Doing Bayesian Integrals

The Reverend Thomas Bayes (c.1702 – 1761)
Philosopher, theologian, mathematician

Presbyterian (non-conformist) minister Tunbridge Wells, UK

Elected FRS, perhaps due to a paper defending(!) the works of Isaac Newton. His bibliography contains one other paper, a theological discussion of happiness.

'An Essay Towards solving a problem in the Doctrine of Chances' (1763), put forward to the Royal Society by Richard Price, after Bayes' death.

We were sorting out Bayesian model choice ....

- We looked again at broadening our range of models via hyper parameters
- We then looked at the case of two models, and Bayesian evidence, working out the Bayes Factor as the full Bayesian way of choosing between models
- We considered a simple but illuminating example
- We considered model simplicity and the so-called Ockham factor, which is not a real factor
- We had a look at how to avoid the serious integrals, via the Laplace approximation
- We tried it out on trying to choose between Gaussian and Lorentz line profiles in the face of noisy data – and we found that it worked
- but it showed us that the BF is a statistic! Surprise! And subject to uncertainty…
- We looked at two other criteria, simpler than the BF, namely AIC and BIC, and we expanded our line-profile example to see how these worked.

Model Choice and Bayesian Evidence: review

Suppose we have just two models H_1 and H_2, with parameter sets α and β.

Set of data D. Then Bayes' Theorem for the posterior probs:

$$\mathbf{prob}(H_1, \vec{\alpha} \mid D) = \frac{\mathbf{prob}(D \mid H_1, \vec{\alpha})\, \mathbf{prob}(\vec{\alpha} \mid H_1)\mathbf{prob}(H_1)}{E}$$

$$\mathbf{prob}(H_2, \vec{\beta} \mid D) = \frac{\mathbf{prob}(D \mid H_2, \vec{\beta})\, \mathbf{prob}(\vec{\beta} \mid H_2)\mathbf{prob}(H_2)}{E}.$$

Note we've doubled up on priors!

- priors on the models – our degree of belief that we've got it right
- priors on parameters – here we put in our known contraints or beliefs

E is normalizing factor to make LHS a probability – its importance is coming…..

Model Choice and Bayesian Evidence 2

We can find E :

$$\int \mathbf{prob}(H_1, \vec{\alpha} \mid D)\, d\vec{\alpha} + \int \mathbf{prob}(H_2, \vec{\beta} \mid D)\, d\vec{\beta} = 1.$$

This may be tough in multi-space but it gives us E:

$$\begin{aligned} E &= \int \mathbf{prob}(D \mid H_1, \vec{\alpha})\, \mathbf{prob}(\vec{\alpha} \mid H_1)\, d\vec{\alpha}\, \mathbf{prob}(H_1) \\ &+ \int \mathbf{prob}(D \mid H_2, \vec{\beta})\, \mathbf{prob}(\vec{\beta} \mid H_2)\, d\vec{\beta}\, \mathbf{prob}(H_2). \end{aligned}$$

Putting together this and our Bayes's setup equations gives the posterior probability of model H_1

$$\mathbf{prob}(H_1) = \frac{1}{1 + \mathcal{BP}}$$ 1

in which $\mathcal{B}$ is the *Bayes factor*, the ratio of the integrals of the Likelihood functions multiplied by their priors:

$$\mathcal{B} = \frac{\int \mathbf{prob}(D \mid H_2, \vec{\beta})\mathbf{prob}(\vec{\beta} \mid H_2)\, d\vec{\beta}}{\int \mathbf{prob}(D \mid H_1, \vec{\alpha})\mathbf{prob}(\vec{\alpha} \mid H_1)\, d\vec{\alpha}}$$ 2

Given the posterior probabilities of the competing models we then also have the *posterior odds* P as their ratio:

$$\mathcal{P} = \frac{\mathbf{prob}(H_2)}{\mathbf{prob}(H_1)}.$$

3

Model Choice and Bayesian Evidence 3

- **last three equations encapsulate the Bayesian model choice method**

- **key ingredient – BAYES FACTOR, a ratio of the terms sometimes called EVIDENCE**

- **EVIDENCE terms are the average of the Likelihood Function over the Prior on the parameters**

- **relative magnitude of the EVIDENCE for each model determines its posterior probability**

- **normalizing term E is sum of EVIDENCE terms, each weighted by Prior on relevant model**

Monte Carlo Integration

Very important use of Monte Carlo!

