{\theta}) = -\ln p(\boldsymbol{\theta}) \tag{8.14}$$

where $p(\boldsymbol{\theta})$ is the target distribution. Think of $\boldsymbol{\theta}$ as being the position ($\boldsymbol{\theta}$ represents a vector $(\theta_1, \ldots, \theta_n)$).

There is also a kinetic energy

$$K(\mathbf{u}) = \frac{1}{2}\mathbf{u} \cdot \mathbf{u} \tag{8.15}$$

where $\mathbf{u}$ is the momentum, with $u_i \sim \mathcal{N}(0, \sigma^2)$ for some variance σ^2 (often taken to be unity).

The Hamiltonian (energy) is

$$H(\boldsymbol{\theta}, \mathbf{u}) = U(\boldsymbol{\theta}) + K(\mathbf{u}) \tag{8.16}$$

We have defined a new parameter space that is twice as large as the original, and we define a new target distribution in the $2n-$dimensional space:

$$T(\boldsymbol{\theta}, \mathbf{u}) = \exp[-H(\boldsymbol{\theta}, \mathbf{u})]. \tag{8.17}$$

HMC explores this phase space using Hamilton's equations:

$$\begin{aligned} \dot{\theta}_i &= \frac{\partial H}{\partial u_i} = u_i \\ \dot{u}_i &= -\frac{\partial H}{\partial \theta_i} = \frac{\partial \ln p}{\partial \theta_i} \end{aligned} \tag{8.18}$$

The equations normally need to be solved numerically, using an integration scheme (which needs to be symmetric forward-back, to satisfy detailed balance. A common choice is the leapfrog method (which is forward-backward symmetric, as required for detailed balance).

After the orbit is integrated for a while, a new proposed sample is generated, and accepted or rejected, then a new random momentum is generated and the procedure repeated.

For HMC, the full algorithm is (from Hajian 2006):

Hamiltonian Monte Carlo

initialize $\theta_{(0)}$
for i = 1 to $N_{samples}$
    $\mathbf{u} \sim \mathcal{N}(0,1)$ (Normal distribution)
    $(\theta^*_{(0)}, \mathbf{u}^*_{(0)}) = (\theta_{(i-1)}, \mathbf{u})$
    for j = 1 to N
        make a leapfrog move: $(\theta^*_{(j-1)}, \mathbf{u}^*_{(j-1)}) \to (\theta^*_{(j)}, \mathbf{u}^*_{(j)})$
    end for
    $(\theta^*, \mathbf{u}^*) = (\theta_{(N)}, \mathbf{u}_{(N)})$
    draw $\alpha \sim$ Uniform(0,1)
     if $\alpha < \min\{1, e^{-(H(\theta^*,\mathbf{u}^*)-H(\theta,\mathbf{u}))}\}$
        $\theta_{(i)} = \theta^*$
     else
        $\theta_{(i)} = \theta_{(i-1)}$
end for

If the derivatives are hard, you might try Sympy (https://www.sympy.org/en/index.html) to differentiate U automatically and produce (quite a few lines of!) python code, and there are other possibilities, such as pymc and Jax which will automatically differentiate under the hood. Stan is also a very powerful language for solving such problems.

Issues to consider are how many integration steps per point in the chain, and how big those steps should be. Small steps yield more accurate integration, so H should change little, and almost all points are accepted. But this is expensive, as many likelihood evaluations are needed. Bigger steps are faster, and the Metropolis step sorts out any issues arising from imperfect integration. An acceptance rate of ~ 0.7 is usually good. For further discussion, see Hajian (2006), astroph/0608679.

If we can sample from T we can get the distribution of p by (trivially) marginalising over u. Since $T = \exp[-U(\theta)]\exp[-K(u)]$, marginalising over u gives p (up to an irrelevant normalisation constant).

Why does this work? Principally, because, if we integrate Hamilton's equations, H should be conserved, so the target density is constant in phase space, and all samples should be accepted. Also, if we integrate for long trajectories, we can travel far in parameter space and explore it well. This is called 'good mixing'.

What challenges are there? Integration is not exact, and we want to do it quickly, so it is usually done with numerical integration (e.g. leapfrog) with big steps. This approximate integration means H is not perfectly conserved. We therefore add a Metropolis-Hastings accept/reject step, and this sorts out any inaccuracies. At the end of the orbit integration, a new random momentum is drawn, and a new orbit then leads to a new sample.

9 Convergence tests

It is vital to know that the chain has enough points in it to represent well the target distribution. It will never be perfect, but asymptotically it approaches the right distribution if the detailed balance condition holds. How do we know? A standard technique is the Gelman-Rubin test (1992) .

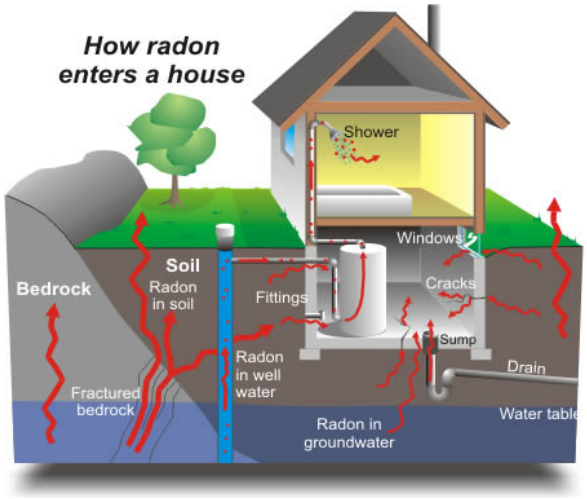

Figure 4: Radon gas (from Pymc website).

