Bayesian Statistics

Alan Heavens

May 11, 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.*

1.2 Syllabus

- Introduction
  - Bayesian and frequentist interpretations of probability
  - Inverse problems and the scientific method
  - Bayes theorem
- Parameter inference
  - Likelihoods
  - Priors
  - - Location parameters
  - - Scale parameters
  - 'Noninformative' discussion
  - Conjugate priors
- The Posterior
  - Marginalisation
  - Nuisance parameters
  - Conditional vs marginal errors
  - Numerical Recipes 15.6
  - Profile likelihood (discussion)
- Sampling
  - Representing distributions with samples
  - Marginalising with samples
  - Detailed balance
  - Markov processes
- MCMC
  - Low-dimensional problems: Metropolis-Hastings MCMC
  - Proposal distribution choice
  - Sample correlations
  - - Autocorrelation function
  - - Effective sample size
  - Burn-in
  - Convergence tests: Gelman-Rubin
- Higher-Dimensional problems
  - Hamiltonian Monte Carlo
  - Gibbs Sampling
- Multi-level models
  - Bayesian Hierarchical models
  - Latent parameters

- Model Comparison
  - Savage-Dickey density ratio
  - AIC,BIC,DIC discussion
- Likelihood Free Inference
  - LFI (or SBI/implicit likelihood)
  - Rejection sampling
  - Kernel density estimation (KDE)
- Extreme data compression
  - MOPED algorithm
- Complications
  - Selection effects:
  - truncation
  - censoring

2 Introduction

Bayesian and frequentist views of probability differ in fundamental ways. The frequentist view is usually expressed in terms of relative occurrences of events in multiply-repeated experiments, such as the fraction of heads thrown in the repeated toss of a coin. In Bayesian statistics, probability is sometimes interpreted as a state of knowledge. To my mind, it is better matched to answering scientific questions, since this notion of probability encapsulates what we often want to know when we do science. To see this we first recognise that most scientific questions are inverse problems - what do we learn about the world from the data that we have collected?

2.1 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)

- Logical proposition: a statement that could be true or false
- $p(A|B)$ is the degree to which truth of a logical proposition B implies that A is also true
- The Bayesian view expresses what we often want to know, e.g.
- given the Planck CMB data, what is the probability that the density parameter of cold dark matter is between 0.3 and 0.4?

4.1 Probability rules and Bayes theorem

5 Probability rules (I will assume that you know these at all times)

- $p(x) + p(\sim x) = 1$ (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 Conditional probabilities

6.1 $p(x|y)$ is not the same as $p(y|x)$

Avoid the probability 101 mistake: $p(x|y)$ is not the same as $p(y|x)$.

e.g.

- $x =$ is male; $y =$ has beard
- $p(y|x) \sim 0.1$
- $p(x|y) \sim 1$

Editor's note: Fig. 6.1 is of Charles Darwin. The photographer was Julia Margaret Cameron.

No-one would make this mistake, surely?

7 Example of $p(x|y) \neq p(y|x)$ error: Virus testing

Medical test A virus test gives a positive result (T) in infected patients (V=true) with $p = 0.8$, and has a false positive rate of 0.1. You get a positive result. If 0.01 of the population have the virus, what is the probability that you have it?

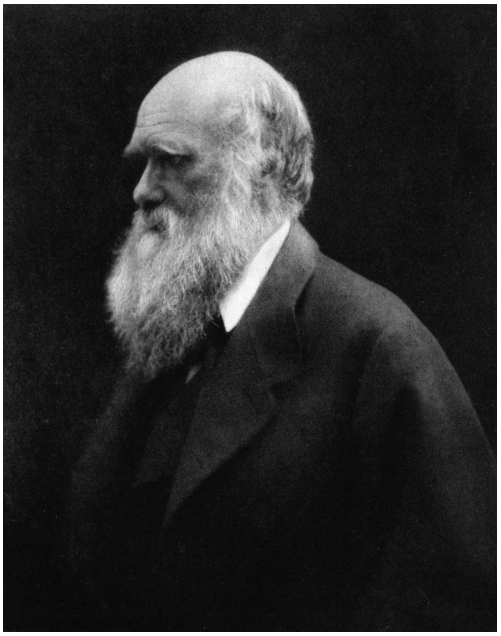

Figure 1: Julia Margaret Cameron

- $$p(V|T) = \frac{p(T|V)p(V)}{p(T|V)p(V) + p(T|\sim V)p(\sim V)}$$
- Put in numbers:
$$p(V|T) = \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.1 \times 0.99} = 0.075$$
- So there is still a 92.5% chance that you do not have the virus.

8 Parameter Inference

8.1 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 $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$

[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.

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

$$p(d|M) = \int p(d|\theta, M)\,\pi(\theta)\,d\theta \tag{8.1}$$

where the integral may be multidimensional (multiple parameters).

8.1.1 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.

8.2 How to set up a problem

8.2.1 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 $\pi(\theta)$?

8.3 Priors

Sometimes there may be previous experimental data that mean that we have some prior knowledge of the model parameters. In fact often the probabilities are specifically written to include this prior information I, so the prior might be written $p(\theta|I, M)$, but I have left this out for clarity. But what do we choose for the prior if we have no prior information?

Summary of priors:

- Prior = (usually) the state of knowledge before the new data are collected.
- For parameter inference, the prior becomes unimportant as more data are added and the likelihood dominates.
- For model comparison, the prior remains important (see later).
- Issues: In the absence of prior experiments, often we want an 'uninformative' prior, but what does this mean?
- Typical choices:

  $\pi(\theta) =$ constant. This is usual for a 'location' parameter, such as the mean of a distribution.

  $\pi(\theta) \propto 1/\theta$ - sometimes called a Jeffreys prior[2] Often applied to 'scale parameters' where we know that it is positive. Each decade is equally likely. Prior is uniform in $\ln\theta$.

For a gaussian distribution, one might reasonably choose a uniform prior for the mean, and a Jeffreys prior for the standard deviation. Note that we sometimes assume a uniform prior over an infinite range, which is an 'improper prior' - it can't be normalised properly to integrate to 1. Provided it yields a proper posterior, it is (usually) fine for parameter inference. (For model comparison, see later - we must have proper priors).

8.3.1 Updating our state of knowledge

If we now obtain some more information, perhaps from a new experiment, then we can use Bayes' theorem to update our state of knowledge of the parameters. The posterior of the last experiment becomes the prior for the next one. This is fine, but it begs the question of what prior did the very first experiment use? This is where the 'uninformative' priors come in.

For this to be a consistent process, then we should show that if we have two experiments, it shouldn't matter if we update the prior after the first experiment, obtaining some data d_1, and then analyse the data d_2 from the second experiment, or take the original prior, and analyse both sets of data together. We show here that they both give the same answer.

Let's do the analysis in two stages, firstly analysing d_1. Let's be explicit about the prior information I (defined to be the state of knowledge before the first experiment is done) in Bayes' theorem applied to the first dataset:

$$p(\theta|d_1, I) = \frac{p(d_1|\theta, I)\, p(\theta|I)}{p(d_1|I)}. \tag{8.2}$$

Now we analyse the second data set. It's similar, but the data are different, $d_1 \to d_2$ of course, and we have some *extra information* from the first experiment, so we should update I to include d_1:

$$I \to d_1, I. \tag{8.3}$$

So, Bayes' theorem applied to the second dataset gives a posterior

$$p(\theta|d_2, d_1, I) = \frac{p(d_2|\theta, d_1, I)\, p(\theta|d_1, I)}{p(d_2|d_1, I)}. \tag{8.4}$$

We now notice that the new prior in this expression is just the old posterior probability from equation (8.2), i.e. we have updated our prior state of knowledge from the original prior, instead using the posterior from the first dataset.

[2] Jeffreys prior is a more general concept, which we will touch on later.

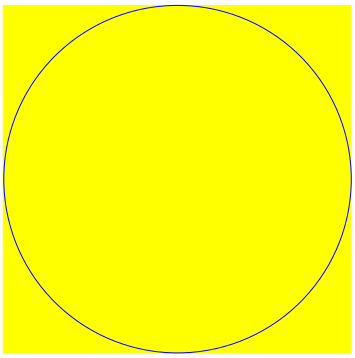

Figure 2: Fraction of area inside the circle is $\pi/4$.

We can also the rules of probability to write the new likelihood as

$$p(d_2|d_1, \theta, I) = \frac{p(d_2, d_1|\theta, I)}{p(d_1|\theta, I)}. \tag{8.5}$$

Substituting in the old posterior probability and this expression for the new likelihood we find

$$p(\theta|d_2, d_1, I) = \frac{p(d_2, d_1|\theta, I)\, p(\theta|I)}{p(d_2, d_1|I)}. \tag{8.6}$$

This has the same form as equation (8.2), the outcome from the initial experiment, but now with the new data incorporated, i.e. the result we would write down if we analysed the data together ($d \to \{d_1, d_2\}$). So, analysing separately and updating the prior after the first dataset gives the same answer as analysing the combined dataset with the original prior.

