(Neva) 3 The Metropolis algorithm

As said, we want to make the acceptance ratio as large as possible to sample as many different states as possible. The Metropolis algorithm does this.* Consider again

$$\frac{P(\mu\to\nu)}{P(\nu\to\mu)} = \frac{P_\nu}{P_\mu} = \frac{g(\mu\to\nu)\,A(\mu\to\nu)}{g(\nu\to\mu)\,A(\nu\to\mu)}$$

If you choose $g(\mu\to\nu) = g(\nu\to\mu)$, then

$$\frac{P_\nu}{P_\mu} = \frac{A(\mu\to\nu)}{A(\nu\to\mu)}$$

or equivalently

$$P_\nu\, A(\nu\to\mu) = P_\mu\, A(\mu\to\nu)$$

The choice made by Metropolis was

$$A(\mu\to\nu) = \begin{cases} 1 \text{ ; } & P_\nu > P_\mu \\ P_\nu/P_\mu \text{ ; otherwise} \end{cases}$$

So let us for an example assume that P_ν is larger than P_μ then: $A(\mu\to\nu) = 1$; $A(\nu\to\mu) = P_\mu/P_\nu$ and

$$P_\nu \cdot \frac{P_\mu}{P_\nu} = P_\mu \cdot 1 \Rightarrow \quad P_\mu = P_\mu$$

Which means that Metropolis choice satisfied our requirement ✓

* The point of metropolis's algorithm is that we do not need to have a <u>functional</u> form for A that holds for all μ and ν's.
E.g.

$$A(\mu \to \nu) = A_0 \, e^{-\frac{\beta}{2}(E_\nu - E_\mu)}$$

would be such a functional choice. This would lead to very low acceptance ratios in cases where E_ν is even slightly large or not much smaller than E_μ (depending on the algorithm of g), because A needs to be $\in [0 \ldots 1]$, so it must be normalized with the largest decrease in E possible by some choice of g: $A_0 = \exp\left[\frac{\beta}{2} \Delta E_{max}\right]$; $\Delta E_{max} < 0$
Anyway. What Metropolis observed was that our requirement

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{P_\nu}{P_\mu} = \frac{A(\mu \to \nu)}{A(\nu \to \mu)}$$

fixes A only for a given pair (μ, ν) and hence, there can be as many functions A as ~~there~~ there are pairs (μ, ν)!

For a thermal system, the Metropolis algorithm is hence

$$A(\mu \to \nu) = \begin{cases} 1 & ; \; E_\nu < E_\mu \\ e^{-\beta(E_\nu - E_\mu)} & ; \text{ otherwise} \end{cases}$$

In cosmology, we want to sample the posterior $\text{prob}(\vec{\lambda} \mid \text{data})$. Or equivalently the likelihood $\text{prob}(\text{data} \mid \vec{\lambda})$ times the prior $\text{prob}(\vec{\lambda} \mid \text{prior knowledge})$. A cosmological MCMC therefore uses

$$A(\mu \to \nu) = \begin{cases} 1 & ; \; \text{Likelihood of } \nu > \text{Likelihood of } \mu \\ \dfrac{\text{prob}(\vec{\lambda}_\nu)}{\text{prob}(\vec{\lambda}_\mu)} & ; \text{ otherwise} \end{cases}$$

In practice, a simple Metropolis search is performed by choosing a Gaussian proposal $g(\mu \to \nu)$ like

$\Omega_m h^2 (t+1)$ is drawn from

$$e^{-\left[\Omega_m h^2(t) - \Omega_m h^2(t+1)\right]^2 / \sigma_{\Omega_m h^2}}$$

etc (there are some refinements like choosing parameter combinations to fight degeneracies, but the idea remains the same) and accept according to $A(\mu \to \nu)$ above, where the probabilities are computed by the likelihood routines available for the different experiments.

(Sivia) 4 MODEL SELECTION

Suppose there is a debate which model describes the data best. Naively, we could look at how well each model fits the data. However, a model with many parameters will always be able to give better fits compared to a simple model. So how do we decide?