10 Bayesian Hierarchical Models

In many practical situations, the likelihood can be difficult to evaluate, since it may be hard to write a direct expression down for the sampling distribution. But we can often make progress by analysing problems as a multilevel system, or Bayesian Hierarchical Model.

A typical example of a BHM is when we have a population of objects, and we use the collection of individual objects to infer something about the population, whose properties may be specific by one or more population parameters θ.

10.0.1 Ordinary Bayes vs Hierarchical Bayes

Ordinary Bayes:

$$p(\theta|d) \propto p(d|\theta)\, p(\theta) \tag{10.19}$$

Hierarchical Bayes: we introduce extra (unobserved) 'latent' variables into the problem

$$p(\theta, \phi|d) \propto p(d|\theta, \phi)\, p(\phi|\theta)\, p(\theta) \tag{10.20}$$

where ϕ are latent variables (or parameters). Often these are marginalised over to obtain

$$p(\theta|d) \propto \int p(\theta, \phi|d)\, d\phi. \tag{10.21}$$

11 Radon data modelling

Radon is a carcinogen and levels of radon in houses in the US have been studied, with a famous dataset collected and analysed in Gelman et al.'s BDA book.

The data are noisy radon measurements, made in different counties in the US, and on different floors (the radon levels will be higher nearer to the ground). The idea is to pool data from many house measurements, to assess the radon risk in a county c, and to extrapolate to living areas if the measurements were taken in the basement.

The data model is as follows:

- We assume that the expected radon level is a linear function of the floor level f,

$$\mu = a_c + b_c f \tag{11.22}$$

  (which is 0 or 1 in the measurements, where 0 is the basement, and 1 the living space).

- We assume the measurement error is a zero mean gaussian, but we don't know the error (variance ϵ^2), i.e.

$$d(f, c) = a_c + b_c f + n \tag{11.23}$$

  where $n \sim \mathcal{N}(0, \epsilon^2)$ and we want to infer ϵ.

- You see that the coefficients a_c and b_c are not fixed, but vary with county. We assume that they are drawn from a normal distribution with a universal mean and variance, i.e. $a_c \sim \mathcal{N}(\mu_a, \sigma_a^2)$, where μ_a and σ_a^2 are unknown. Similarly for b_c.

A good starting point is to draw a diagram that represents what you would need to do to generate the data. This Directed Acyclic Graph (DAG) for this is shown in Fig. 5. It is a Bayesian Hierarchical Model, with variability at several levels.

As usual, we analyse the problem systematically:

- Rule 1: what do we want to know? Quite a few things: risk levels for each county (a_c); extra risk in basements (b_c); variability from house to house (or measurement device error) ϵ; variation across country (σ_a) etc., all conditioned on the data (i.e. posterior probabilities).
- Data: radon measurements $\hat{r}$.
- Model. See the DAG.
- Parameters: $\mu_a, \sigma_a, \mu_b, \sigma_b, a_c, b_c, \epsilon$
- Likelihood (of the final level of the DAG): $\hat{r} \sim \mathcal{N}(r_{\text{true}}, \epsilon^2)$.

Sample from all of the unknowns.

12 Model Comparison

- A higher-level question than parameter inference, in which one wants to know which theoretical framework ('model') is preferred, given the data (regardless of the parameter values)
- The models may be completely different (e.g. compare Big Bang with Steady State, to use an old example),
- or variants of the same idea. E.g. comparing a simple cosmological model where the Universe is assumed to be flat, with a more general model where curvature is allowed to vary (i.e. adding an extra parameter can be considered as a new model).
- The sort of question asked here is essentially 'Do the data favour a more complex model?'
- Clearly in the latter type of comparison the likelihood itself will be of no use - it will always increase if we allow more freedom.

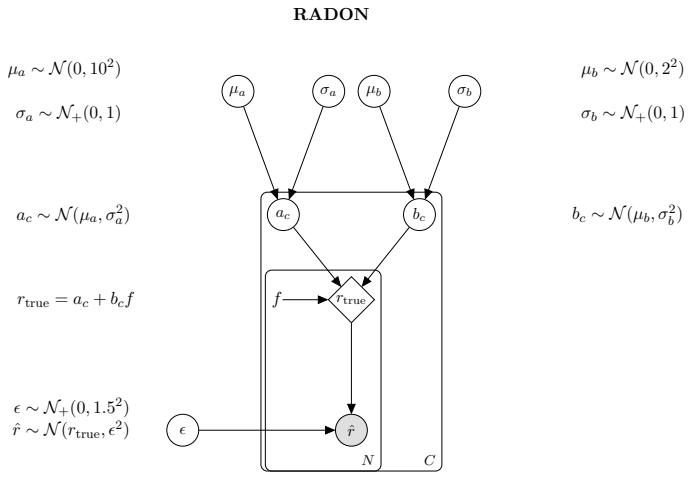

Figure 5: Radon DAG with probability distributions. $\mathcal{N}_+$ indicates a positive half-gaussian distribution.