Bayes' theorem gives us a natural way of improving our statistical inferences as our state of knowledge increases.

8.4 Noninformative priors?

Using previous data to define our state of knowledge is fine, but the very first dataset that was used to determine our state of knowledge will have had to have a prior with no previous data to go on. For such situations, we often try to choose an 'uninformative' prior, which is not as easy as it sounds (and its meaning may not be particularly well-defined).

A uniform prior may seem natural, but it is worth thinking a bit more. Consider this problem: imagine cartesian coordinates in N dimensions, with the prior range being $(-\frac{1}{2}, \frac{1}{2})$ for all coordinates. The prior probability of being inside the N-sphere which just fits inside the prior volume is

$$\frac{\pi^{N/2}}{2^N \Gamma(1 + N/2)}$$

An apparently uninformative prior may be highly informative when viewed a different way.

8.5 Conjugate priors

Sometimes a prior is chosen for mathematical convenience, where, when combined with a given form for the likelihood, the posterior can be calculated analytically and has the same mathematical form

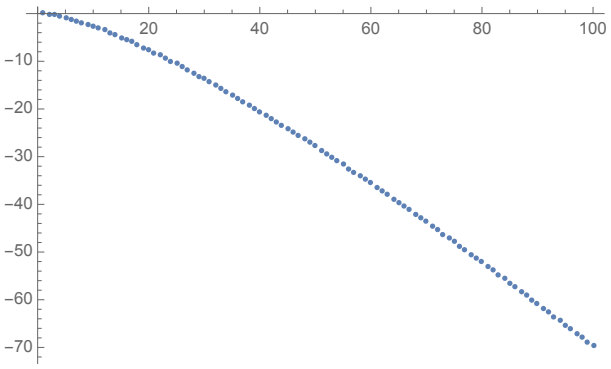

Figure 3: Probability of being within inscribed N-sphere, as a function of the dimension N. What is plotted is $\log_{10} p$ vs N

as the prior. Examples are:

8.5.1 Gaussian likelihood with gaussian prior

If the likelihood is gaussian, and the variance is known, such that the mean μ is the parameter of interest, then a conjugate prior is also a gaussian. The conjugate prior can have any mean and variance, so is flexible.

Note that there is nothing inherently special about conjugate priors; they are just convenient mathematically, but they may be flexible enough to specify sensible location and scale constraints.

8.5.2 Gaussian likelihood with known mean

If on the other hand the mean μ is known, but the variance σ^2 is unknown, an inverse gamma distribution is a conjugate prior for $x = \sigma^2$:

$$f(x; \alpha, \beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x} \tag{8.7}$$

Exercise: show that the posterior is also an inverse gamma distribution, with parameters updated as follows $\alpha \to \alpha + n/2$, and $\beta \to \beta + \sum_{i=1}^{n} (x_i - \mu)^2/2$ for n data.

End of Lecture 3

9 Prior becomes less important with more data

In parameter inference problems, as more data are collected, the likelihood gets progressively more peaked around the true parameter values. For a sufficiently narrow likelihood, the prior is almost constant over the relevant range, and it becomes unimportant (the height of the prior there is irrelevant as it is normalised away by the evidence in the denominator). Let us see this with an example (from Sivia & Skilling).

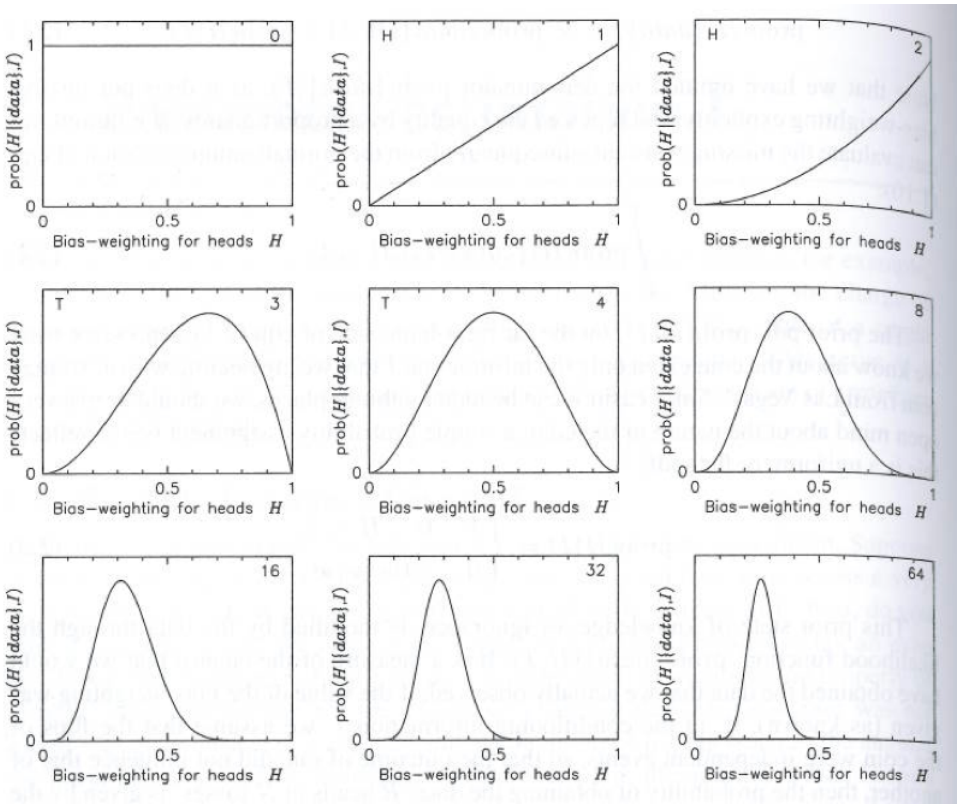

Figure 4: Top left: uniform prior. Subsequent panels - posterior after N coin tosses (labelled in the upper right; result of coin toss is shown in upper left for small N). From Sivia & Skilling.

9.1 Case Study: fair coin?

Model: probability of a head is θ. Uniform prior in θ assumed. Sequence is HHTT...

Second panel: $p(\theta|H) \propto p(H|\theta)\,\pi(\theta) = 2\theta$

3 different priors, shown in first panel. After many data are collected, the posterior becomes insensitive to the prior. In this case, the highly informative prior that supposes the coin is almost fair needs more data to overrule the prior.

10 Posterior

The posterior is the natural outcome of a Bayesian inference problem. It encapsulated our current state of knowledge of the model parameters. It may be very high dimensional, if there are many parameters, and we may want to put it into a more digestible form. It is common to marginalise over all but two parameters, and plot marginal posteriors as a function of each pair of parameters. These are often plotted in 'corner plots'. An example (from Planck cosmological analysis) is shown in Fig. 6.

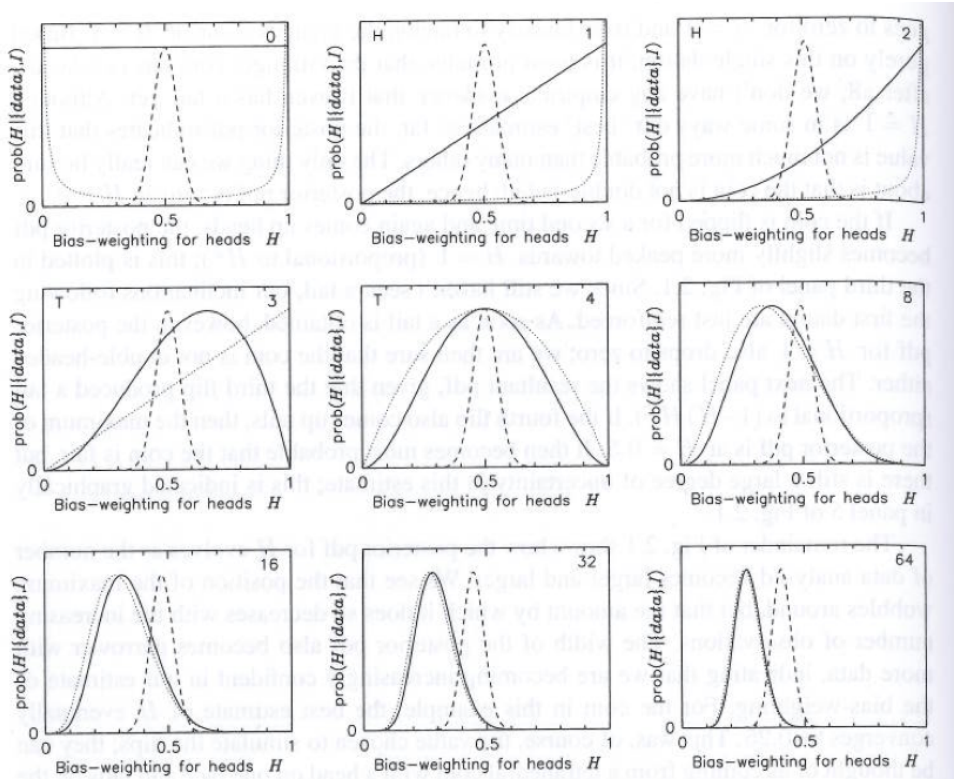

Figure 5: As in Figure 4, but with three different priors (representing uniform, and prior belief of a fair coin, and unfair coin). Note that the posteriors approach each other, but this may require a lot of data, if the prior is very 'informative', as here for the strong prior belief that the coin is fair. From Sivia & Skilling.

