Afternoon: Teaching Tower PT212

M1b-1

6/10/19 Bayes2019 TALENT Lecture M1b

- Notebooks:
  - (A) • from M1a - topics/basics-of-bayesian-statistics/Exploring_pdfs.ipynb
  - • topics/bayesian-parameter-estimation/
    - (B) parameter_estimation_in_bayesTALENT_intro.ipynb
    - (C) parameter_estimation_fitting_straight_line_I.ipynb

- Topic: Parameter Estimation I (of 3)

- Overview comments
  - In general terms, "parameter estimation" in physics means obtaining values for parameters (constants) that appear in a theoretical model that describes data. (Exceptions exist, of course)
  - Conventionally this process is known as "fitting the parameters" and the goal is to find the "best fit".
  - We will make particular interpretations of these phrases from our Bayesian point of view.
  - Today we'll set up the problem and look at how familiar ideas like "least-squares fitting" show up from a Bayesian perspective.

- There are many examples of where parameter estimation is needed in low-energy nuclear physics (as examples) and every other subfield of physics.
  - the parameters in an effective field theory Hamiltonian (chiral, pionless, halo, deformed nuclei) - usually called low-energy constants or LECs

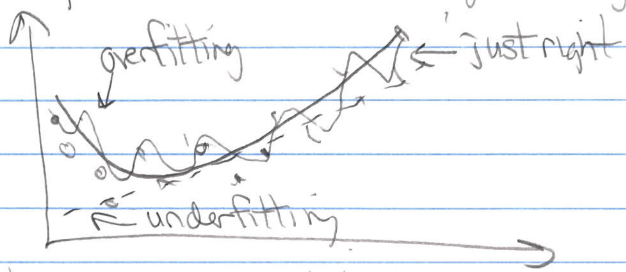
    we'll do a toy model of this to explore issues
  - parameters that define an energy density functional (eg. Skyrme type)
  - parameters in an optical potential used in reactions calculations
  - parameters in a model for extrapolating to an infinite model space (eg. no-core shell model)
  - and so on.

- As we proceed, we will make the case that a Bayesian approach is the way to go.

M1b-2

6/10/19

As a teaser, lets ask: what can go wrong in a fit?



Bayesian methods can prevent/identify both underfitting (model is not complex enough to describe the data) or overfitting (model tunes to data fluctuations or terms are underdetermined, leading to them playing off each other).

- We'll see how this plays out!

Let's step through part of the notebook you were sent last month — with some supplementary material.

- Load notebook (B). We'll run using RISE. But you just use it normally. :)
- We'll include "footnotes" here on Python, Jupyter, Bayesian statistics, physics

- Import of modules
  - Note "cell magic" %matplotlib inline (alternative %matplotlib notebook, has interactive figures)
  - use of seaborn here is just to make the graphs look good.
  - We'll use emcee (cf. MC) to do "sampling" later. corner is used to make a particular type of plot.

- Example from Sivia's book: Gaussian noise and averages.
  - This is an excellent book!

- $p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ ← μ, σ are given. Probability density of x given μ, σ. Normalized $\Rightarrow \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$

  dimensions of $1/x$

- Justification as theoretical model in Bayesian circles from maximum entropy. Usually justified from "central limit theorem". How many know about that?

6/10/19

- M measurements $D \equiv \{x_k\} = (x_1, \ldots, x_M)$ eg. $M=100$
  distributed according to $p(x|\mu,\sigma)$.

- How do we get them here? As in Exploring-pdfs.ipynb.
  - "Sample" from $N(\mu,\sigma^2)$ $\Rightarrow$ we'll see this.

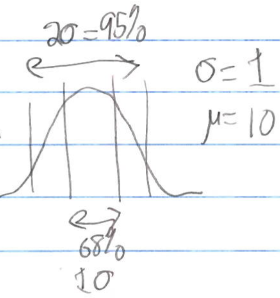
- Goal: find approximate μ, σ
  Frequentist: maximum likelihood method
  Bayesian: compute posterior pdf $p(\mu,\sigma|x,I)$ ← other information

- Random seem of 1 means same series of random numbers. If you put 2 or 42, then different from 1, but still the same with every run.

- stats.norm.rvs as in Exploring-pdfs.ipynb
  - size=M is a "keyword argument" (often kw ≡ keyword)
    $\Rightarrow$ optional and there is a default value (here 1).

- shift-tab-tab after evaluating cell.
  - e.g. place on "norm" or "rvs"

- Output D is a numpy array. ← everything in Python is an object. So more than just a datatype $\Rightarrow$ extra methods.
  - Put cursor after D (here) and shift-tab-tab
  - [...] when printed.



- Discuss about number of entries in tails amongst selves.
  - Hint: "tail" of Gaussian, say beyond 2σ
    $\Rightarrow x > 12$ or $x < 8$.
  - How many to you expect on average?
    2σ $\Rightarrow$ 95% so about 5/100.
  - Here 4 in that range. If there were zero is there a bug?
    No, there is a chance that will happen.