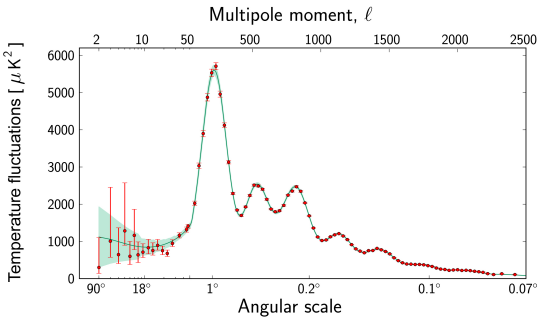

Figure 6: The Planck power spectrum, with the theoretical model with best fitting cosmological parameters. Models other than the Big Bang ΛCDM model may struggle to reproduce the data as well as this, so $p(\mathbf{d}|M)$ would be smaller than $p(\mathbf{d}|M = \Lambda CDM)$.

12.1 Bayesian Evidence, or Marginal Likelihood

- We denote two competing models by M and M'.
- We denote by $\mathbf{d}$ the data vector, and by θ and θ' the parameter vectors (of length n and n').
- Rule 1: Write down what you want to know.
- Here it is $p(M|\mathbf{d})$ - the probability of the model, given the data.
- Use Bayes' theorem:
$$p(M|\mathbf{d}) = \frac{p(\mathbf{d}|M)\pi(M)}{p(\mathbf{d})}$$
- The Bayesian Evidence is
$$p(\mathbf{d}|M) = \int d\theta\, p(\mathbf{d}|\theta, M)\pi(\theta|M),$$
- **If a model has no parameters,** then the integral is simply replaced by $p(\mathbf{d}|M)$, which is just the sampling distribution in this simple case.
- The relative probabilities of two models is
$$\frac{p(M'|\mathbf{d})}{p(M|\mathbf{d})} = \frac{\pi(M')}{\pi(M)} \frac{\int d\theta'\, p(\mathbf{d}|\theta', M')\pi(\theta'|M')}{\int d\theta\, p(\mathbf{d}|\theta, M)\pi(\theta|M)}.$$
- With 'uninformative' (equal) priors on the models, $\pi(M_1) = \pi(M)$, this ratio simplifies to the ratio of evidences, called the **Bayes Factor**,
$$B \equiv \frac{\int d\theta'\, p(\mathbf{d}|\theta', M')\, \pi(\theta'|M')}{\int d\theta\, p(\mathbf{d}|\theta, M)\, \pi(\theta|M)}.$$

Challenges: The evidence requires a multidimensional integration over the likelihood and prior, and this may be very expensive to compute.

- Algorithms: we won't study these in this course, but the most used method is nested sampling (examples are multinest, polychord, dynesty), where one tries to sample the likelihood in an efficient way.

12.2 Gaussian Example

In this gaussian example, we can evaluate the integrals analytically.

Let M_0 be $x \sim \mathcal{N}(0, \sigma^2)$, and M_1 be $x \sim \mathcal{N}(\mu, \sigma^2)$, where the prior on μ is gaussian with variance Σ^2. Let the measurement be $x = \lambda\sigma$.

$$p_0(x|M_0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}$$

and

$$p_1(x|\mu, M_1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

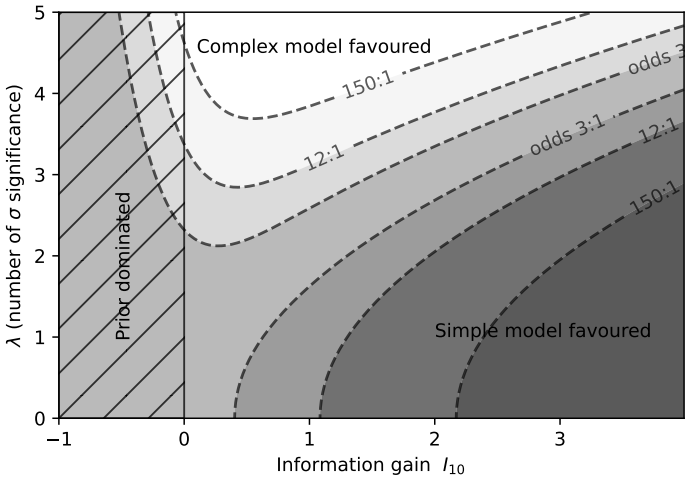

Figure 7: The Bayes Factor for a gaussian likelihood (variance σ^2), and a gaussian prior (variance Σ^2). The x axis $=\log_{10}(\Sigma/\sigma)$; the y axis is datum$/\sigma$. Figure: R. Trotta.

Hence

$$B_{01} = \frac{p_0(x|M_0)}{\int_{-\infty}^{\infty} p_1(x|\mu, M_1)\, p_1(\mu|M_1)\, d\mu}$$

so

$$B_{01} = \sqrt{1 + \frac{\Sigma^2}{\sigma^2}} \exp\left[-\frac{\lambda^2}{2(1 + \frac{\sigma^2}{\Sigma^2})}\right]$$

If $\lambda \gg 1$, then $B_{01} \ll 1$ and M_1 is favoured. If $\lambda \simeq 1$ and $\sigma \ll \Sigma$, then M_0 is favoured (Occam's razor). If likelihood is much broader than prior, $\sigma \gg \Sigma$ then $B_{01} \simeq 1$ and nothing has been learned.

This diagram is very interesting and instructive, and somewhat counter-intuitive. To favour the more complicated model with high probability (say 10 times the probability of the simple model), then the deviation from the simple model parameter value needs to be at least about 3σ. So a 3σ 'result' is really not very significant in a model comparison context, since a probability of 10% is not particularly small.