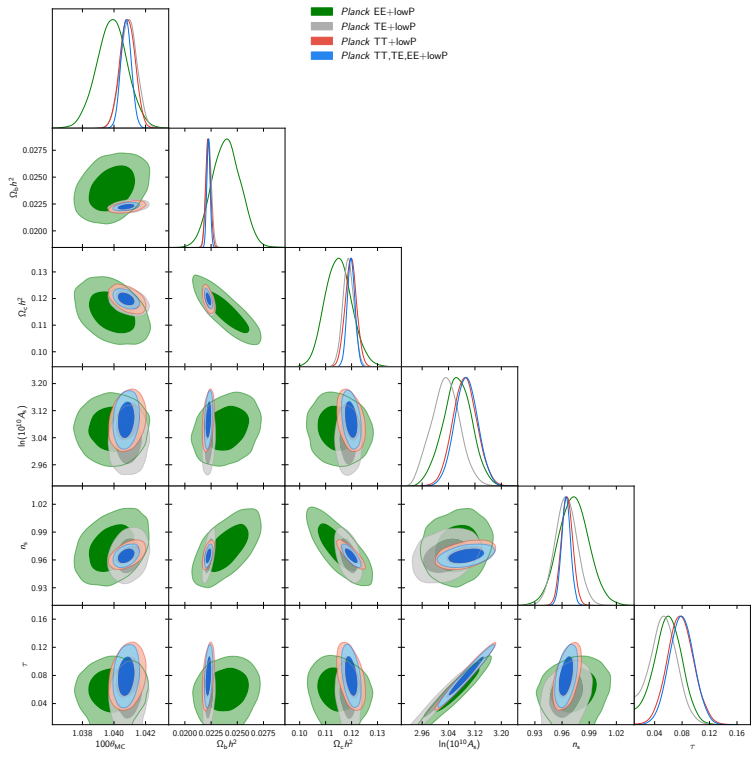

Figure 6: Planck parameters. Model: ΛCDM. Parameters: density of baryons, density of dark matter, etc. Data: measurements of temperature and polarisation of the cosmic background radiation. From Planck Collaboration, arxiv:1502.01589

10.1 Marginalisation

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

$$p(\theta_1, \theta_2|d) = \int p(\theta_1, \dots, \theta_n|d)\, d\theta_3 \dots d\theta_n \tag{10.8}$$

10.2 Conditional errors

If all parameters are kept fixed (typically at the maximum posterior values), and the posterior distribution computed as a function of the remaining parameter, this is a conditional distribution, with an associate conditional error (e.g. standard deviation). It is rarely relevant, since it does not reflect the additional uncertainty that arises from incomplete knowledge of the other parameters.

10.3 Credible regions

From the posterior, it is often convenient to define credible regions, which are defined as being any region that contains some percentage of the posterior. i.e., a X% credible region is any volume Ω in parameter space such that

$$\int_\Omega p(\theta|d)\, d\theta = \frac{X}{100}. \tag{10.9}$$

In 2D these are often shown as contour plots, for $X = 68.3, 99.5, 99.7$ (corresponding to the probabilities enclosed in a gaussian distribution by $\pm 1\sigma, 2\sigma, 3\sigma$ points, but it's arbitrary. Make sure you identify what the contours represent if you use them.

Of course, there is a lot of freedom in how Ω is chosen. A common choice it the highest posterior density (HPD) credible interval where the pdf shown has the same value on the boundary of Ω. For unimodal posteriors, this is a sensible choice, but for multimodal posteriors, it may not be so useful - the HPD region may consist of several islands.

10.3.1 Gaussian posteriors

With multivariate Gaussian posteriors, the contour levels that contain X% of the posterior can be calculated, and depend on the number of parameters. A useful figure and table comes from 'Numerical Recipes', by Press et al. Fig. 7 shows an ellipse that contains 68.3% of the posterior (dashed line), corresponding to χ^2 being larger than the peak value by $\Delta\chi^2 = 2.30$. This is the '1σ' credible region for the joint parameters. For a single parameter, the range containing 68.3% is different, being between A and A', corresponding to the projection of $\Delta\chi^2 = 1$. It is good practice to specify exactly what the contours contain.

Fig. 8) is a handy table for choosing contour levels. Note that for a Gaussian likelihood, $\ln \mathcal{L} = -\chi^2$ + constant, and the relative posterior probability of parameters (assuming uniform priors), with respect to the mode, is $\exp(-\Delta\chi^2/2)$, where $\Delta\chi^2 \equiv \chi^2 - \chi^2_{\rm min}$.

Note that in the non-Gaussian case, it is better to find numerically the HPD regions that contain X% of the posterior, rather than adopting the contour levels for a Gaussian.

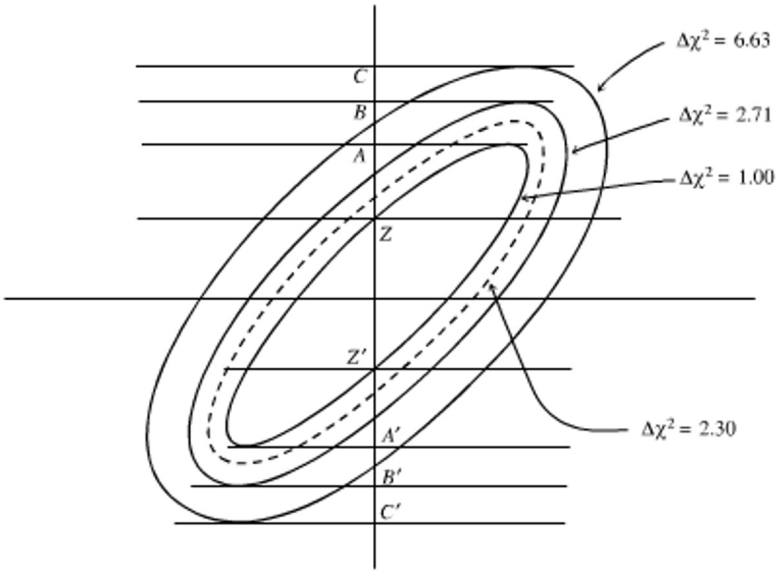

Figure 7: Ellipses showing regions that contain 68.3% of the posterior probability. Regions are within the dashed ellipse, and between the lines A and A'. Other regions are described in Fig. 8. From Press et al, Numerical Recipes, CUP.

$\Delta\chi^2$ as a Function of Confidence Level p and Number of Parameters of Interest ν

	ν					
p	1	2	3	4	5	6
68.27%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.45%	4.00	6.18	8.02	9.72	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.9

Figure 8: Contours that contain X% of the posterior probability (assuming uniform priors) as a function of dimensionality of the parameter space. From Press et al, Numerical Recipes, CUP.

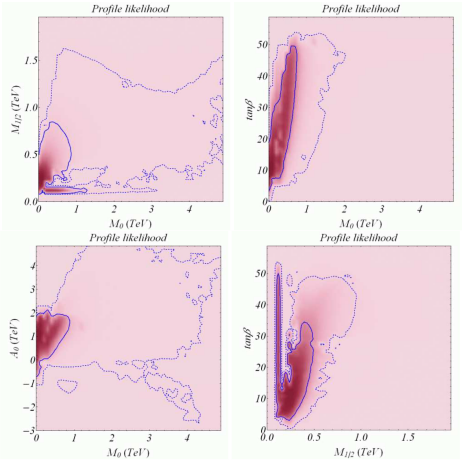

Figure 9: Profile likelihood for LHC data. From Balazs & Carter (2009) arxiv:0906.5012.