Here's a simplistic way to start:

Suppose we have a probability distribution **f(x)** defined for **a < x < b**

- Draw **N** random numbers **X**, uniformly distributed between **a** and **b**.

- Calculate the function at these points.

- Add these values of the function up, normalize - and

$$\int_a^b f(x)\,dx \simeq \frac{(b-a)}{N}\sum_i f(X_i).$$

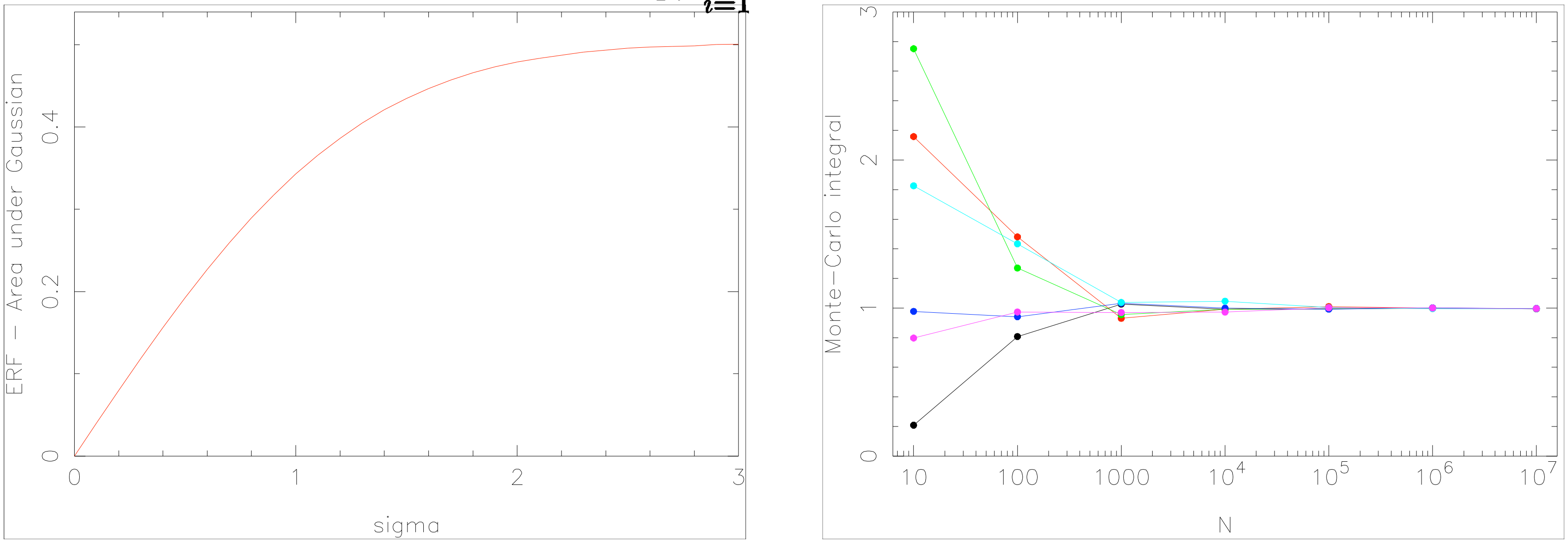
This is **Monte Carlo integration** in its simplest form, **grossly inefficient** because we may not be sampling at points where the function has much value>

But **if the X_i are drawn from the distribution f itself**, then they will sample the regions where **f** is large and the integration will be more accurate. This technique is called **importance sampling.**

