10/9/19

Physics 8805: Learning from Data: Lecture 14

On board: Winter break this week → no class Friday.
- Mini-project 2a is released. There will be a mini-project 2b.
- Both are are Model Selection. 2a is due Friday, Oct. 18.

- Notebooks for today:
  - model-selection_mini-project-IIa.ipynb
  - Evidence_for_model_EFT_coefficients.ipynb

- Comments on Assignment 2 and mini-project I
  - Review mini-project I via a key notebook.
    - Using pandas to read data into a dataframe
    - starting guesses histogram
    - What do traces tell us? Should check Gelman-Rubin ratio.
    - Explain uniform correlations.
    - Explain Gaussian correlations (not bad that they exist!)
    - Table III features
    - Figure 3 and bands ⇒ see how xfit, yfit then mu and std calculated
      - notice failure to extrapolate with good error.

emcee affine-invariant sampler can handle.

10/9/19

Bayesian Model Selection

- The discussion here is based heavily on Sivia, Chapter 4.

- We've mostly focused on parameter estimation: given a model with parameters, what is the joint posterior for those parameters given some data — this is what we Bayesians mean by fitting the parameters: $p(\vec{\theta}|D, I)$.

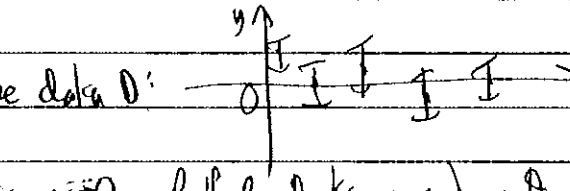
- Now we turn to an analysis of the model itself, or, more precisely, the comparison of models.
  - Remember that: model selection will always be about comparisons.
  - We can think of many possibilities: (all given data)
    - Is the signal a Gaussian or Lorentzian line shape?
    - When fitting to data, what order polynomial is best?
    - Given two types of Hamiltonian, which is better?
    - In an EFT, which power counting (organization of Feynman diagrams) is favored?

not best fit but most suitable

- If you decided by how well polynomials fit the data, by having the least residual, then higher order will always be better (or at least equal, as you can set coefficients to zero).

- So let's think how a Bayesian would proceed. As is usually best, we'll start with the simplest possible example, dating back to Jeffreys (1939) → Gull (1988) → Sivia (2006) → 8805 (2019):
  The story of Dr. A and Prof. B.:
  "Dr A has a theory with no adjustable parameters. Prof. B also has a theory, but with an adjustable parameter λ. Whose theory should we prefer on the basis of data D?"

- Eg. could be data D:



Dr. A thinks $y \equiv 0$. Prof B thinks $y = \lambda$, with λ to be determined.

10/9/19

Bayesian: consider ratio of posteriors

$$\frac{p(A|D,I)}{p(B|D,I)}$$

again, not prob. for particular instance of B, like $\lambda = 0.2$, but prob. that the theory is correct. λ doesn't appear yet.

Ok, let's proceed with Bayes' Theorem:

$$\frac{p(A|D,I)}{p(B|D,I)} = \frac{p(D|A,I)\,p(A|I)/p(D|I)}{p(D|B,I)\,p(B|I)/p(D|I)}$$

(likelihood: $p(D|A,I)$; prior: $p(A|I)$; $p(D|I) \to 0$) so denominator cancels in the ratio.

- The ratio $\frac{p(A|I)}{p(B|I)}$ might be given by our opinion of the two scientists based on their track records. But it is more typically taken to be 1 (before seeing data, no preference).

- Now $p(D|A,I)$ would seem to be straightforward, but what do we do about the λ in B? Marginalize!

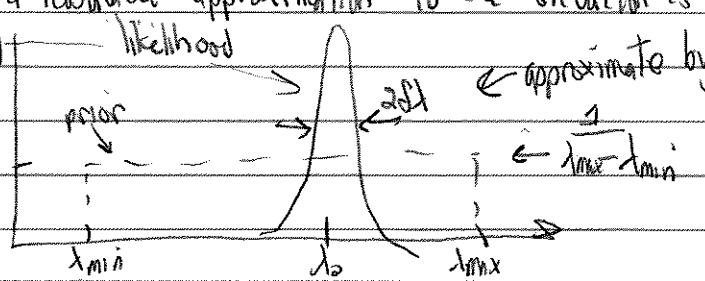
$$p(D|B,I) = \int p(D,\lambda|B,I)\,d\lambda = \int p(D|\lambda,B,I)\,p(\lambda|B,I)\,d\lambda$$

ordinary likelihood; prior pdf for λ; integrates over all λ

- Let's suppose a uniform prior for $p(\lambda|B,I)$:

$$p(\lambda|B,I) = \frac{1}{\lambda_{max} - \lambda_{min}} \quad \text{for } \lambda_{min} \le \lambda \le \lambda_{max} \text{ and zero otherwise.}$$

- Suppose also we can find λ_0 that maximizes the likelihood $p(D,\lambda,B,I)$. So we imagine a reasonable approximation to the situation is



10/9/19
Ptm

$$p(D|\lambda,B,I) \approx p(D|\lambda_0,B,I)\, e^{-\frac{(\lambda-\lambda_0)^2}{2\delta\lambda^2}}$$

(value at $\lambda=\lambda_0$)

Note that $p(\lambda|B,I)$ is normalized wrt λ, but $p(D|\lambda,B,I)$ is not → $p(D|\lambda_0,B,I)$ is not equal to $\frac{1}{\lambda_{max}-\lambda_{min}}$ (in general), [because λ is to the right of the |].

