9/11/19

Physics 8805: Learning from Data: Lecture 6

On board:

Notebooks for today:

- parameter_estimation_fitting_straight_line_I.ipynb
- amplitude_in_presence_of_background.ipynb
- Metropolis_Poisson_example.ipynb
- MCMC-random-walk-and-sampling.ipynb

Websites for mcmc visualization (mcmc visualization page on Canvas):
http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/
https://chi-feng.github.io/mcmc-demo/

** Questions or problems for Assignment 1?

For today: go back first to (26), (33), (34) [fitting straight line 1 notebook]

- Follow-up to priors: where does $p(m) \propto 1/(1+m^2)^{3/2}$ come from?
- We can consider the line as $y = mx + b$ or $x = m'y + b'$ where $m' = 1/m$ and $b' = -b/m$ gives the same result.
- Let us require that the priors on m, b have the same functional form as m', b', because the labeling is arbitrary.
- Then for the probabilities we must have

$$p(m,b)\,dm\,db = p(m',b')\,dm'\,db' = p\left(\frac{1}{m}, -\frac{b}{m}\right)\left|\begin{matrix} \frac{\partial m'}{\partial m} & \frac{\partial m'}{\partial b} \\ \frac{\partial b'}{\partial m} & \frac{\partial b'}{\partial b} \end{matrix}\right| dm\,db$$

- Evaluating the Jacobian gives $\frac{1}{m^3}$, so we need to solve the functional equation:

$$p\left(\frac{1}{m}, -\frac{b}{m}\right) = m^3\, p(m,b)$$

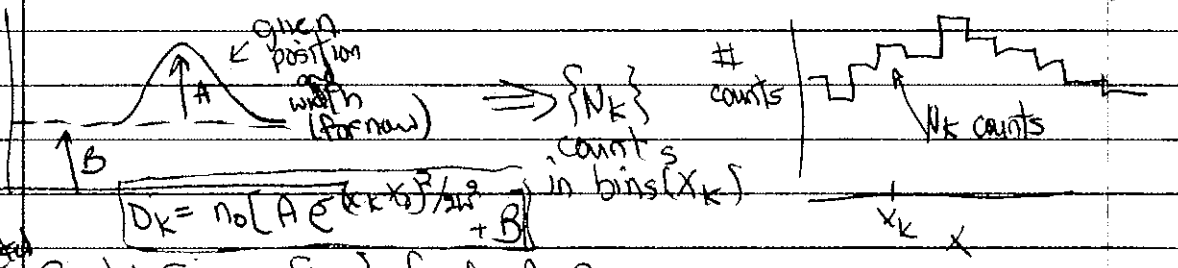
- A solution is $p(m,b) = \frac{1}{(1+m^2)^{3/2}}$, as claimed. ← arbitrary ∈

- How would you solve this without guessing the answer?

- Continue with the Sivia example on signal on top of background

9/11/19

Sina example: Amplitude of a signal in the presence of background



Goal: Given $\{N_k\}$ find A, B

So what is the posterior we want?

$p(A, B | \{N_k\}, I)$ with $I = x_0, w$, Gaussian, flat background

Actual counts we get will be integers, and we can expect a Poisson distribution

$$\Rightarrow p(n|\mu) = \frac{\mu^n e^{-\mu}}{n!} \quad \text{for } n \geq 0 \text{ integer}$$

or

$n \to N_k, \mu \to D_k$ $\quad p(N_k|D_k) = \frac{D_k^{N_k} e^{-D_k}}{N_k!}$ for k^{th} bin at x_k

* What do we learn from the plots of the Poisson distribution?

$$p(A,B|\{N_k\}, I) \propto p(\{N_k\}|A,B,I) \times p(A,B|I)$$

posterior $\propto$ likelihood $\times$ prior

$$\Rightarrow L = \log[p(A,B|\{N_k\}, I)] = \text{constant} + \sum_{k=1}^{n} [N_k \log(D_k) - D_k]$$

- Choose constant for convenience: independent of A, B.
- Best point estimate: maximize $L(A,B)$ to find A_0, B_0.
- Look at code for likelihood and prior

$$p(A,B|I) = \begin{cases} \text{constant} & 0 \leq A \leq A_{max},\ 0 \leq B \leq B_{max} \\ 0 & \text{otherwise} \end{cases}$$

- Uniform flat prior for $0 \leq A \leq A_{max}$, $0 < B \leq B_{max}$
  - Not sensitive to A_{max}, B_{max} if larger than support of likelihood

Fig #	data bins	Δx	$(x_k)_{max}$	D_{max}
1	15	1	7	100
2	15	1	7	10
3	31	1	15	100
4	7	1	3	100

9/11/19

Comments on Figures.