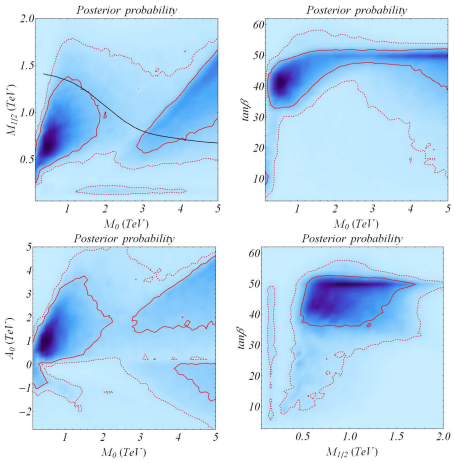

Figure 10: Posterior for LHC data. From Balazs & Carter (2009) arxiv:0906.5012

10.4 Nuisance parameters

Sometimes there are parameters that we are not in the least bit interested in, but affect the measurements, so are part of the data model, and have to be included in the inference problem. They would normally be marginalised out and posteriors of only the interesting model presented. Examples might be the gain of a detector instrument, which may have a prior set by making some lab measurements before the main experiment is done.

10.5 Profile likelihoods

Note that the 'conditional' posterior distribution, where some parameters are fixed (e.g. at most probable values) and the posterior plotted as a function of the others, is not obviously useful, as it does not include variability coming from the imprecisely known parameters that have been fixed. Sometimes this is done though, especially with nuisance parameters, which may be fixed at most probable values (which may vary with , and the resulting pdf is called a profile likelihood (from a Bayesian perspective, in simple cases this is proportional to the posterior for uniform priors). It may be useful to get some feeling for the structure in what may be a very high-dimensional posterior, but it has no simple interpretation and may be misleading.

A useful lesson is that the natural outcome of a Bayesian inference problem is the posterior (which may be high-dimensional). It may be possible to summarise it sensibly with a few numbers (mean, variance, covariance in multidimensional cases), but it may not, and in the end you may need to present the full posterior.

End of Lecture 4

11 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

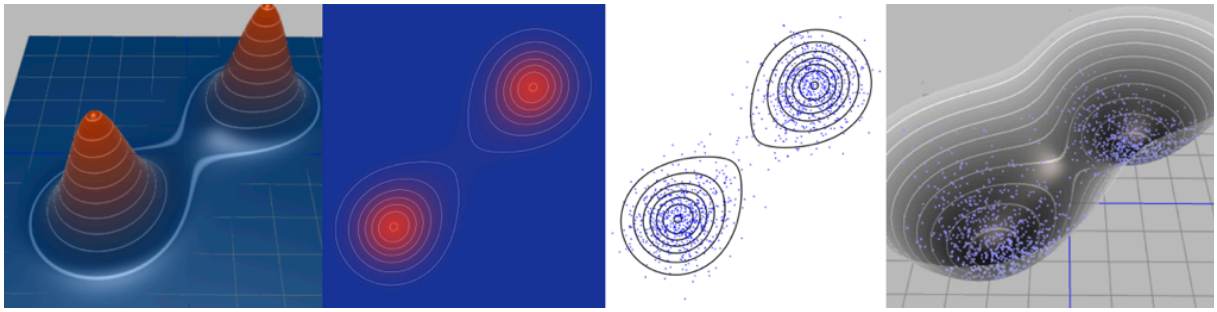

Figure 11: Credit: Alex Rogozhnikov

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:

$$p(\theta) \simeq \frac{\sum_{s=1}^{S} w_s \delta(\theta - \theta_s)}{\sum_{s=1}^{S} w_s}. \tag{11.10}$$

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

$$\hat{\mu} = \langle \theta \rangle = \int p(\theta)\theta \, d\theta \simeq \int \frac{\sum_{s=1}^{S} w_s \delta(\theta - \theta_s)}{\sum_{s=1}^{S} w_s} \theta \, d\theta = \frac{\sum_{s=1}^{S} w_s \theta_s}{\sum_{s=1}^{S} w_s}. \tag{11.11}$$

This is generalised to any function $f(\theta)$,

$$\langle f(\theta) \rangle = \int p(\theta) f(\theta) \, d\theta = \frac{\sum_{s=1}^{S} w_s f(\theta_s)}{\sum_{s=1}^{S} w_s}. \tag{11.12}$$

This is Monte Carlo integration. For example, one might want the covariance of the distribution, whose elements are estimated by

$$\hat{\Sigma}_{ij} = \frac{\sum_{s=1}^{S} w_s (\theta_{i,s} - \hat{\mu}_i)(\theta_{j,s} - \hat{\mu}_j)}{\sum_{s=1}^{S} w_s}. \tag{11.13}$$

12 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.

12.1 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).

12.1.1 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 (see Convergence Tests later). 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)[3]
- 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 idea is convergence, which we will cover later.

12.1.2 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.[4] 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

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

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

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

$$\rho_i P_{ij} = \rho_j P_{ji}. \tag{12.14}$$

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.

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

$$\sum_i \rho_i P_{ij} \tag{12.15}$$

including all the routes to populate j from the 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.

12.2 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.

$$q(\theta'|\theta) \tag{12.16}$$

$=$ probability of a proposed sample at θ' from a previous state θ. Typically this is a function of $\theta' - \theta$, 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

$$\alpha = \min\left[1, \frac{\rho(\theta')q(\theta'|\theta)}{\rho(\theta)q(\theta|\theta')}\right]. \tag{12.17}$$

Let us see if this satisfies detailed balance. Let θ be labelled by i, θ' by j. For concreteness, let us assume that

$$\rho_j q_{ji} \leq \rho_i q_{ij} \tag{12.18}$$

The probability of an accepted transition from i to j is

$$P_{ij} = q_{ij}\min\left[1, \frac{\rho_j q_{ji}}{\rho_i q_{ij}}\right] = \frac{\rho_j q_{ji}}{\rho_i} \tag{12.19}$$

where the first term is the probability that the transition is proposed, and the second is the probability that it is accepted. The reverse probability is

$$P_{ji} = q_{ji}\min\left[1, \frac{\rho_i q_{ij}}{\rho_j q_{ji}}\right] = q_{ji} \tag{12.20}$$

since the proposed sample is accepted with probability 1 in this case. Hence the detailed balance relation is satisfied 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).

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.

12.2.1 Burn-in

For convergence tests (see later) it is often necessary to run two or more chains, with 'dispersed' starting points (i.e. not similar). Each chain may take some time to find the region(s) where the target distribution is high. This exploratory phase is called burn-in and these samples are discarded. There is no golden rule about how many samples to throw away, but one common and useful technique is to find the first sample which is within some factor (say 0.1) of the highest value of the target in the entire chain, and discard all the previous samples.

12.3 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.

12.4 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

$$\hat{C}_\Delta \equiv \frac{1}{S-\Delta} \sum_{s=1}^{S-\Delta} \frac{(\theta_s - \hat{\mu})(\theta_{s+\Delta} - \hat{\mu})}{\hat{\Sigma}} \tag{12.21}$$

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. 12 shows some examples.

12.4.1 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

$$S_{\text{eff}} \equiv \frac{1}{1 + 2\sum_{\Delta=1}^{\Delta_0 - 1} \hat{C}_\Delta}, \tag{12.22}$$

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

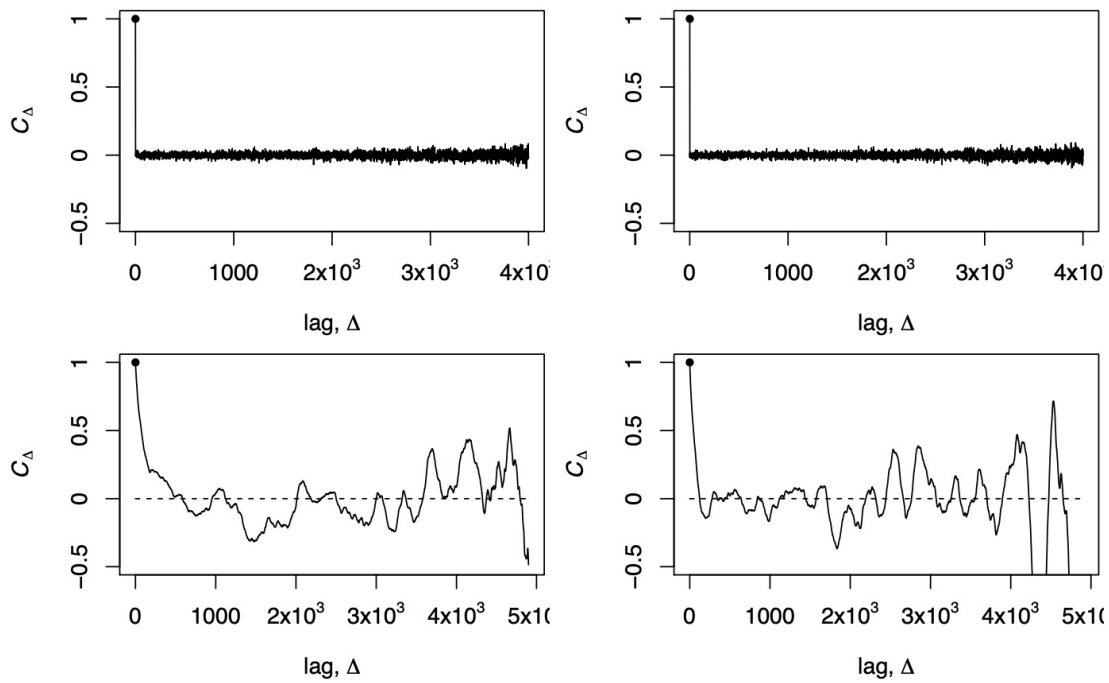

Figure 12: Correlation coefficient of samples for uncorrelated samples (top) and badly-correlated samples (bottom). From D. Mortlock.