The story of A and B:

Mr. A has a theory; Mr B also has one but with an adjustable parameter λ. Whose theory should we prefer given data $\mathcal{D}$?

To settle the issue, we should look at the relative merit of the two theories. If the

$$\text{posterior ratio} = \frac{\text{prob}(A|\mathcal{D},I)}{\text{prob}(B|\mathcal{D},I)}$$

is > 1 : prefer A

< 1 : prefer B

≈ 1 : no preference

Bayes theorem yields:

$$\frac{\text{prob}(A|\mathcal{D},I)}{\text{prob}(B|\mathcal{D},I)} = \frac{\text{prob}(\mathcal{D}|A,I)}{\text{prob}(\mathcal{D}|B,I)} \times \frac{\text{prob}(A|I)}{\text{prob}(B|I)}$$

Without λ, we can't quantify $\text{prob}(\mathcal{D}|B,I)$, so we proceed

marginalize.

$$\text{prob}(D|B,I) = \int \text{prob}(D,\lambda|B,I)\,d\lambda$$

$$= \int \underbrace{\text{prob}(D|\lambda,B,I)}_{\text{not an ordinary pdf.}} \cdot \text{prob}(\lambda|B,I)\,d\lambda$$

product rule

Assume that MC B only knows that $\lambda \in [\lambda_{min}, \lambda_{max}]$

so

$$\text{prob}(\lambda|B,I) = \frac{1}{\lambda_{max}-\lambda_{min}} \quad \text{for } \lambda_{min} \le \lambda \le \lambda_{max}$$

Assume further that $\exists\ \lambda_0$ which yields best fit and well approximated by Gaussian

$$\text{prob}(D|\lambda,B,I) = \text{prob}(D|\lambda_0,B,I)\exp\left[-\frac{(\lambda-\lambda_0)^2}{2\delta\lambda^2}\right]$$

In addition, we can take the prior $\text{prob}(\lambda|B,I)$ out of the $\int d\lambda$ integral (it does only depend on $\lambda_{max}, \lambda_{min}$)

So

$$\text{prob}(D|B,I) = \frac{1}{\lambda_{max}-\lambda_{min}} \int_{\lambda_{min}}^{\lambda_{max}} \text{prob}(D|\lambda,B,I)\,d\lambda$$

Furthermore, we assume that $\lambda_{max}, \lambda_{min}$ does not truncate the λ-peak around λ_0 so that we can perform the Gaussian integral yielding

$$\text{prob}(D|B,I) = \frac{1}{\lambda_{max}-\lambda_{min}} \int_{\lambda_{min}}^{\lambda_{max}} \text{prob}(D|\lambda,B,I)\,d\lambda$$

$$= \frac{\text{prob}(D|\lambda_0,B,I) \times \delta\lambda\sqrt{2\pi}}{\lambda_{max}-\lambda_{min}}$$

So

$$\frac{\text{prob}(A \mid D, I)}{\text{prob}(B \mid D, I)} = \underbrace{\frac{\text{prob}(A \mid I)}{\text{prob}(B \mid I)}}_{=1 \text{ (blackjack)}} \times \underbrace{\frac{\text{prob}(D \mid A, I)}{\text{prob}(D \mid \lambda_0, B, I)}}_{\text{ratio of goodness to fit}} \frac{\lambda_{max} - \lambda_{min}}{\delta\lambda} \leftarrow \text{penalty}$$

penalty $\gg 1$, because $\lambda_{max} - \lambda_{min}$ usually much larger than $\delta\lambda$. This is called a Ockham factor or Ockham's razor.

Problem: If no knowledge about additional parameter known then infinite penalty. No big problem in practice, though, because usually boundaries given and then goodness to fit decides. If goodness to fit undecisive then Ockham effect kicks in.

If you take a look back at Bayes theorem:

$$\text{prob}(\lambda \mid D, B, I) = \frac{\text{prob}(D \mid \lambda, B, I)\ \text{prob}(\lambda \mid B, I)}{\text{prob}(D \mid B, I)}$$

you'll recognize that the denominator played a crucial role in ascertaining the relative merit of B over A. It is therefore called "evidence" or "marginal likelihood".