Monte Carlo Integration: Example - Gaussian

Use a uniform random-number generator such as the function routine *ran1* of *Numerical Recipes*; make N calls to it, scaling the $(0 \rightarrow 1)$ random numbers to the range of σs required, say $k\sigma$. For each resulting value x_i, compute $f(x_i) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{x_i^2}{2\sigma^2}]$. The integral from 0 to $k\sigma$ is simply

$$\frac{k}{N}\sum_{i=1}^{N} f_i(x). \qquad \textit{(if } \sigma = 1\textit{)}$$



Left - the result, using N=10^6. Right - using +/- 10σ, and varying N. The different curves are the results of different starting indices for the random-number generator. This mindless sum shows how stable MC integration is for well-behaved functions; we have uniformly sampled +/- 10σ, and the function is really a spike between +/- 1σ.

Importance Sampling I

If the X_i are drawn from the distribution **f** itself, then they will sample the regions where **f** is large and the integration will be more accurate.

Suppose **f** is a posterior distribution of some parameter, and we want the expectation value of some function **g** of this parameter. If we can get random X_i drawn from **f** then the MC integral is simply

$$\boxed{\int g(x) f(x)\, dx \simeq \frac{1}{N} \sum_i g(X_i)}$$

- because if the X_i are drawn from **f**, so $f(X_i)$ is uniformly distributed 0.0 -> 1.0. Works for multivariate case. But how to get X_i?

Importance Sampling 2

Take a random number **Y_i** from **h**,
a distribution ~like **f**. Then

$$\int g(x) f(x)\, dx \simeq \frac{1}{N} \sum_i \frac{f(Y_i)}{h(Y_i)}\, g(Y_i).$$

We need this **h** function to 'cover' **f** so that the denom does not explode.
Thus if **f** is a posterior from a Bayes solution we can estimate the Evidence integral:

$$\int f(x)\, dx \simeq \frac{1}{N} \sum_i \frac{f(Y_i)}{h(Y_i)}.$$

Useful if we can find an OK **h** – often the famous **multivariate Gaussian or t dist**.

But frequently we are in many dimensions. Because of how volume multiplies with dimensions a large fraction of the random numbers are wasted, i.e. **$f(Y_i)$** is very small in the above numerator. **We need a better way.**

The Metropolis - Hastings Algorithm

We want to generate random numbers from f/C, C unknown, f multivariate

Metropolis-Hastings algorithm invented to compute equation of state of interacting particles in a box; the algorithm produces thermal equilibrium.

If f is the un-normalized distribution of interest ('**target**') and h a suitable transition probability distribution (the '**proposal**') then

1. **Draw a random number X_i from h**
2. **Draw a random number U_i, uniformly distributed 0.0 to 1.0**
3. **Compute α, the minimum of 1.0 and $f(X_i)/f(X_{i-1})$**
4. **if $U_i < \alpha$ then accept X_i**
5. **Otherwise set $X_i = X_{i-1}$**

The random numbers delivered will (eventually) be randoms drawn from f/C. These randoms are generated sequentially and dependently.

The string is a **Markov Chain => Markov Chain Monte Carlo, MCMC**

The Proposal Function

h engineers the jump from position x_{i-1} to position x_i.

In original algorithm h must be symmetric:

- either in the sense that the prob of a reverse jump is the same
- or if f depends only on absolute value of difference $(x_i - x_{i-1})$

A Gaussian proposal would be of this type.

Acceptance rates of 0.25 to 0.5 give good balance:

- proposal too narrow => too much correlation; chain must be thinned
  => structure of target may not be explored
- proposal too wide => excessive rejection rate, much comp time

A good proposal function is the key

Markov Chain Properties I

Burn-in: serial correlation, so starting point matters, but becomes lost during the burn-in period.

Has burn-in been achieved? Has target region been adequately sampled?

For former: consider ℓ chains each n long (say the last n numbers from a much longer chain which may also have been thinned).