Now the prior doesn't depend on λ, so pull it out of the integral

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

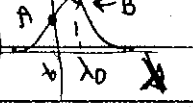
take $\lambda_{max}\to\infty$ and $\lambda_{min}\to-\infty$ with negligible error $\doteq \frac{1}{\lambda_{max}-\lambda_{min}} p(D|\lambda_0,B,I)\cdot \delta\lambda\sqrt{2\pi}$ ← integral over the Gaussian

Now put this together $\frac{p(A|D,I)}{p(B|D,I)} = \frac{p(A|I)}{p(B|I)} \times \left(\frac{p(D|A,I)}{p(D|\lambda_0,B,I)}\right) \times \left(\frac{\lambda_{max}-\lambda_{min}}{\delta\lambda\sqrt{2\pi}}\right)$

a priori ratio (usually unity) | i) ratio of likelihoods | ii) "Ockham factor" (Occam in latin)

⇒ competing factors

i) B with λ_0 will usually have better likelihood than A.
(always if nested theories)



ii) Another term! Penalize B for additional parameter since $\lambda_{max}-\lambda_{min} > \delta\lambda\sqrt{2\pi}$ (usually)

⇒ This is a formalization of Occam's razor: add parameters as long as the gain in likelihood beats the cost in complexity.

- Ockham factor is the ratio of volumes before and after data is known. A greater collapse means a greater penalty - cf. many parameters; full prior volume vs. effective likelihood volume.
- Doesn't always get interpreted so easily: think Gaussian vs. Lorentzian.

10/9/19

Jeffreys first worried about the prior: if the new parameter has infinite range, isn't the Ockham factor an infinite penalty? Indeed an issue if there are no constraints.

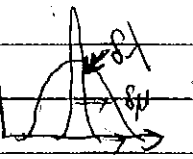
- A more reasonable case is the naturalness prior in Mini-project-I.
- In most physical cases, there are limits prescribed by some a priori knowledge.
- What happens if likelihood is wider than prior? ← caution! recognize the integral rather than using our formula

⇒ in model-selection_Mini-project-IIa.ipynb you will take an exploration of evidence for a simple case and test intuition.

Suppose Dr. A now also has an adjustable parameter μ.

$$\Rightarrow \frac{p(A|D,I)}{p(B|D,I)} = \frac{p(A|I)}{p(B|A)} \frac{p(D|\mu_0,A,I)}{p(D|\lambda_0,B,I)} \times \frac{\delta\mu\,(\lambda_{max}-\lambda_{min})}{\delta\lambda\,(\mu_{max}-\mu_{min})}$$

Suppose same prior range and equal a priori theories $\Rightarrow \frac{p(A|D,I)}{p(B|D,I)} \approx \frac{p(D|\mu_0,A,I)}{p(D|\lambda_0,A,I)} \times \frac{\delta\mu}{\delta\lambda}$



If data is good, likelihood ratio is most likely to dominate. If comparable, then shape with larger error bar for its parameter will be favored.
Why? Because in model selection, this means more parameter values are consistent with a good fit.

Finally: same theory but different prior range. Then $\frac{p(A|D,I)}{p(B|D,I)} \approx \frac{\lambda_{max}-\lambda_{min}}{\mu_{max}-\mu_{min}}$

⇒ prefer narrower prior range ⇒ must have more insight to give a narrow range for the parameter.

10/9/19

Compare to parameter estimation...

$$p(\lambda|D,B,I) = \frac{p(D|\lambda,B,I)\,p(\lambda|B,I)}{p(D|B,I)}$$

↑ this is what we've been calculating for model selection

Called evidence for B or marginal likelihood ← (marginal ⇒ integrate over all parameters in likelihood) or global likelihood or prior predictive.

Parameter estimation focuses on the maximum of the likelihood while model selection calculates an average of it.

Evidence calculations: Laplace's method

Suppose an unnormalized probability density $P^*(\theta)$ has a peak at θ_0. Let θ be K dimensional. Then the evidence is

$$Z_p = \int P^*(\theta)\, d^K\theta$$

If it is okay to expand $\log P^*(\theta)$ around its peak:

$$\log P^*(\theta) = \log P^*(\theta_0) - \tfrac{1}{2}(\theta-\theta_0)^T \Sigma^{-1} (\theta-\theta_0) + \dots$$

(saddle point approx.)

where $\Sigma^{-1} = H$ is the Hessian matrix $H_{ij} = -\dfrac{\partial^2}{\partial\theta_i \partial\theta_j} \log P^*(\theta)\Big|_{\theta=\theta_0}$

$\Rightarrow$ approximate $Q^*(\theta) = P^*(\theta_0)\, e^{-\frac{1}{2}(\theta-\theta_0)^T \Sigma^{-1} (\theta-\theta_0)}$

and approximate $Z_p \approx Z_Q = P^*(\theta_0)\sqrt{\dfrac{(2\pi)^K}{\det(\Sigma^{-1})}}$

If $P^*(\theta) = e^{-\frac{1}{2}\chi^2(\theta)} \Rightarrow Z_p \approx e^{-\frac{1}{2}\chi^2(\theta_0)} \sqrt{\dfrac{(2\pi)^K}{\det(\Sigma^{-1})}}$.

10/9/19

Quick look at model-selection-mini-project-II a.ipynb

- We have a non-polynomial function
  - Given data with errors, what polynomial is "best" to use.
  - Higher order polynomial will always reduce sum of residuals.
- Compare χ^2/dof in least-squares fits to use of Bayesian evidence ratios.
  - Using Laplace method here.
- Look at questions.
- How do the results depend on the data? # of data pts, range of data, size of error bars.

Quick look at Evidence_for_model_EFT_coefficients.ipynb

- Evidence with linear algebra and Gaussian integrals.
- Saturation of evidence. How to explain? (next time!)