Hypothesis testing

Suppose we have a hypothesis H_1. To quantify how much we believe that this is true given the data D and information I, we need to evaluate the posterior

$$\text{prob}(H_1 \mid D, I) = \frac{\text{prob}(D \mid H_1, I) \times \text{prob}(H_1 \mid I)}{\text{prob}(D \mid I)}$$

Suppose we have another hypothesis H_2.

$$\frac{\text{prob}(H_1 \mid D, I)}{\text{prob}(H_2 \mid D, I)} = \frac{\text{prob}(D \mid H_1, I)}{\text{prob}(D \mid H_2, I)} \times \frac{\text{prob}(H_1 \mid I)}{\text{prob}(H_2 \mid I)}$$

let $H_2 = \overline{H_1}$ then

$$\text{prob}(D \mid I) = \text{prob}(D \mid H_1, I)\,\text{prob}(H_1 \mid I) + \text{prob}(D \mid \overline{H_1}, I)\,\text{prob}(\overline{H_1} \mid I)$$

where $\text{prob}(H_1 \mid I) + \text{prob}(\overline{H_1} \mid I) = 1$

The problem is that we can't compute $\text{prob}(D \mid \overline{H_1}, I)$, because even if we know that it is not H_1, we still don't know what $\overline{H_1}$ is. We would need a <u>specific</u> alternative to give quantitative answers.

Usual statistics would say that we should compute the χ^2 and compare to the expected deviation $\sqrt{N}$ for N measurements.

We would say that the misfit statistics serves the purpose to think about alternatives.

4.3 Examples

Common problem: classification. Archeologist finds two skeletons differing by a million years. He wants to know if there is significant evolutionary change. Suppose the two sets yield N_1 and N_2 measurements represented by $\vec{D}_1$ and $\vec{D}_2$. Consider the following hypotheses:

A: no change over period of time. Both sets characterized by some (unknown) mean μ and standard deviation σ.

B: There is change with unknown μ_1, μ_2 and σ_1, σ_2.

We need to compute the evidence. So we need to compute $\text{prob}(\vec{D}_1, \vec{D}_2 \mid A, I)$ and $\text{prob}(\vec{D}_1, \vec{D}_2 \mid B, I)$. Let's start.

$$\text{prob}(\vec{D}_1, \vec{D}_2 \mid A, I) = \iint \text{prob}(\vec{D}_1, \vec{D}_2 \mid \mu, \sigma, A, I) \, \text{prob}(\mu, \sigma \mid A, I) \, d\mu \, d\sigma$$

We use a uniform prior

$$\text{prob}(\mu, \sigma \mid A, I) = \frac{1}{(\mu_{max} - \mu_{min}) \sigma_{max}}$$

$N \equiv N_1 + N_2$ treating as independent N measurements:

$$\text{prob}(\vec{D}_1, \vec{D}_2 \mid \mu, \sigma, A, I) = (\sigma \sqrt{2\pi})^{-N} \exp\left[-\frac{1}{2\sigma^2} \sum^{N} (x_k - \mu)^2\right]$$

and the $\int d\mu\, d\sigma$ integral yields

$$\int_{\mu_{min}}^{\mu_{max}} \int_{0}^{\sigma_{max}} \exp\left[-\frac{1}{2}\left(\alpha(\mu-\mu_0)^2 + \beta(\sigma-\sigma_0)^2\right)\right] d\mu\, d\sigma \approx \frac{2\pi}{\sqrt{\alpha\beta}}$$

$$\Rightarrow \text{prob}(\vec{D}_1, \vec{D}_2 \mid A, I) \approx \frac{\left(\sigma_0\sqrt{2\pi}\right)^{2-N} \exp(-N/2)}{(\mu_{max}-\mu_{min})\,\sigma_{max}\,N\sqrt{2}}$$