1. Look at all within-chain std devs; these should not be evolving with n

2. The ratio of the std dev of the ℓ means to the individual std devs should be $1/\sqrt{n}$. (This needs proving, as we're not dealing with indep samples.)

Markov Chain Properties

Failing the basic tests?

1. Lengthen burn-in period
2. Examine proposal for width, rejection rate, extent of correlation
3. Vary the thinning
4. Look at power spectrum of number => severity of correlations, thinning requirements

Simple Example

We'll integrate $1/(1 + x^4)$

Take proposal as simple Gaussian centred at current value in chain. Make chains 10000 long; discard first 25% of each, thin by taking every 15th. Two cases:

1. Narrow proposal (σ=1).

- 60% success in making transition

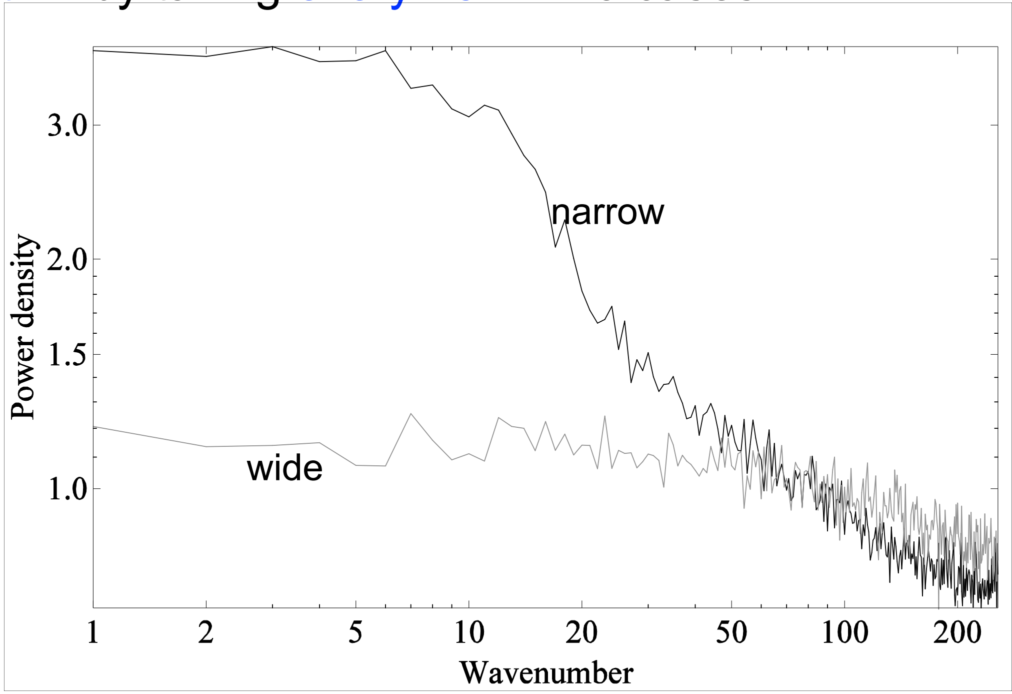
- many repeats: std dev scatter 18%

2. Wide proposal (σ=10)

- 90% failure to make transition

– but maybe less correlated?

- many repeats std dev scatter 4%



Power spectrum shows why – as expected the chain from the narrow proposal shows much correlation, even after thinning, and this outweighs having more distinct numbers in the chain. Clear inefficiency: we would expect 1% scatter in std dev from 10000 samples – but much better than simple MC integration!

The Multi-Dimensional Problem I

Straightforward in one dimension? But we usually want random numbers from a multivariate **$f(\alpha, \beta, \gamma, \ldots.)$;** much more likely in the Bayesian context.

Same arguments for the M-H algorithms, but suitable proposal distributions? Hard.