Summary

- Bayesian formalism can easily be generalised to model comparison
- Resulting integrals over parameter space may be challenging to compute
- Evidence ratios have sensitivity to the prior, even asymptotically. Beware of using the Bayes factor in high dimensions, since the prior volume may be highly uncertain and the Bayes factor can be very sensitive to the limits that are placed on the parameters

13 Likelihood-free inference, or Simulation-based inference

Likelihood-free inference (LFI), or Implicit Likelihood, or Simulation-based inference (SBI) are alternative names for a very different approach to Bayesian parameter inference. It is particularly suitable for cases where the likelihood is either very expensive, or impossible to compute. It requires a way to simulate the data, usually via a computer program, where the model parameters can be adjusted.

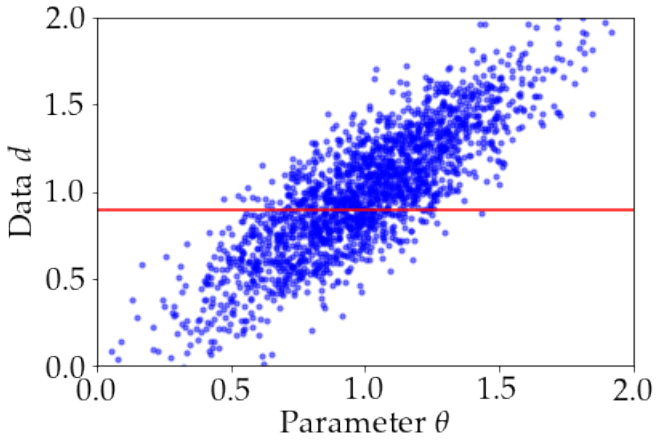

Figure 8: Samples from the joint distribution of parameter θ and data d, $p(\theta, d)$.

The basic idea is to run a very large number of simulations with random parameters (drawn from some prior), and to keep only those that match the experimentally obtained real data. One then inspects the distribution of the parameters that gave rise to the matching data, and this is the posterior.

There are some obvious challenges to this approach. The first is that if the data are continuous, rather than discrete, the probability of obtaining exactly the real data is zero (a set of measure zero), so a certain tolerance may be needed. Secondly, with many data points, the probability of matching all the measured data (even allowing a certain tolerance) is extremely small, especially as the dimensionality of the data increases. For example, running a simulation of the Universe and expecting to reproduce the Milky Way with its neighbour Andromeda, and all the dwarf galaxies of the Local Group, is vanishingly small.

As a result, one demands much less than a perfect match, and typically one requires only that certain summary statistics are reproduced approximately.

Example summary statistics are: correlation function, power spectrum.

13.1 ABC

Let is look at a very simple case, of a model with one parameter θ, and one data point d. We draw θ from some prior $\pi(\theta)$, and run a simulation with that parameter value, generating a data point. We repeat this many times, and sample from the joint distribution $p(\boldsymbol{\theta}, \mathbf{d})$. See Fig. 8.

The simplest way to obtain the posterior is to select those points that lie close to the measured d, within some tolerance ϵ. This is (rejection sampling) ABC (Approximate Bayesian Computation). The distribution of θ values will approach the posterior $p(\theta|d)$ as $\epsilon \to 0$, but at some point one runs out of points, and it gets noisy. To get enough points in the strip, a very large number of simulations need to be run, so it can be expensive. See Fig. 9.

Notice that one can also obtain an estimate of the likelihood (or rather, the sampling distribution), by selecting points at (almost) fixed θ. Sometimes this approach is called implicit likelihood rather than likelihood-free, since the likelihood is in there somewhere. See Fig. 10.

As an alternative to ABC, the distribution of points can be fitted with a continuous function, using machine learning techniques generically called kernel density estimation, or KDE. DELFI is a package

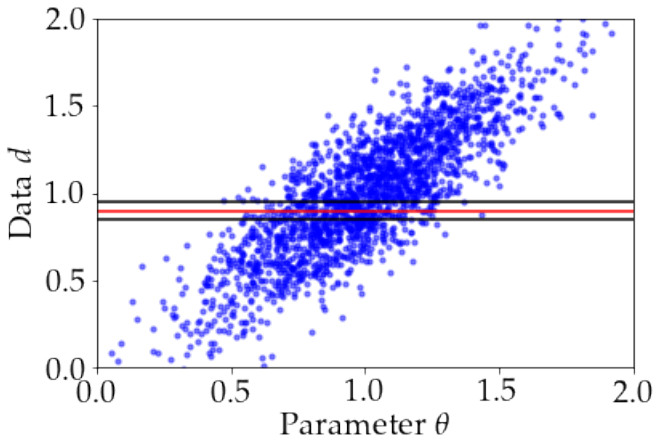

Figure 9: Keeping samples that are close to the measured datum gives an approximation to the posterior, $p(\theta|d)$, which is what we want (Rule 1).