- Note the pattern (or lack) and repeat to get different numbers. How?
  Change the random seed from 1. (You are invited to try.) Always play!

6/10/19

- Questions about plotting?
  - We'll repeatedly use constructions like this, so get used to it!
  - ; means we put on same line. Not necessary.
  - alpha=0.5 just makes the (default) color lighter.
  - try color='red' on your own in scatter plot (as in vlines)
  - might prefer side-by-side ⇒ alternative code.
  - An "axis" in matplotlib means an entire subfigure, not just the x-axis or y-axis.
  - If you want to know about a plotting command already there, shift-tab-tab (usually, sometimes not).
  - To find vlines (vertical lines), google "matplotlib vertical line". (try it).
  - fig.tight_layout() for good spacing with subplots.

- Ask questions this afternoon and throughout if you are confused by code!
- Observations on graphs?
  - scatter plot shows tail ⇒ in this case there are 5, but rerun and it will be more or less ⇒ everything is a pdf
  - histogram is imperfect. Problem? cf. Exploring_pdfs at end (sampling 1D pdfs)
  - tails fluctuate

- Frequentist approach
  - true value for parameters μ, σ, not a pdf
  - Use of $\mathcal{L}$ is notation commonly used
  - * Why the product? Assumed independent. Reasonable?
  - $\log \mathcal{L}$ for several reasons (note: "log" always means ln. If we want base 10, then $\log_{10}$)

$\mathcal{L} \doteq (\text{const})\, e^{-\chi^2}$ so maximizing $\mathcal{L}$ = maximizing $\log \mathcal{L}$ = minimizing χ^2

- You can all carry out the maximization

e.g. $\dfrac{\partial \log \mathcal{L}}{\partial \mu} = -\dfrac{1}{2}\sum_{i=1}^{M} 2\cdot\dfrac{(x_i-\mu)}{\sigma^2}\cdot -1 = \dfrac{1}{\sigma^2}\sum_{i=1}^{M}(x_i-\mu) = \dfrac{1}{\sigma^2}\left(\left(\sum_{i=1}^{M} x_i\right) - M\mu\right)$

⇒ set to zero ⇒ $M\mu_0 = \sum_{i=1}^{M} x_i$ or $\mu_0 = \frac{1}{M}\sum_{i=1}^{M} x_i$. You do σ_0^2. (easier than $\frac{\partial}{\partial\sigma}$ is $\frac{\partial}{\partial\sigma^2}$)

M16-5

6/10/19

* Do these make sense?
  - μ_0 is mean of data → estimator for "true mean"
  - σ_0 gives spread about μ_0.

- Note use of .sum to add up D array elements
- Printing with f strings
  f'...' or f""" ... """ ← multiline string
  - .2f means float with 2 decimal points

- Note comment on "unbiased estimator"
  - an accurate statistic
  - Here compare μ_0 from $\frac{1}{M}$ and $\frac{1}{M-1}$
  - If you do this many times, you'll find $\frac{1}{M}$ doesn't quite give μ_{true} correctly (take mean of μ_0's from many trials), but $\frac{1}{M-1}$ does. (Try it!)
  - The difference is $O(\frac{1}{M})$, so small for large M.

- Compare estimates to true. Are they good estimates? How can you tell? E.g. should they be within .1, .01, or what?
  ⇒ more as we go!

- Bayesian approach
  $p(\mu, \sigma \mid D, I)$ is posterior: probability (density) of finding some μ, σ given data D and what else we know (I).
  - "I" could be that $\sigma > 0$ or μ should be near zero.

Frequentist probability: long-run frequency of (real or imagined) trials.
⇒ data is probabilistic (repeat experiment and get different result) but model parameters are not (universe stays the same with more observation)

Bayesian probability: quantification of information (what you know, often said "what you believe"). Data are fixed (it's what you found) but knowledge of true model parameters is fuzzy (and gets update with more trials — coin flipping).

M1b6

6/10/19

class: you label each term

Bayes' theorem

$$\text{posterior} \rightarrow p(\mu,\sigma|D,I) = \frac{\overset{\text{likelihood}}{p(D|\mu,\sigma,I)}\, p(\mu,\sigma|I) \leftarrow \text{prior}}{p(D|I)} \leftarrow \text{data probability}$$

(or "fully marginalized likelihood" or evidence or ...)

It will become intuitive!

- tells you how to flip $\underset{\text{hard}}{p(\mu,\sigma|D,I)} \longleftrightarrow \underset{\text{easy}}{p(D|\mu,\sigma,I)}$

Aside on denominator

general vector of parameters

$$p(D|I) = \int p(D|\theta,I)\, p(\theta)\, d\theta$$

so integrate ("marginalize") over all values of θ. Numerically costly $\Rightarrow$ more later on how to do it.

(parameter)

- For model fitting, we don't need $p(D|I)$ calculated. Find the posterior and just normalize that function (or we might only need relative probabilities).