Having computed this, we also need

$$\text{prob}(\vec{D}_1, \vec{D}_2 \mid B, I) = \text{prob}(\vec{D}_1 \mid B, I) \times \text{prob}(\vec{D}_2 \mid B, I)$$

where we are allowed to factorize, because our hypothesis said that the two are independent.
Again, each of the two can be written as

$$\text{prob}(\vec{D}_i \mid B, I) = \iint \text{prob}(\vec{D}_i \mid \mu_i, \sigma_i, B, I)\, \text{prob}(\mu_i, \sigma_i \mid D, I)\, d\mu_i d\sigma_i$$

Both priors $i = 1, 2$ can be set the same, because we have no evidence for the contrary.
So the result for each looks exactly like the one for A:

$$\text{prob}(\vec{D}_i \mid B, I) \approx \frac{\left(\sigma_{0i}\sqrt{(2\pi)}\right)^{2-N_i} \exp\left(-\frac{N_i}{2}\right)}{(\mu_{max}-\mu_{min})\,\sigma_{max}\,N_i\sqrt{2}}$$

Finally, we divide the two evidences:

$$\frac{prob(\vec{D}_1, \vec{D}_2 \mid A, I)}{prob(\vec{D}_1, \vec{D}_2 \mid B, I)} = \frac{(\mu_{max} - \mu_{min})\sigma_{max}}{\pi\sqrt{2}} \quad \frac{N_1 N_2 \ (\sigma_0)^{2-U}}{N \ (\sigma_{01})^{2-N_1} \ (\sigma_{02})^{2-N_2}}$$

THE LIGHT BULB EXAMPLE

Consider 2 manufacturers of light bulbs. We know that their mean life-time and variance will be different. But which are better, given some samples we received?

What we need is the posterior probability of the hypothesis $\mu_1 > \mu_2$:

$$prob(\mu_1 > \mu_2 \mid \vec{D}_1, \vec{D}_2, I) = \int_0^{\infty} d\mu_1 \int_0^{\mu_1} d\mu_2 \ prob(\mu_1, \mu_2 \mid \vec{D}_1, \vec{D}_2, I)$$

Surely, the expected lifetime factorizes:

$$prob(\mu_1, \mu_2 \mid \vec{D}_1, \vec{D}_2, I) = prob(\mu_1 \mid \vec{D}_1, I) \ prob(\mu_2 \mid \vec{D}_2, I)$$

If the number of samples "N_i" is reasonably large, then

$$prob(\mu_i \mid \vec{D}_i, I) \approx \frac{\sqrt{N_i}}{2\pi S_i} \exp\left[-\frac{1}{2} \frac{N_i (\mu_i - \mu_{0i})^2}{S_i^2} \right]$$

Where $\mu_{0i} = \frac{1}{N_i} \sum$ lifetimes

$$S_i = \frac{1}{N_i - 1} \sum (\text{lifetimes} - \mu_{0i})^2$$

So

$$\text{prob}(\mu_1>\mu_2 \mid \vec{D}_1, \vec{D}_2, I) = \int_0^{\infty} d\mu_2 \int_0^{\mu_1} d\mu_2 \frac{\sqrt{N_1 N_2}}{2\pi S_1 S_2} \exp\left[-\frac{1}{2}\frac{N_1(\mu_1-\mu_{01})^2}{S_1^2} - \frac{1}{2}\frac{N_2(\mu_2-\mu_{02})^2}{S_2^2} -\right.$$

Make a change of variables $z = \mu_1 - \mu_2$,

then the integral becomes

$$\text{prob}(\mu_1>\mu_2 \mid \vec{D}_1, \vec{D}_2) = \frac{1}{S_z\sqrt{2\pi}} \int_0^{\infty} \exp\left[-\frac{(z-z_0)^2}{2S_z^2}\right] dz$$

where $z_0 = \mu_{01} - \mu_{02}$ $\qquad S_z^2 = \frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}$

So if the difference $z_0 = \mu_{01} - \mu_{02}$ (which we inferred from our samples) equals S_z, then $\text{prob}(\mu_1>\mu_2) = 0.84$ and so on...