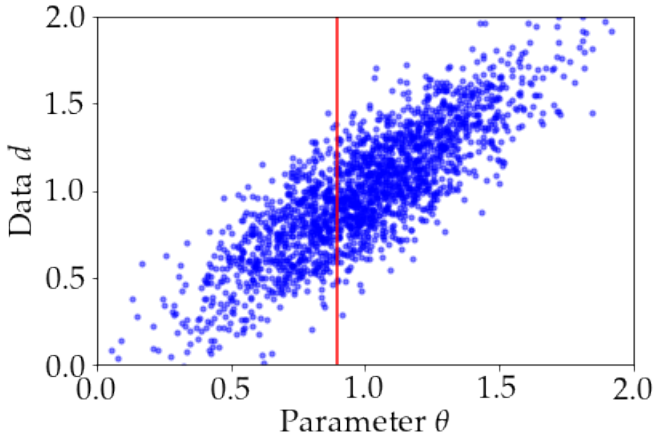

Figure 10: Cutting vertically learns the sampling distribution, $p(d|\theta)$.

that does this. With the distribution fitted, the posterior can be obtained from the approximated probability density evaluated at $\mathbf{d}$, as a function of $\boldsymbol{\theta}$.

In this 2D case, all works well, but as the dimensionality increases, this technique rapidly becomes unfeasible, as too few points will be close to the data. The size of a typical physics experiment dataset will be far too large to handle (If there are N data, and m parameters, the joint distribution is $N + m$-dimensional, which can be **huge**), and even the summary statistics are likely to be too numerous, so we need to compress these radically down to a handful.

We usually need some *massive* data compression.

14 Extreme Data Compression

If we have m parameters, then the maximum compression of the summary statistics, without leading to degenerate solutions, is down to m numbers. Can we do this in a way that preserves information? The simplest is the MOPED algorithm

Assume:

- gaussian data (sampling distribution);
- information is in the mean $\boldsymbol{\mu}(\boldsymbol{\theta})$
- data covariance matrix Σ is independent of parameters
- derivatives of $\boldsymbol{\mu}$ w.r.t. $\boldsymbol{\theta}$ beyond the gradient are not important.

Even if the assumptions are not satisfied precisely, the resulting data compression can still contain almost all the information on the model parameters.

14.1 Derivation of MOPED compression

MOPED was originally derived a different way, for a different purpose (Heavens et al. 2000, MNRAS, 317, 965). This derivation, from Alsing & Wandelt, MNRAS, 2018, 476, 60 is easier.

The log likelihood is

$$\ln p(\mathbf{d}|\boldsymbol{\theta}) = cst. - \frac{1}{2}\left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta})\right]^T \Sigma^{-1} \left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta})\right] \tag{14.24}$$

Expanding about some fiducial point $\boldsymbol{\theta}_*$, this is approximately

$$\ln p(d|\boldsymbol{\theta}) = cst. - \frac{1}{2}\left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*) - \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_\alpha}\tilde{\boldsymbol{\theta}}_\alpha\right]^T \Sigma^{-1} \left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*) - \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_\beta}\tilde{\boldsymbol{\theta}}_\beta\right] \tag{14.25}$$

where $\tilde{\boldsymbol{\theta}}_\alpha \equiv \boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_{*\alpha}$ and we are using the summation convention over α and β. Expanding the brackets (the two cross terms give the same):

$$\begin{aligned} \ln p(\mathbf{d}|\boldsymbol{\theta}) &= cst. - \frac{1}{2}\left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*)\right]^T \Sigma^{-1} \left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*)\right] \\ &+ \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_\alpha}^T \Sigma^{-1} \left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*)\right] \tilde{\boldsymbol{\theta}}_\alpha \\ &- \frac{1}{2}\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_\alpha}\right]^T \Sigma^{-1} \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_\beta}\right] \tilde{\boldsymbol{\theta}}_\alpha \tilde{\boldsymbol{\theta}}_\beta. \end{aligned} \tag{14.26}$$

It's instructive to pause and comsider this expression. The first term after the constant depends only on the data, not on the parameters $\boldsymbol{\theta}$. It is just $\ln p(\mathbf{d}|\boldsymbol{\theta}_*)$ which doesn't vary with $\boldsymbol{\theta}$, so we can ignore it as we are interested in the parameter dependence (i.e. we want the $\boldsymbol{\theta}$ dependence of the likelihood).

The last line does not depend on the data, only on the parameters (as does the prior).

So how do the data change the prior to the posterior? The data coupling to the parameters is only in the second line. Remarkably it comes in only in the combinations

$$\mathbf{y}_\alpha = \mathbf{b}_\alpha^T.(\mathbf{d} - \boldsymbol{\mu}_*) \tag{14.27}$$

where the MOPED vectors are

$$\mathbf{y}_\alpha = \mathbf{b}_\alpha = \Sigma^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_\alpha} \tag{14.28}$$

We don't need all N original data, $\mathbf{d}$, but only the M values $\mathbf{y}_\alpha$! M can be $\ll N$. We have massively compressed the data, and if the assumptions hold, the likelihood is the same - no information has been lost.

Alternatively, we can translate y_α to point estimates of the parameters, with the same assumptions as above. Maximising $\ln p(\mathbf{d}|\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ gives:

$$0 = \frac{\partial \ln p(\mathbf{d}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\gamma} = \mathbf{y}_\gamma - (\mathbf{b}_\alpha^T \Sigma^{-1} \mathbf{b}_\gamma)\tilde{\boldsymbol{\theta}}_\alpha \tag{14.29}$$

where I've used $\partial\tilde{\boldsymbol{\theta}}_\beta/\partial\boldsymbol{\theta}_\gamma = \delta^K_{\beta\gamma}$ (Kronecker delta). Hence point estimates (maximum likelihood) are

$$\tilde{\boldsymbol{\theta}} = (\mathbf{b}^T\Sigma^{-1}\mathbf{b})^{-1}\mathbf{y}. \tag{14.30}$$

i.e.