- If $p(\mu,\sigma|I) \propto 1 \Rightarrow$ "flat prior" (more later)
then
$$p(\mu,\sigma|D,I) \propto \mathcal{L}(D|\mu,\sigma)$$

then F and B get same answer for most likely values μ_0, σ_0 (called "point estimates" as opposed to a full pdf)

- Back to the prior $\Rightarrow$ include additional information.
What you know before a measurement.
  - We will talk much more.
  - F says it is nonsense; subjective, individual

$\Rightarrow$ discuss this amongst yourselves.

- How to compute $p(\mu,\sigma|D,I)$ in practice? Often with MCMC. Just look now and return Wednesday

MIb-7

6/10/19

Now turn to parameter_estimation_fitting_straight_line_I.ipynb
Fitting a straight line.

Annotations:

- same imports as before
- assume we create data from underlying

$$y_{exp}(x) = m_{true}x + b_{true} + \text{gaussian noise}$$

mean zero

$\Theta_{true} = [b_{true}, m_{true}] = [\text{intercept}, \text{slope}]_{true}$ fixed $\sigma = dy$ in code

- x_i points are also chosen randomly according to uniform distribution ⇒ rand.rand(N)
- errors are normal ⇒ y += dy * rand.randn(N)
- These use the numpy random number generators & while we will mostly use scipy.stats (see other codes)

- Theoretical model: $y_{th}(x) = mx + b$ with $\Theta = [b, m]$

$$y_{exp} = y_{th} + \delta y_{exp} + \delta y_{th}$$

← critically important but has often (mostly?) been neglected

normally distributed

$$\Rightarrow y_i \sim N(y_{th}(x_i;\Theta), \sigma^2)$$

should be ε_i^2

mean

usually squared

- Is independent a good assumption?

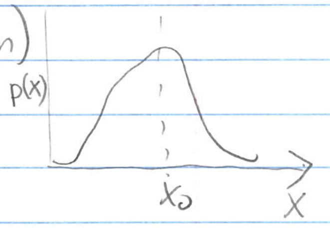
- Priors ~ quick run through
  - Major point: when does prior matter?

6/10/19

General reason why Gaussians may show up:

- Given $p(x|D,I)$, then our "best estimate" from

$$\frac{dp}{dx}\Big|_{x_0}=0 \text{ with } \frac{d^2p}{dx^2}\Big|_{x_0}<0 \text{ (maximum)}$$



Look nearby to characterize posterior $p(x)$.

$p(x)$ varies too fast, so characterize $\log p$

$$\Rightarrow L(x)=\log p(x|D,I)=L(x_0)+\cancel{\frac{dL}{dx}\Big|_{x_0}}^{\,0}+\frac{1}{2}\frac{d^2L}{dx^2}(x-x_0)^2+\ldots$$

- If we can neglect higher order terms, then

$$p(x|D,I)\approx A\, e^{\frac{1}{2}\frac{d^2L}{dx^2}\big|_{x_0}(x-x_0)^2}$$

$\nwarrow$ normalization

$\Rightarrow$ very generally looks like Gaussian.

$$p(x|D,I)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{\sigma^2}} \Rightarrow \mu=x_0,\ \sigma=\left(-\frac{d^2L}{dx^2}\Big|_{x_0}\right)^{-1/2}$$

- we usually quote $x=x_0\pm\sigma$, because if Gaussian, this is <u>sufficient</u> to tell us the entire distribution.

- For Bayesian: full posterior $p(x|D,I)$ for $\forall x$ is general result, and $x=x_0\pm\sigma$ may be an approximate characterization.

- What if asymmetric $p(x|D,I)$? Multimodal?

6/10/19

Bayesian vs. Frequentist 95% confidence interval

- Bayesian is easy; a credible interval or Bayesian confidence interval or degree-of-belief (DOB) interval is: given some data, 95% chance (probability) that the interval contains the true parameter.

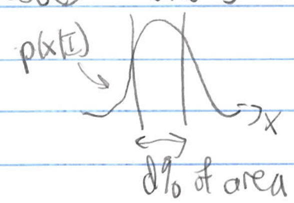
- Frequentist 95% confidence interval
  - If large # of repeat samples, 95% of these intervals include the true value of the parameter
  - So the parameter is fixed (no pdf) and the confidence interval depends on data. (random sampling)

"There is a 95% probability that when I compute a confidence interval from data of this sort that the true value of θ will fall within the (hypothetical) space of observations."

- What?

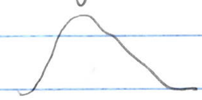
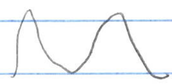
- One key difference: Bayesian includes prior.

- Issues: One-D → if symmetric pdf, then clear how to define d% confidence interval.



Algorithm: start from center, step outward adding area, stop at d%.

Two-D: need a way to integrate from top.

- What if asymmetric or multimodal?  or 

Two of the possible choices:

- Equal-tailed interval (central interval): area above and below interval are equal

- Highest posterior density (HPD) region: Posterior density for every point is higher than the posterior density for any point outside the interval.