12.5 Gibbs sampling

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

Algorithm:

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

Repeat, randomizing (or reversing) the order.

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).

12.6 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 (X, Y) (in fact for simplicity we will have just one pair)

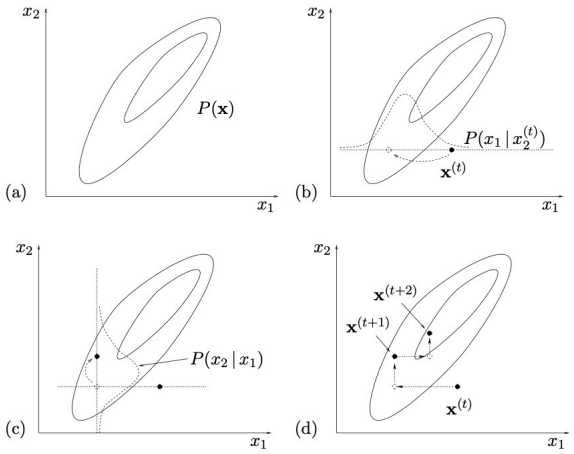

Figure 13: Illustration of Gibbs sampling (from Mackay 2003)

- X and 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:

  $$p(m|X, Y)$$

  (strictly, this is also conditional on knowing the error distribution for 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 X and Y, which we will call x and y.
- Note that the model connects the true variables. i.e.,

  $$y = mx.$$

  (i.e. NOT $Y = mX$).

- 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 X and Y, i.e. we assume we know

  $$p(X, Y|x, y) = p(X|x)p(Y|y)$$

  where the equality holds if the errors are independent.

- Let us now analyse the problem. First we use Bayes' theorem:

$$p(m|X, Y) = \frac{p(X, Y|m)\, p(m)}{p(X, Y)} \propto p(X, Y|m)\, p(m)$$

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

$$p(m|X, Y) \propto \int p(X, Y, x, y|m)\, p(m)\, dx\, dy$$

- Manipulate using the product rule

$$p(m|X, Y) \propto \int p(X, Y|x, y, m)\, p(x, y|m)\, p(m)\, dx\, dy$$

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

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

- Secondly, the product rule gives

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

- Next: the model is deterministic:

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

- Also, we assume the prior on x is independent[5] of m, so

$$p(x|m) = p(x).$$

- Putting these together, we find

$$\begin{aligned} p(m|X, Y) &\propto \int p(X, Y|x, y)\, p(y|x, m)\, p(x)\, p(m)\, dx\, dy \\ &\propto \int p(X, Y|x, y)\, \delta(y - mx)\, p(x)\, p(m)\, dx\, dy \end{aligned} \tag{12.23}$$

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

$$p(m|X, Y) \propto \int p(X, Y|x, mx)\, p(x)\, p(m)\, dx.$$

- Assume errors in X and Y are independent Gaussians, and take uniform priors for x and m. For simplicity, let us take $\sigma_x^2 = \sigma_y^2 = 1$.

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

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

$$p(m|X, Y) \propto \frac{1}{\sqrt{1 + m^2}}\, e^{-\frac{(-mX+Y)^2}{2(1+m^2)}}.$$

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

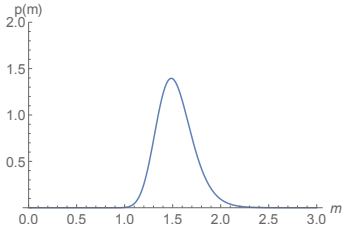

Figure 14: Unnormalised posterior distribution of the slope m, for $X = 10,\ Y = 15$.

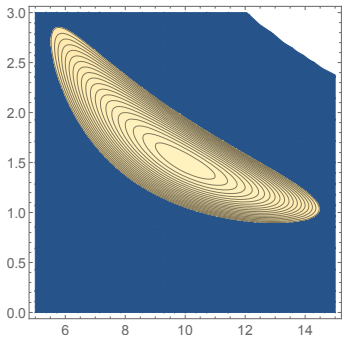

Figure 15: Unnormalised posterior distribution of the latent variable x, and the slope m (on the y-axis).

This is shown in Fig. 14.

12.6.1 Results

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

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

This is shown in Fig. 15.

12.6.2 Gibbs Sampling

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

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

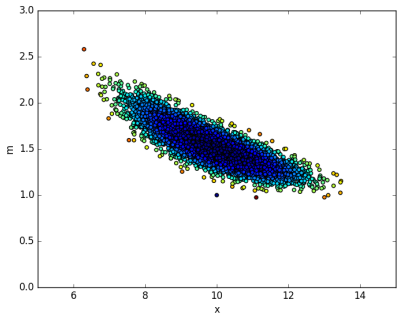

Figure 16: Gibbs sampling of the latent variable x, and the slope m.

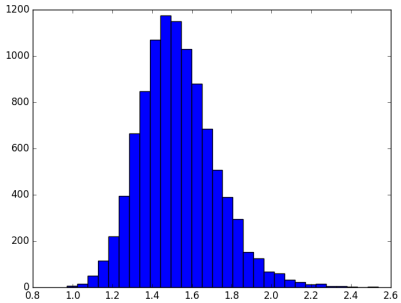

Figure 17: Gibbs sampling of the slope m.

Hence we can sample alternately from m and x, using the conditional distributions, to sample $p(m, x|X, Y)$, 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.

End of Lecture 8

12.7 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

$$U(\theta) = -\ln p(\theta) \tag{12.24}$$

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

There is also a kinetic energy

$$K(u) = \frac{1}{2}\mathbf{u} \cdot \mathbf{u} \tag{12.25}$$

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

$$H(\theta, u) = U(\theta) + K(u) \tag{12.26}$$

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:

$$T(\theta, u) = \exp[-H(\theta, u)]. \tag{12.27}$$

HMC explores this phase space using Hamilton's equations:

$$\begin{aligned} \dot{\theta}_i &= \frac{\partial H}{\partial u_i} = u_i \\ \dot{u}_i &= -\frac{\partial H}{\partial \theta_i} = \frac{\partial \ln p}{\partial \theta_i} \end{aligned} \tag{12.28}$$

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 (see below)).

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: $(\theta^*_{(j-1)}, \mathbf{u}^*_{(j-1)}) \rightarrow (\theta^*_{(j)}, \mathbf{u}^*_{(j)})$
    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 pymc3 and Jax which will automatically differentiate under the hood. Stan is also a very powerful language for solving such problems.

You should use the leapfrog algorithm (which is forward-backward symmetric, as required for detailed balance)

$$\begin{aligned} u_i\left(t+\frac{\epsilon}{2}\right) &= u_i(t) - \frac{\epsilon}{2}\left(\frac{\partial U}{\partial \theta_i}\right)_{\theta_{(t)}} \\ \theta_i(t+\epsilon) &= \theta_i(t) + \epsilon u_i\left(t+\frac{\epsilon}{2}\right) \\ u_i(t+\epsilon) &= u_i\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\left(\frac{\partial U}{\partial \theta_i}\right)_{\theta_{(t+\epsilon)}} . \end{aligned} \tag{12.29}$$

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.

End of Lecture 9

13 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) . Here, two or more chains that begin at 'dispersed' starting points are compared, after their burn-ins are removed. The idea is that if the chains have converged, then their means and variances should agree, except for fluctuations due to there being a finite number of samples. The test is applied separately to each parameter of the model.