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_* + D^{-1}\mathbf{y}. \tag{14.31}$$

where the matrix D has elements $D_{\alpha\beta} = (\mathbf{b}_\alpha^T\Sigma^{-1}\mathbf{b}_\beta)$.

You can use either $\mathbf{y}$ or $\hat{\boldsymbol{\theta}}$ as the 'data' in SBI. Both are 'statistics' (a statistic is a function of the data $\mathbf{d}$).

14.2 Alternatives to MOPED

We can use neural networks to find informative summaries (especially when the signal is coming from Σ, not $\boldsymbol{\mu}$). IMNN (information maximizing neural network; Charnock et al. 2018, PRD, 97, 3004). Also Graph NN (Makinen et al. arxiv 2207.05202).

15 Validating the model

Bayesian analysis assumes that the model is the correct one (or, in model comparison, that one of the models is correct). How do we know that the model is a good description of the data. I don't know a Bayesian way to do this, but we can use a frequentist/Bayesian approach: Posterior Predictive Distributions (PPDs).

From our data $\mathbf{x}$ we have the posterior $p(\boldsymbol{\theta}|\mathbf{x})$. We can draw samples from this (in fact we already have some), and generate simulated data $\mathbf{y}$ according to the model, and ask if these simulated

datasets look anything like the original data. For example, we might make a statistic $T(\mathbf{y})$ and plot its frequentist distribution, and see whether $T(\mathbf{x})$ is consistent with being drawn from the distribution.

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \int p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \\ &= \int p(\mathbf{y}|\boldsymbol{\theta}, \not{\mathbf{x}})\, p(\boldsymbol{\theta}|\mathbf{x} d\boldsymbol{\theta} \end{aligned} \tag{15.32}$$

which involves the sampling distribution and the posterior. This works for SBI as well as MCMC. For SBI I would not choose the compressed data for T, but something closer to the raw data (e.g. mean, variance, other standard stats).

16 Selection effects: non-detections and the like

What do we do when our experiment does not always return a result? This is a pretty common situation, when for example the signal is too small (or indeed, too large, if the instrument can't make a measurement there). What should we do? The Bayesian approach is again to assess the problem logically, and a useful approach is to start with a DAG to mimic generating the data.

We'll consider two types of missing data:

- Censoring: the experiment informs us that a measurement was attempted, but no detection was made
- Truncation: if no measurement is made, we don't even know if there is anything there

To give an example of each from astronomy. In the first case we might make a catalogue of bright stars that are visible at optical wavelengths in a patch of sky. Then we see if any of them emit radio waves. We point a radio telescope in the direction of the field of stars, and measure the radio flux from the locations of each star. For some of them, we don't detect anything, and simply report that the flux is below the detection limit. In this case we know how many stars are undetected in radio wavelengths.

For the second case, we don't have the optical image, and only make the radio observations. We simply don't see the stars with no detectable radio emission, and don't know how many there are.

Let's consider this example:

16.1 Measuring the mean from censored data

An experiment measures the mass of (a known number) N identical objects, whose true mass is μ. The measurement error distribution is gaussian, with zero mean and (known) variance σ^2. The measurements are independent of each other. For $M \leq N$ of the objects, the mass is returned by the experiment as detected (included: $I = 1$), but for $N - M$ of the objects, the experiment tells us that it can't measure it - the mass is too small. Its criterion is that it thinks the mass is less than $x_{\text{min}} = 3\mu$ and it is not confident of the measurement, so it is not included ($I = 0$). It does not tell us what it thinks the mass is.

How do we approach this in a Bayesian way? Much the same as before. It's helpful to start with a DAG that describes the generation of data by the model, as shown in Fig. 11. μ is drawn from a

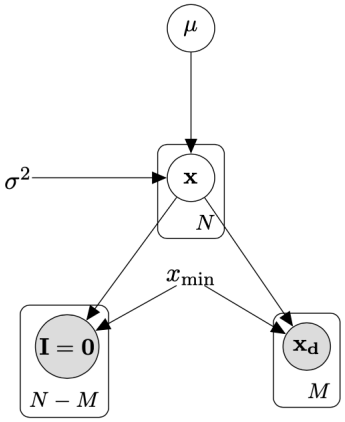

Figure 11: DAG for the censored data model.

prior, and generates N copies of x, each with an error drawn from a gaussian. These x values are either returned as is, as detected objects ($x_d = x$), or a no detection ($I = 0$) is returned if $x < x_{\min}$.

We start with Rule 1: we want the posterior probability of μ, given the M detected data x_d, plus the $N - M$ non-detections. Using Bayes, and a prior $\pi(\mu)$ on μ:

$$p(\mu|x_d, I) \propto p(x_d, I|\mu)\pi(\mu) \tag{16.33}$$

Now, as usual, we introduce the latent variables x, and marginalise over them. Let us do this a little formally, since $x = x_d$ if detected. For notational convenience, let us assume that the experiment returns $x_d = 0$ if not detected (so we can drop I and just use the value of x_d)

$$\begin{aligned} p(\mu|x_d) &\propto \pi(\mu) \int p(x_d, x|\mu)\, dx \\ &\propto \pi(\mu) \prod_{i=1}^{N} \int p(x_{d,i}|x_i, \mu)\, p(x_i|\mu)\, dx_i \end{aligned} \tag{16.34}$$

Now we split the sample into detections and non-detections. For the detections, $p(x_d|x) = \delta^D(x - x_d)$ so the integral is trivial, and for the non-detections, x can be any value below $x_{\min}$, so $p(x_{d,i} = 0|x_i) = 1$ if $x_i < x_{\min}$, so for the undetected objects, the integral is

$$\int_{-\infty}^{x_{\min}} \mathcal{N}(x_i|\mu, \sigma^2)\, dx_i \equiv \Phi(x_{\min}), \tag{16.35}$$

where $\Phi(x) \equiv \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}\sigma}\right)\right]$ and erf is the error function.