So – **the Gibbs sampler**

1. Guess at starting vector **$(\alpha_0, \beta_0, \gamma_0, \ldots.)$**
2. Draw **α_1** from **$f(\alpha_0, \beta_0, \gamma_0, \ldots.)$**
3. Draw **β_1** from **$f(\alpha_1, \beta_0, \gamma_0, \ldots.)$, γ_1** from **$f(\alpha_1, \beta_1, \gamma_0, \ldots.)$,** etc.
4. => first multivariate sample

The Multi-Dimensional Problem 2

We may use one iteration of M-H to make each draw from **f**, or find some other simpler way to sample from the distributions.

Check burn-in! May be slowed considerably via correlation between variables.

It can be useful to change to variable combinations which are less correlated, ie approximations to Principal Components.

This combo equips us to do the multi-dimensional integrals often needed in Bayesian problems, e.g. marginalizations, and deriving stats, eg means, percentiles.

Computation of Evidence by MCMC

All this and the problem is not solved: f is a prob dist, but we only know f/C.
We can get samples efficiently from f, but we can't get rid of C this way.
The evidence E is a number, the integral over parameters of the product of the likelihood and its priors:

$$E = \int \mathcal{L}(\vec{\theta})\, p(\vec{\theta})\, d\vec{\theta}.$$

By analogy with thermodynamics, partition functions, dependence of the mean energy of a system on its temperature, we introduce a parameter λ which is going to play a similar role to the inverse temperature in thermodynamics:

$$E(\lambda) = \int \mathcal{L}(\vec{\theta})^{\lambda}\, p(\vec{\theta})\, d\vec{\theta}$$

so that $E(0) = 1$, because our prior at least is normalized to unity. By analogy with physics, we calculate the rate of change with λ:

$$\frac{\partial \ln E(\lambda)}{\partial \lambda} = \frac{1}{E(\lambda)} \int \ln \mathcal{L}(\vec{\theta})\, \mathcal{L}(\vec{\theta})^{\lambda}\, p(\vec{\theta})\, d\vec{\theta}.$$

The right-hand-side is just the expectation of the log-likelihood, with respect to the probability distribution that is proportional to $\mathcal{L}(\vec{\theta})^{\lambda}\, p(\vec{\theta})$. We can sample from this with a M-H algorithm + Gibbs sampler, to get an expectation $< \ln \mathcal{L} >_{\lambda}$. Our solution for $E = E(\lambda = 1)$ is then

$$\ln E = \int_0^1 < \ln \mathcal{L} >_{\lambda}\, d\lambda.$$

In practice, we would generate chains of random numbers for a set of discrete values of λ, compute the respective values of $< \ln \mathcal{L} >_{\lambda}$, and numerically integrate a function fitted to these values.

Example - M-H + Gibbs + thermo-integration

To illustrate both the Gibbs sampler and thermodynamic integration, we generate numbers following a bivariate Gaussian

$$\mathbf{g}(x,y) \propto \exp -\frac{\gamma}{2}\left(x^2 - \frac{9}{5}xy + y^2\right).$$

The correlation coefficient is 9/10. The inverse temperature γ, normally 1, will be used for the thermodynamic integration. The second step is to integrate to get the normalizing factor. Here, as noted, we have to use a proper prior. We used the elliptical prior

$$\mathbf{p}(x,y) = 0 \ \ \textbf{if} \ \ 20 - \left(x^2 - \frac{9}{5}xy + y^2\right) > 0, \quad \mathbf{p}(x,y) = 1/\mathcal{N} \ \ \textbf{otherwise.}$$

Here $\mathcal{N}$ is defined so that

$$\int\!\!\int dx\, dy\, \mathbf{p}(x,y) = 1.$$

The prior is non-zero over the whole region where ***g*** has any signal; we need it.