The algorithm is

1. Calculate the mean of each chain:
$$\bar{x}_c = \frac{1}{S}\sum_{s=1}^{S} x_{c,s}$$

2. Calculate the variance of each chain:
$$\sigma_c^2 = \frac{1}{S-1}\sum_{s=1}^{S}(x_{c,s} - \bar{x}_c)^2$$

3. Calculate the mean of all the chains (i.e., the best combined estimate for the mean of the distribution):
$$\bar{x} = \frac{1}{C}\sum_{c=1}^{N_c}\frac{1}{S}\sum_{s=1}^{S} x_{c,s} = \frac{1}{C}\sum_{c=1}^{C}\bar{x}_c$$

4. Calculate the average of the individual chains' variances:
$$\sigma^2_{\text{chains}} = \frac{1}{C}\sum_{i=1}^{C}\sigma_c^2$$

5. Estimate the variance of the chains' means:

$$\sigma^2_{\text{means}} = \frac{1}{C-1}\sum_{i=1}^{C}(\bar{x}_c - \bar{x})^2$$

6. The estimated posterior variance is a weighted average of σ^2_{means} and σ^2_{chains}:

$$\hat{V} = \frac{S-1}{S}\sigma^2_{\text{chains}} + \frac{C+1}{C}\sigma^2_{\text{means}}$$

7. We calculate the ratio

$$\hat{R} = \frac{\hat{V}}{\sigma^2_{\text{chains}}} = 1 - \frac{1}{S} + \frac{C+1}{C}\frac{\sigma^2_{\text{means}}}{\sigma^2_{\text{chains}}}.$$

The test statistic $\hat{R}$ can be used to assess convergence. If the chains are well mixed and have all sampled the target distribution then $\sigma^2_{\text{chains}} \simeq \sigma^2_{\text{means}}$ and $\hat{R} \simeq 1$. Whereas if the chains have sampled different parts of the target distribution then their individual variances will be less than the variance between the estimates of the chains and $\hat{R} > 1$. The common heuristic approach is to regard the chains as converged if $\hat{R} \lesssim 1.2$ (we often look for 1.03 or less). The use of means and variances in calculating $\hat{R}$ means it is most appropriate to target densities that are close to be normal and do not have heavy tails. The statistic is also useful in general, even though its distribution under correct sampling is then more difficult to calculate.

14 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. We have in fact already seen one of these - fitting a straight line to data with errors in x and y.

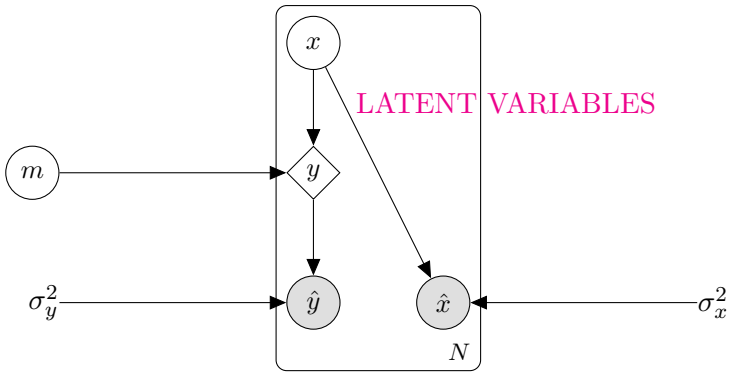

Figure 18: Directed Acyclic Graph (DAG) for straight-line fitting with errors on x and y. Circles indicate values drawn from a probability distribution, grey shaded are the measured data (called X and Y perviously). The diamond indicates that y is a deterministic function of the inputs. The rectangle ('plate') with N inside says there are N repetitions of the data. The variances are fixed, so have no symbol (sometimes a dot is used).

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) is a representation of the data model. Fig. 18 shows

the DAG for the straight line fitting problem. x and y are latent variables that are not measured, but which are part of the data model. The general way to deal with these is to introduce them into the probabilities, and marginalise over them. Very often, HMC techniques are used to sample jointly from the parameters of interest and the latent variables simultaneously.

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.

14.0.1 Ordinary Bayes vs Hierarchical Bayes

Ordinary Bayes:

$$p(\theta|d) \propto p(d|\theta)\, p(\theta) \tag{14.30}$$

Hierarchical Bayes:

$$p(\theta, \phi|d) \propto p(d|\theta, \phi)\, p(\phi|\theta)\, p(\theta) \tag{14.31}$$

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

$$p(\theta|d) \propto \int p(\theta, \phi|d)\, d\phi. \tag{14.32}$$

This is a two-level system, but it can be extended to more levels. To make this concrete, let us look at an example (done in lectures) of a population of pairs x, y where the population is drawn from a gaussian distribution $y_i \sim \mathcal{N}(\bar{y}_i, \sigma^2)$, where the mean grows linearly with x, $\bar{y}_i = mx_i + c$, and the variance of the population around this line is fixed at σ^2. m, c, σ^2 are the parameters of the problem. In this case, for simplicity, we take x_i as fixed and known.

You may like to think about how you would tackle this problem. Details will be given in lectures.

15 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,

$$\mu = a_c + b_c f \tag{15.33}$$

  (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.

$$d(f, c) = a_c + b_c f + n \tag{15.34}$$

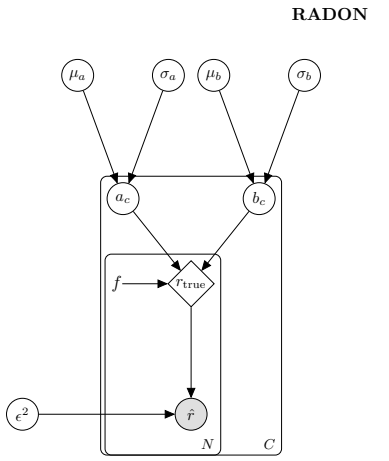

Figure 19: Radon Directed Acyclic Graph. There are C counties, and N houses in each (depends on c)

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.

The Directed Acyclic Graph (DAG) for this is shown in Fig. 19. 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.

16 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),

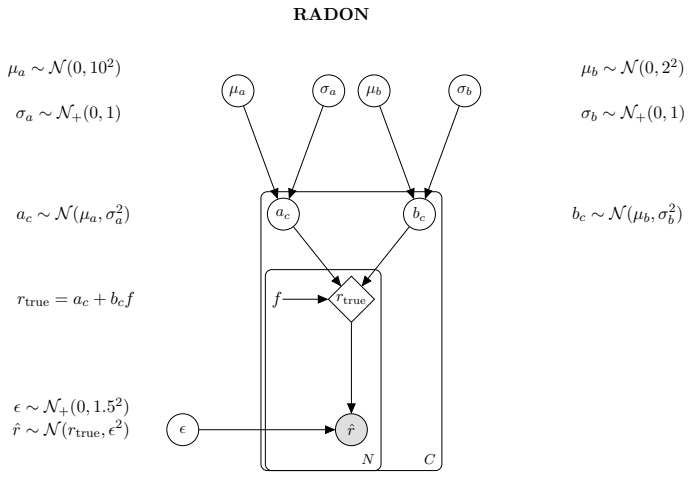

Figure 20: Radon DAG with probability distributions. $\mathcal{N}_+$ indicates a positive half-gaussian distribution.

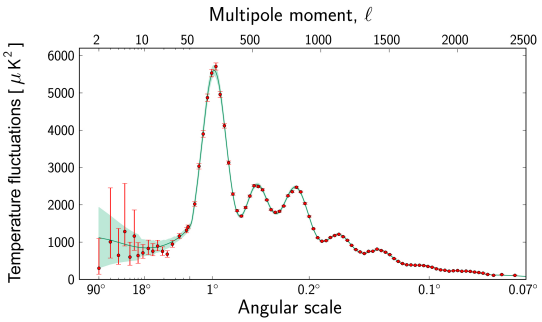

Figure 21: The Planck power spectrum, with the theoretical model with best fitting cosmological parameters. Models other than the Big Bang ΛCDM model may struggle to reproduce the data as well as this, so $p(\mathbf{d}|M)$ would be smaller than $p(\mathbf{d}|M = \Lambda CDM)$.

- 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.

16.1 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**,
$$B \equiv \frac{\int d\theta'\, p(\mathbf{d}|\theta', M')\, \pi(\theta'|M')}{\int d\theta\, p(\mathbf{d}|\theta, M)\, \pi(\theta|M)}.$$

16.1.1 Nested models

- We assume that M' is a simpler model, which has fewer parameters in it ($n' < n$)
- We further assume that it is nested in Model M, i.e. the n' parameters of model M' are common to M, which has $p \equiv n - n'$ extra parameters in it. These parameters are fixed to fiducial values in M'.
- Note that the a complicated model M will (if M_0 is nested) inevitably lead to a higher likelihood (or at least as high), but the evidence may favour the simpler model if the fit is nearly as good, through the smaller prior volume.
- We assume uniform (and hence separable) priors in each parameter, over ranges $\Delta\theta$ (or $\Delta\theta'$). Hence $p(\theta|M) = (\Delta\theta_1 \ldots \Delta\theta_n)^{-1}$
- $$B = \frac{\int d\theta'\, p(\mathbf{d}|\theta', M')}{\int d\theta\, p(\mathbf{d}|\theta, M)} \frac{\Delta\theta_1 \ldots \Delta\theta_n}{\Delta\theta'_1 \ldots \Delta\theta'_{n'}}.$$
- Note that if the prior ranges are not large enough to contain essentially all the likelihood, then the position of the boundaries would influence the Bayes factor. In what follows, we will assume the prior range is large enough to encompass all the likelihood.
- In the nested case, the ratio of prior hypervolumes simplifies to
$$\frac{\Delta\theta_1 \ldots \Delta\theta_n}{\Delta\theta'_1 \ldots \Delta\theta'_{n'}} = \Delta\theta_{n'+1} \ldots \Delta\theta_{n'+p},$$

where $p \equiv n - n'$ is the number of extra parameters in the more complicated model.