Hence the posterior is

$$p(\mu|x_d) \propto \pi(\mu)\, \Phi^{N-M}(x_{\min}) \binom{N}{M} \prod_{i=1}^{M} \mathcal{N}(x_{d,i}|\mu, \sigma^2). \tag{16.36}$$

Notice that we have included a combinatorial factor, to account for the multiple ways that M detections can be drawn from N. For fixed N and M it is a constant and can be absorbed into the proportionality. We have all we need to compute the posterior once we specify a prior for μ (as a location parameter, a uniform prior is appropriate).

16.2 Truncation

Let us now modify the experiment, such that we don't know how many non-detections there are - the experiment returns only the detections. The generative model is the same, except that we don't know N, and we see only the $x_d > 0$ data.

The data model has an extra parameter in it, N, and we are not very interested in it, so it is a nuisance parameter, and we marginalise over it.

The joint posterior for μ and N is

$$p(\mu, N|x_d) \propto p(x_d|\mu, N)\pi(\mu)\pi(N). \tag{16.37}$$

The maths follows as before, but we need to keep the combinatorial term, since it depends on N. After (discrete!) marginalising the posterior over N (which needs to be at least M, obviously), we get

$$p(\mu|x_d) \propto \pi(\mu) \sum_{N=M}^{\infty} \pi(N)\, \Phi^{N-M}(x_{\min}) \binom{N}{M} \prod_{i=1}^{M} \mathcal{N}(x_{d,i}|\mu, \sigma^2). \tag{16.38}$$

A suitable prior for N would be the Jeffreys prior, since N is a scale parameter. $\pi(N) \propto 1/N$.

A Appendix: Bayesian Hierarchical Model example

Here I expand on the example of a simple Bayesian Hierarchical Model, derive an analytic solution and show how it can be solved via Gibbs sampling.

A.1 Straight line fitting with errors in x and y

Let's consider a more complex parameter inference problem, which we can solve analytically, but also via Gibbs sampling. Let's suppose we want to fit a straight line $y = mx$ to some data points with errors in both x and y.

- Data: we have a set of data pairs $(\hat{x}, \hat{y})$ (in fact for simplicity we will have just one pair)
- $\hat{x}$ and $\hat{y}$ are the observed data, related to (unknown) true values x and y
- Model: y is linearly related to x, $y = mx$. Errors are gaussian and independent.
- Parameter: m.
- First, apply Rule 1: write down what you want to know:

  $$p(m|\hat{x}, \hat{y})$$

  (strictly, this is also conditional on knowing the error distribution for $\hat{x}$ and Y, but let us omit it for clarity).

- This is a problem that we can solve analytically, given simple priors, but we will also illustrate how to sample from m using Gibbs sampling.
- Break problem into two steps.

- There are extra unknowns in this problem (so-called latent variables), namely the unobserved true values of $\hat{x}$ and $\hat{y}$, which we will call x and y.
- Note that the model connects the true variables. i.e.,

$$y = mx.$$

  (i.e. NOT $\hat{y} = m\hat{x}$).
- The latent variables x and y are nuisance parameters - we are (probably) not interested in them, so we will end up marginalising over them.

Analysis

- We assume we know the sampling distribution of $\hat{x}$ and $\hat{y}$, i.e. we assume we know

$$p(\hat{x}, \hat{y}|x, y) = p(\hat{x}|x)p(\hat{y}|y)$$

  where the equality holds if the errors are independent.
- Let us now analyse the problem. First we use Bayes' theorem:

$$p(m|\hat{x}, \hat{y}) = \frac{p(\hat{x}, \hat{y}|m)\, p(m)}{p(\hat{x}, \hat{y})} \propto p(\hat{x}, \hat{y}|m)\, p(m)$$

- Now we introduce the latent variables x, y, and write the likelihood above as a marginal integral over x and y:

$$p(m|\hat{x}, \hat{y}) \propto \int p(\hat{x}, \hat{y}, x, y|m)\, p(m)\, dx\, dy$$

- Manipulate using the product rule

$$p(m|\hat{x}, \hat{y}) \propto \int p(\hat{x}, \hat{y}|x, y, m)\, p(x, y|m)\, p(m)\, dx\, dy$$

- The first probability is not dependent on m, i.e.

$$p(\hat{x}, \hat{y}|x, y, m) = p(\hat{x}, \hat{y}|x, y)$$

- Secondly, the product rule gives

$$p(x, y|m) = p(y|x, m)p(x|m)$$

- Next: the model is deterministic:

$$p(y|x, m) = \delta(y - mx)$$

- When multiplied by $p(m)$, $p(x|m)\, p(m) = p(x, m)$, the (joint) prior on x and m. We'll later assume that this is a constant.[4]

[4] One could also reasonably put a prior on the angle, which would lead to a slightly different calculation

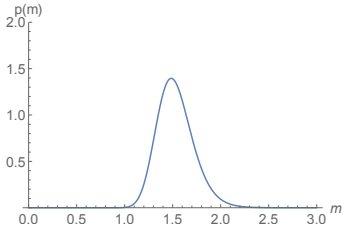

Figure 12: Unnormalised posterior distribution of the slope m, for $\hat{x} = 10$, $\hat{y} = 15$.