Fig 1: 15 bins and $D_{max} = 100$

- Contours are at 20% intervals showing height.
- Read off best estimates and compare to True.
  - does find signal is about half background
- Marginalization of B
  - what if we don't care about B? "nuisance parameter"

$$p(A|\{N_k\}, I) = \int_0^\infty p(A, B|\{N_k\}, I)\, dB$$

compare to $p(A|\{N_k\}, B_{true}, I)$ ⇒ plotted on graph

- Also can marginalize over A

$$p(B|\{N_k\}, I) = \int_0^\infty p(A, B|\{N_k\}, I)\, dB$$

- See how these are done in code: B_marginalized
  B_true fixed
  - note the normalization at the end.

- Set extra plots to true
  - different representations of same info and contours in first 3. Last one is attempt at 68%, 95%, 99.7% but looks wrong.
* - note difference between contours showing pdf height and showing integrated volume.

- Look at the other figures and draw conclusions!

- How should you design your experiments?
  Eg, how should you bin data, how many counts are needed, what $(x_k)_{max}$?, and so on.

9/11/19

Why MCMC? (Based on Gregory, Chp 12)

We have been emphasizing that in the Bayesian approach, everything is a pdf. One type of pdf is for the parameters of theory, which we'll denote by the vector $\vec{\theta}$, given data D:

$$p(\vec{\theta} \mid D, I)$$

and information I. Suppose we have a theoretical model for this.

- Maybe these are the LECs for an effective field theory Hamiltonian.

And now we want to calculate the expectation value of a function of $\vec{\theta}$: $\langle f(\vec{\theta}) \rangle$. Or θ characterizes a signal and background.

- As we discussed in doing the central limit theorem:

$$\langle f(\vec{\theta}) \rangle = \int f(\vec{\theta})\, p(\vec{\theta} \mid D, I)\, d\vec{\theta} \equiv \int g(\vec{\theta})\, d\vec{\theta}$$

Note that this is more than traditional calculations, in which we would have single values of $\vec{\theta}$, e.g. denoted $\hat{\vec{\theta}}$, that we might have found by minimizing a χ^2. E.g. we identified the particular values of $\theta_1, \theta_2, \ldots \theta_n$ that best reproduced scattering data. Then we would calculate $f(\hat{\vec{\theta}})$, which might be the binding energy of a nucleus.

- But $\langle f(\vec{\theta}) \rangle$ means we do a multidimensional integral over the full range of possible $\vec{\theta}$ values, weighted by the probability density function $p(\vec{\theta} \mid D, I)$, which we have worked out.
- This is a lot more work!
- We frequently also have a situation where we want to integrate (marginalize) over a subset of parameters $\vec{\theta}_B$ to find a probability for the rest $\vec{\theta}_A$. E.g. over parameters for the width of a signal and other parameters characterizing our model for the Higgs mass.
- These multidimensional integrals then become a necessity to do, but conventional methods for low dimension (e.g. Gaussian quadrature or Simpson's rule) become inadequate rapidly with the increase of dimension.

9/11/19

· The integrals are particular problematic because the posteriors are very small in much of the integration volume, which will typically be a challenging shape.

· To approximate such integrals one turns to Monte Carlo methods. The straight naive MC integration evaluates the integral by randomly distributing n points in the multidimensional volume V of possible $\vec{\theta}$'s. V has to be large enough to cover where $p(\vec{\theta}|D,I)$ is significantly different from zero.

· Then $\langle f(\vec{\theta})\rangle = \int g(\vec{\theta})\,d\vec{\theta} \doteq V \times \langle g(\vec{\theta})\rangle \pm V\sqrt{\frac{\langle g^2(\vec{\theta})\rangle - \langle g(\vec{\theta})\rangle^2}{n}}$

where $\langle g(\vec{\theta})\rangle = \frac{1}{n}\sum_{i=1}^{n} g(\vec{\theta}_i)$ (not components of $\vec{\theta}$, but separate vectors) $\langle g^2(\vec{\theta})\rangle = \frac{1}{n}\sum_{i=1}^{n} g^2(\vec{\theta}_i)$

[e.g. in one-D, the average of a function " $\overline{g(\theta)} = \frac{1}{b-a}\int_a^b g(\theta)d\theta$ " from calculus. But we can estimate $\overline{g(\theta)}$ by averaging over a set n of random samples $\overline{g(\theta)} \doteq \frac{1}{n}\sum_{i=1}^{n} g(\theta_i)$

Then $\langle f(\theta)\rangle \doteq \int_a^b g(\theta)d\theta \doteq \frac{b-a}{n}\sum_{i=1}^{n} g(\theta_i)$. Here b−a is the volume V.