Example - M-H + Gibbs + thermo-integration

To illustrate both the Gibbs sampler and thermodynamic integration, we generate numbers following a bivariate Gaussian

$$\mathbf{g}(x,y) \propto \exp -\frac{\gamma}{2}\left(x^2 - \frac{9}{5}xy + y^2\right).$$

The correlation coefficient is 9/10. The inverse temperature γ, normally 1, will be used for the thermodynamic integration. The second step is to integrate to get the normalizing factor. Here, as noted, we have to use a proper prior. We used the elliptical prior

$$\mathbf{p}(x,y) = 0 \;\; \textbf{if} \;\; 20 - \left(x^2 - \frac{9}{5}xy + y^2\right) > 0, \quad \mathbf{p}(x,y) = 1/\mathcal{N} \;\; \textbf{otherwise.}$$

Here $\mathcal{N}$ is defined so that

$$\int\int dx\, dy\, \mathbf{p}(x,y) = 1.$$

The prior is non-zero over the whole region where ***g*** has any signal; we need it.

M-H + Gibbs + thermo-integration: Example, con't

The integral of ***gp*** is to be calculated with our random numbers.

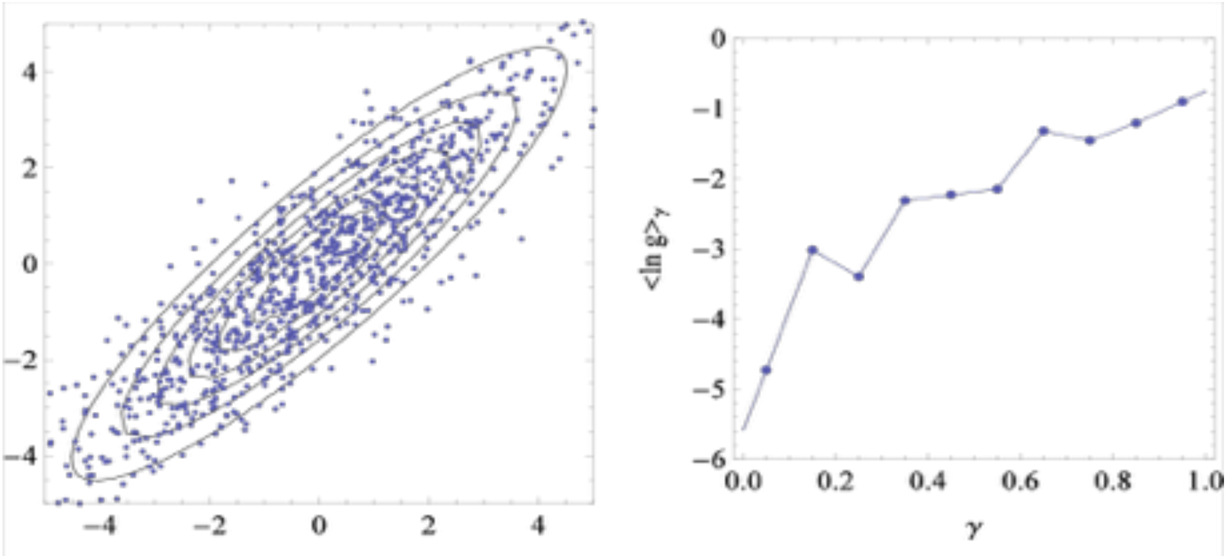
(We could do this case with plain numerical integration because there are only two variables.)

The ***(x,y)*** pairs are generated with the Gibbs sampler, with a Gaussian of standard deviation 5 as the univariate proposal distribution.

What do we get?

M-H + Gibbs + thermo-integration: Example, con't

What do we get?



Success rate ~ 40%.

Chain of 100,000, thinned to 1 in 100, gives correl coeff within 0.5% of 0.90; LH fig shows theoretical contours and some of the chain samples. Power spectrum ~ white. Variance in ***y*** estimated by chain to within 5%.

Thermo integration (RH fig) uses 10 values of ***γ*** 0.0 – 1.0. For each value, a chain is generated and the average of ***ln g*** is calculated, ***g*** in the role of the likelihood function described in the derivation. Well-behaved curve which when integrated 0.0 – 1.0 gives values like 0.099, close to true value of 0.10.