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), where one tries to sample the likelihood in an efficient way.
- There are some approximations: e.g., Akaike information criterion (AIC) and Bayesian information criterion (BIC) may be unreliable as they are based on the best-fit χ^2, and from a Bayesian perspective we want to know how much parameter space would give the data with high probability. Also they don't include the prior and are not really Bayesian.

16.2 Bayesian Information Criterion (BIC)

Since the computation of the Bayesian evidence can be very expensive, some short cuts have been derived. There are various 'information criteria' that have been devised as fast substitutes for the Bayesian evidence, but none is completely satisfactory. Perhaps the nearest to a Bayesian approach is the Bayesian Information Criterion (BIC).

- With very constraining data, the likelihood $\mathcal{L}$ will be approximately gaussian near a narrow peak at $\theta_{\rm max}$.
- $$\mathcal{L} \simeq \mathcal{L}_{\rm max} \exp\left[-\frac{(\theta-\theta_{\rm max})_i F_{ij} (\theta-\theta_{\rm max})_j}{2}\right].$$
  The matrix F generally scales in proportion to the number of data points, N.
- The evidence
$$p(\mathbf{d}|M) = \int d\theta\, p(\mathbf{d}|\theta)\pi(\theta)$$
$$p(\mathbf{d}|M) \simeq \mathcal{L}_{\rm max}\pi_{\rm max} \int d\theta \exp\left[-\frac{(\theta-\theta_{\rm max})_i F_{ij} (\theta-\theta_{\rm max})_j}{2}\right]$$
- In k parameter dimensions, this is
$$p(\mathbf{d}|M) \simeq (2\pi)^{k/2} |F|^{-1/2} \mathcal{L}_{\rm max}\pi_{\rm max}$$
- Taking logs:
$$\ln p(\mathbf{d}|M) \simeq \ln \mathcal{L}_{\rm max} + \ln \pi_{\rm max} + \frac{k}{2}\ln 2\pi - \frac{1}{2}\ln|F|$$
- Since $F \propto N$, $|F| \propto N^k$, then up to some constants which are (assumed to be!) unimportant in comparison to $k \ln N$ when N is large,
- $\ln p(\mathbf{d}|M) \simeq \ln \mathcal{L}_{\rm max} - \frac{k}{2}\ln N =$ -BIC/2.
- So maximising the evidence is roughly equivalent to minimising the BIC, except that it is only true asymptotically, and assumes different models have the same $\pi_{\rm max}$

- The BIC is independent of the prior, whereas the full Bayesian evidence depends on the prior, so BIC cannot be accurate in general.
- These assumptions often do not hold in practice. Beware!

16.3 Nested Models: Savage-Dickey Density Ratio

- Let M_0 and M_1 be nested models, such that M_0 is a subset of M_1, e.g., where one of M_1's parameters is fixed to a particular value.
- Some notation. Let the parameters for M_0 be ψ, and those of M_1 be ψ, ϕ.
- M_0 has $\phi = \phi_0$ (fixed).
- Assume that all probabilities are continuous, so

$$\lim_{\phi \to \phi_0} \pi_1(\psi|\phi) = a\pi_0(\psi). \quad i.e.\ \pi_1(\psi|\phi = \phi_0) = a\pi_0(\psi).$$

- The factor a is needed to ensure that the priors are normalised, which we absolutely have to have. e.g. If M_1 has a 2D flat prior with ranges $\Delta\theta_1, \Delta\theta_2$,

$$\pi_1 = \frac{1}{\Delta\theta_1 \Delta\theta_2}; \quad \pi_0 = \frac{1}{\Delta\theta_1}$$

  so $a = \frac{1}{\Delta\theta_2}$.

- The Bayes factor is

$$B_{01} \equiv \frac{p(x|M_0)}{p(x|M_1)} = \frac{\int p_0(x|\psi)\, \pi_0(\psi)\, d\psi}{\int p_1(x|\psi, \phi)\, \pi_1(\psi, \phi)\, d\psi d\phi}$$

- With the continuity, we have then

$$B_{01} = \frac{\int p_1(x|\psi, \phi = \phi_0)\, \pi_1(\psi, \phi = \phi_0)\, d\psi}{\int p_1(x|\psi, \phi)\, \pi_1(\psi, \phi)\, d\psi d\phi} = \frac{a\, p_1(x|\phi = \phi_0)}{p_1(x)}.$$

  where the last step uses $p(A|B, C)p(B|C) = p(A, B|C)$ and $\int p(A, B|C)dB = p(A|C)$ for the numerator. Work out what is going on in the denominator.

- Now, using Bayes' theorem,

$$p_1(x|\phi = \phi_0) = \frac{p_1(\phi = \phi_0|x)\, p_1(x)}{\pi_1(\phi = \phi_0)}$$

Hence

- $$B_{01} = \frac{a\, p_1(\phi = \phi_0|x)}{\pi_1(\phi = \phi_0)}.$$

  This is the Savage-Dickey Density Ratio (SDDR). It looks very simple, but we need to think how to use it, since the numerator is a posterior, not a likelihood.

- However, if we have sampled it, we can estimate the SDDR.

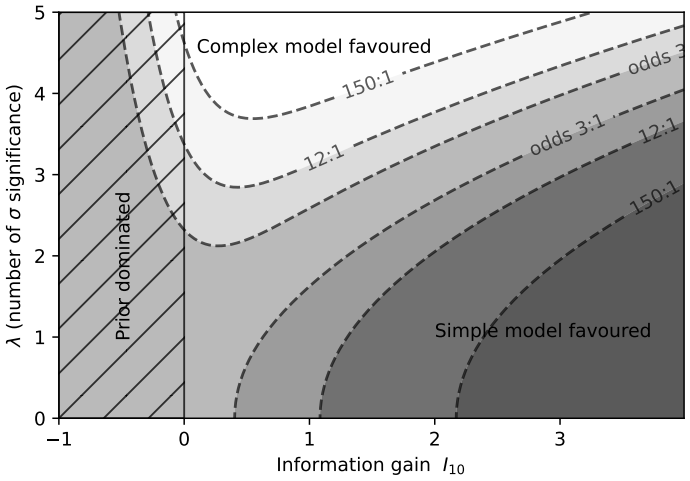

Figure 22: The Bayes Factor for a gaussian likelihood (variance σ^2), and a gaussian prior (variance Σ^2). The x axis $=\log_{10}(\Sigma/\sigma)$; the y axis is datum$/\sigma$. Figure: R. Trotta.

- The denominator is easy if the prior has a simple functional form.
- The numerator may be estimated from samples of the posterior in model 1, e.g. if $f(\Delta\phi, \phi_0)$ is the fraction of samples within (a small range) $\pm\Delta\phi$ of ϕ_0, then $p_1(\phi_0|x) \simeq f(\Delta\phi, \phi_0)/(2\Delta\phi)$.

16.4 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

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

Hence

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

i.e.,

$$B_{01} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}}{\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\Sigma}\int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)} e^{-\mu^2/(2\Sigma^2)} d\mu}$$

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
- Approximations such as BIC may not be accurate
- 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
- SDDR may be useful for nested models

17 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. 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.

17.1 ABC

Let is look at a very simple case, of a model with one parameter θ, and one data point d (called α and $\tilde{\alpha}$ in the plots). 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 $p(\theta, d)$. See Fig. 23.

The simplest way to obtain the posterior is to select those points that lie close to $d = d_m$, the measured data, within some tolerance ϵ. This is ABC (Approximate Bayesian Computation). The

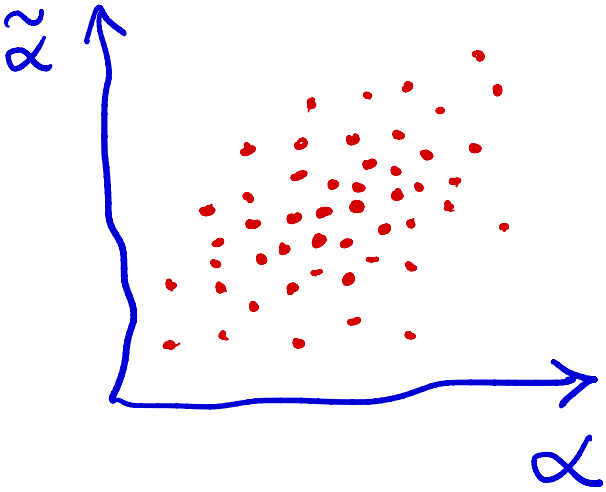

Figure 23: Samples from the joint distribution of parameter (α) and data ($\tilde{\alpha}$), $p(\alpha, \tilde{\alpha})$.