(cf lighthouse)

Does this always work? No! But if CLT is ok, it should work, but maybe not efficiently.)

· The uncertainty is assuming a Gaussian approximation is valid.

· Note the dependence on $\frac{1}{\sqrt{n}}$, which means you can get a more precise answer by increasing n. Slowly better, each additional decimal point accuracy costs you a factor of 100 in n!

· Key problem: too much time is wasted sampling regions where $p(\vec{\theta}|D,I)$ is very small. If for one parameter the fraction of significant is 10^{-1}, in M-parameter problem the fraction of the volume is 10^{-M}. This necessitates importance sampling which reweights the integrand to more appropriately distribute points (e.g. VEGAS) but this is difficult to accomplish.

9/11/19

Bottom line: its not feasible to draw a series of independent random samples from $p(\vec{\theta}|D,I)$ for larger $\vec{\theta}$.

↳ remember, independent means if $\vec{\theta}_1, \vec{\theta}_2, \ldots$ is the series, knowing $\vec{\theta}_1$ doesn't tell us anything about $\vec{\theta}_2$

** But the samples don't need to be independent, they just need to generate $p(\vec{\theta}|D,I)$ in the correct proportions (eg. as indicated by histogramming the samples, it approximates $p(\vec{\theta}|D,I)$.

* ⇒ Do a random walk in the parameter space of $\vec{\theta}$ so that the probability for being in a region is proportional to $p(\vec{\theta}|D,I)$ for that region.

- $\vec{\theta}_{i+1}$ follows from $\vec{\theta}_i$ by a transition probability (kernel) ⇒ $p(\vec{\theta}_{i+1}|\vec{\theta}_i)$
- assumed to be "time independent", so same $p(\vec{\theta}_{i+1}|\vec{\theta}_i)$ no matter when you do it.

⇒ Markov chain and method is Markov chain Monte Carlo.

Basic structure of algorithm:

("candidate")

(1) Given $\vec{\theta}_i$, propose a value for $\vec{\theta}_{i+1}$, call it $\vec{\phi}$, sampled from $q(\vec{\phi}|\vec{\theta}_i)$. This q could take many forms, so for concreteness imagine it as a multivariate normal with mean given by $\vec{\theta}_i$ and variance $\vec{\sigma}^2$.

- decreased probability as you get away from current sample
- $\vec{\sigma}$ determines the step size.

(2) Decide whether or not to accept candidate $\vec{\phi}$ for $\vec{\theta}_{i+1}$. Here we'll use a Metropolis condition (later we'll see other ways that may be better).

- This dates from the 1950's in physics but didn't become widespread in statistics until almost 1980.
- Enabled Bayesian methods to take off.

9/11/19

Calculate Metropolis ratio:

$$r = \frac{p(\vec{\phi}|D,I)\, q(\vec{\theta}_i|\vec{\phi})}{p(\vec{\theta}_i|D,I)\, q(\vec{\phi}|\vec{\theta}_i)}$$

($\vec{\phi}$: proposed; $\vec{\theta}_i$: current)

q may be symmetric $q(\vec{\theta}_1|\vec{\theta}_2) = q(\vec{\theta}_2|\vec{\theta}_1)$

if so ⇒ "Metropolis"

If not, Run "Metropolis-Hastings."

Decision:

if $r \geq 1$, set $\vec{\theta}_{i+1} = \vec{\phi}$ accept

if $r < 1$, so less probable, don't always reject!

accept with probability r (remember $0 \leq r \leq 1$)

by sampling a uniform $(0,1)$ distribution.

If $U \sim \text{Uniform}(0,1)$ is $U \leq r$, then $\vec{\theta}_{i+1} = \vec{\phi}$

else $\vec{\theta}_{i+1} = \vec{\theta}_i$

Note that the last case means you do have a $\vec{\theta}_{i+1}$, but it is the same as $\vec{\theta}_i$ (so the chain continues to grow).

Acceptance probability is the minimum of 1, r

Algorithm pseudo code:

1. initialize $\vec{\theta}_i$, set $i = 0$
2. Repeat { Obtain new candidate $\vec{\phi}$ from $q(\vec{\phi}, \vec{\theta}_i)$
   Sample $U \sim \text{uniform}(0,1)$
   If $U \leq r$ set $\vec{\theta}_{i+1} = \vec{\phi}$, else set $\vec{\theta}_{i+1} = \vec{\theta}_i$
   $i++$
   }

Plan:
1. look at visualizations
2. look at a basic example for Poisson distribution
3. consider MCMC-random-walk and sampling, notebook.
4. look at emcee example from Assignment 1.