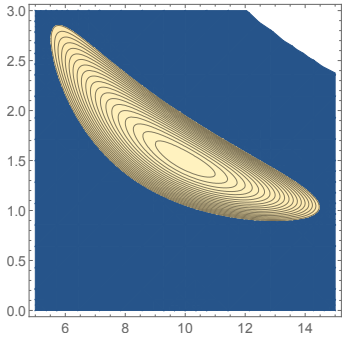

Figure 13: Unnormalised posterior distribution of the latent variable x, and the slope m (on the y-axis).

- Putting these together, we find

$$\begin{aligned} p(m|\hat{x},\hat{y}) &\propto \int p(\hat{x},\hat{y}|x,y)\, p(y|x,m)\, p(x,m)\, dx\, dy \\ &\propto \int p(\hat{x},\hat{y}|x,y)\, \delta(y - mx)\, p(x,m)\, dx\, dy \end{aligned} \tag{A.39}$$

- The integration over y is trivial with the Dirac delta function:

$$p(m|\hat{x},\hat{y}) \propto \int p(\hat{x},\hat{y}|x,mx)\, p(x,m)\, dx.$$

- Assume errors in $\hat{x}$ and $\hat{y}$ are independent Gaussians, and assume the uniform prior for $p(x,m)$. For simplicity, let us also take $\sigma_x^2 = \sigma_y^2 = 1$.

- $$p(m|\hat{x},\hat{y}) \propto \int e^{-\frac{1}{2}(\hat{x}-x)^2} e^{-\frac{1}{2}(\hat{y}-mx)^2}\, dx$$

- Completing the square and integrating (exercise for the student)

$$p(m|\hat{x},\hat{y}) \propto \frac{1}{\sqrt{1+m^2}}\, e^{-\frac{(-m\hat{x}+\hat{y})^2}{2(1+m^2)}}.$$

This is shown in Fig. 12.

A.1.1 Results

We have marginalised analytically over x, but if we want, we can investigate the joint distribution of x and m:

$$p(x,m|\hat{x},\hat{y}) \propto p(\hat{x},\hat{y}|x,mx)\, p(x)\, p(m) \propto e^{-\frac{1}{2}(\hat{x}-x)^2} e^{-\frac{1}{2}(\hat{y}-mx)^2}.$$

This is shown in Fig. 13.

A.1.2 Gibbs Sampling

Let us see how we would set this up as a Gibbs sampling problem.

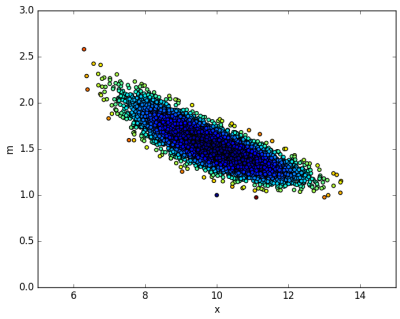

Figure 14: Gibbs sampling of the latent variable x, and the slope m.

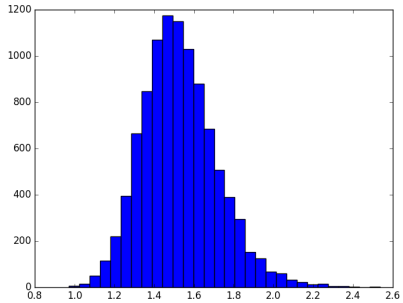

Figure 15: Gibbs sampling of the slope m.

- At fixed x, the conditional distribution on m given x is (note that everything is conditional on the data $\hat{x}, \hat{y}$, but we partly suppress this dependence for clarity - we will be alternately Gibbs sampling x and m so need the conditionals $p(x|m)$ and $p(m|x)$.):
- $$p(m|x,[\hat{x},\hat{y}]) \propto \exp\left[-\frac{(\hat{y}-mx)^2}{2}\right] \propto \exp\left[-\frac{x^2\left(m-\frac{\hat{y}}{x}\right)^2}{2}\right],$$
- i.e.
$$p(m|x,[\hat{x},\hat{y}]) \sim \mathcal{N}\left(\frac{\hat{y}}{x}, \frac{1}{x^2}\right)$$
is a normal $\mathcal{N}(\mu,\sigma^2)$ distribution (in m).
- The conditional distribution of x given m is
$$p(x|m,[\hat{x},\hat{y}]) \propto \exp\left[-\frac{(\hat{x}-x)^2}{2} - \frac{(\hat{y}-mx)^2}{2}\right].$$
- After completing the square, this becomes another normal distribution (in x now):
$$p(x|m,[\hat{x},\hat{y}]) \sim \mathcal{N}\left(\frac{\hat{x}+\hat{y}m}{1+m^2}, \frac{1}{1+m^2}\right)$$

Hence we can sample alternately from m and x, using the conditional distributions, to sample $p(m,x|\hat{x},\hat{y})$, and marginalise over x in the normal MCMC way by simply ignoring the values of x.

Gibbs is only one option for sampling. MCMC with Metropolis-Hastings, or Hamiltonian Monte Carlo, would also be perfectly viable.