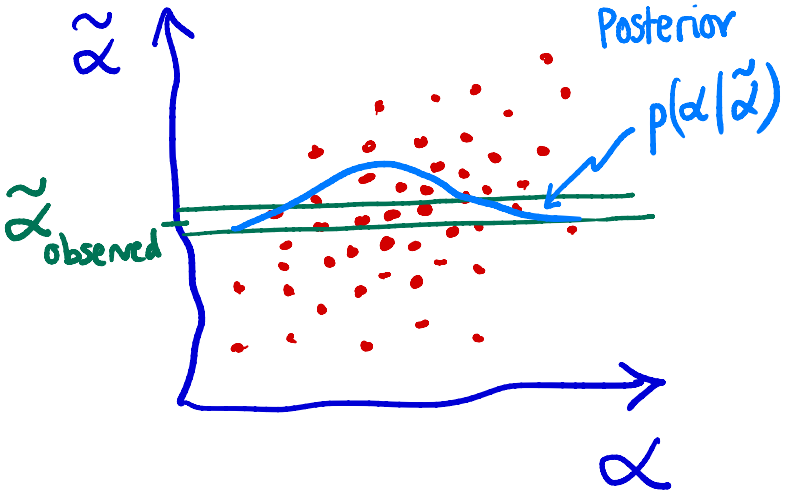

Figure 24: Keeping samples that are close to the measured datum gives an approximation to the posterior, $p(\alpha|\tilde{\alpha})$, which is what we want (Rule 1).

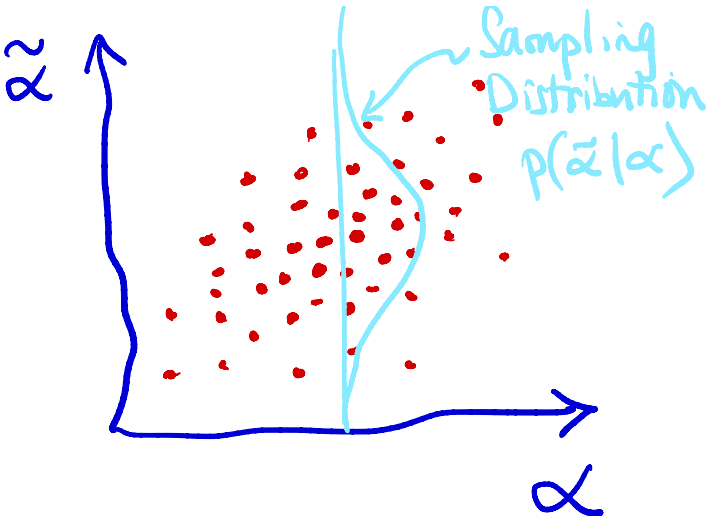

Figure 25: Cutting vertically learns the sampling distribution.

distribution of θ values will approach the posterior $p(\theta|d_m)$ 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. 24.

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. 25.

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 that does this. With the distribution fitted, the posterior can be obtained from the approximated probability density evaluated at $d = d_m$, as a function of θ.

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.

18 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 $\mu(\theta)$
- data covariance matrix Σ is independent of parameters
- derivatives of μ w.r.t. θ 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.

18.1 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

$$\ln p(d|\theta) = cst. - \frac{1}{2}\left[d - \mu(\theta)\right]^T \Sigma^{-1} \left[d - \mu(\theta)\right] \tag{18.35}$$

Taylor expanding μ to linear order about some fiducial point θ_*, this is approximately

$$\ln p(d|\theta) = cst. - \frac{1}{2}\left[d - \mu(\theta_*) - \frac{\partial \mu}{\partial \theta_\alpha}\tilde{\theta}_\alpha\right]^T \Sigma^{-1} \left[d - \mu(\theta_*) - \frac{\partial \mu}{\partial \theta_\beta}\tilde{\theta}_\beta\right] \tag{18.36}$$

where $\tilde{\theta_a} \equiv \theta_\alpha - \theta_{*\alpha}$ and we are using the summation convention over α and β. Expanding the brackets:

$$\begin{aligned} \ln p(d|\theta) &= cst. - \frac{1}{2}\left[d-\mu(\theta_*)\right]^T \Sigma^{-1}\left[d-\mu(\theta_*)\right] \\ &+ \frac{\partial\mu}{\partial\theta_\alpha}^T \Sigma^{-1}\left[d-\mu(\theta_*)\right]\tilde{\theta}_\alpha \\ &- \frac{1}{2}\left[\frac{\partial\mu}{\partial\theta_\alpha}\right]^T \Sigma^{-1}\left[\frac{\partial\mu}{\partial\theta_\beta}\right]\tilde{\theta}_\alpha\tilde{\theta}_\beta. \end{aligned} \tag{18.37}$$

(The two cross terms give the same). The first term (with the constant) is just $\ln p(d|\theta_*)$ which doesn't vary with θ, so we can ignore it as we are interested in the parameter dependence (i.e. we want the θ dependence of the likelihood).

Now we see something interesting, the data comes in only in the combinations

$$y_\alpha \equiv \mathbf{b}_\alpha^T.(\mathbf{d}-\mu_*) \tag{18.38}$$

where the MOPED vectors are

$$\mathbf{b}_\alpha = \Sigma^{-1}\frac{\partial\mu}{\partial\theta_\alpha}. \tag{18.39}$$

We don't need all N original data, $\mathbf{d}$, but only the M values y_α! 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 $\ln p(d|\theta)$ w.r.t. θ gives:

$$0 = \frac{\partial \ln p(d|\theta)}{\partial\theta_\gamma} = y_\gamma - (\mathbf{b}_\alpha^T\Sigma^{-1}\mathbf{b}_\gamma)\tilde{\theta}_\alpha \tag{18.40}$$

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

$$\tilde{\theta} = (\mathbf{b}^T\Sigma^{-1}\mathbf{b})^{-1}\mathbf{y}. \tag{18.41}$$

i.e.

$$\hat{\theta} = \theta_* + D^{-1}\mathbf{y}. \tag{18.42}$$

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 $\hat{\theta}$ as the 'data' in SBI. Both are 'statistics' (= known functions of the data $\mathbf{d}$). These are *highly informative summary statistics* which contain (in ideal cases) as much information as the entire original dataset, but are *extremely* compressed in number.

18.2 Alternatives to MOPED

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

19 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:

19.1 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 $x_{\rm min} = 3\mu$ 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. 26. μ is drawn from a 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 $x < x_{\rm min}$.

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:

$$p(\mu|x_d, I) \propto p(x_d, I|\mu)\pi(\mu) \tag{19.43}$$

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

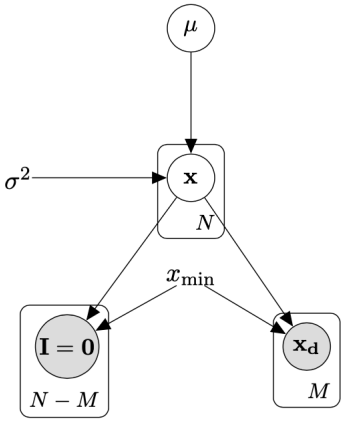

Figure 26: DAG for the censored data model.

returns $x_d = 0$ if not detected (so we can drop I and just use the value of x_d)

$$\begin{aligned} p(\mu|x_d) &\propto \pi(\mu) \int p(x_d, x|\mu)\, dx \\ &\propto \pi(\mu) \prod_{i=1}^{N} \int p(x_{d,i}|x_i, \mu)\, p(x_i|\mu)\, dx_i \end{aligned} \tag{19.44}$$

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

$$\int_{-\infty}^{x_{\min}} \mathcal{N}(x_i|\mu, \sigma^2)\, dx_i \equiv \Phi(x_{\min}), \tag{19.45}$$

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

$$p(\mu|x_d) \propto \pi(\mu)\, \Phi^{N-M}(x_{\min}) \binom{N}{M} \prod_{i=1}^{M} \mathcal{N}(x_{d,i}|\mu, \sigma^2). \tag{19.46}$$

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).

19.2 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

$$p(\mu, N|x_d) \propto p(x_d|\mu, N)\pi(\mu)\pi(N). \tag{19.47}$$

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

$$p(\mu|x_d) \propto \pi(\mu) \sum_{N=M}^{\infty} \pi(N)\, \Phi^{N-M}(x_{\min}) \left(\begin{array}{c} N \\ M \end{array} \right) \prod_{i=1}^{M} \mathcal{N}(x_{d,i}|\mu, \sigma^2). \tag{19.48}$$

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