10/15/19

Physics 8805: Learning from Data: Lecture 15

On board:

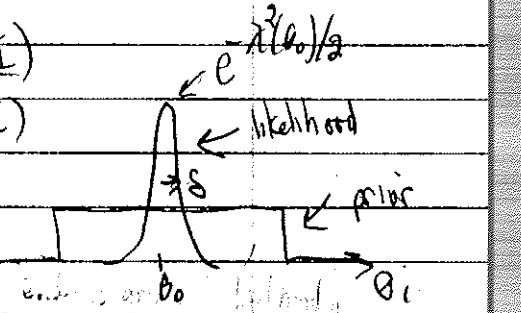
- Mini-project 2a is due Friday (really by end of Monday).

Notebooks for today:

model-selection_mini-project-IIa.ipynb
Evidence_for_model-EFT_coefficients.ipynb

- Review evidence ratio calculation ⇒ top of (78):

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



If model A depends on parameters $\vec{\theta}$, then

$$p(D|A,I) = \int p(D|\vec{\theta},A,I)\,p(\vec{\theta}|A,I)\,d\vec{\theta}$$

Occam factor = $p(D|\vec{\theta},A,I)$ (usual likelihood) × $p(\vec{\theta}|A,I)$ (prior for those particular parameters); integrate over all parameters

model A will have greater evidence when peak of likelihood increases more than the ratio of the parameter space pre-data volume & the post-data volume

- We did an approximate evaluation of $p(A|D,I)/p(B|D,I)$ in a special case. We'll see a similar argument today for the toy model in Mini-project-I.

- Do comparison to parameter estimation on (81)

- Do evaluation with Laplace's method on (81)

- Another look at model-selection-mini-project-IIa.ipynb. (82)

prior $p(\theta_i|I) = \frac{1}{\theta_{max}-\theta_{min}}$ if $\theta_{min} \le \theta_i \le \theta_{max}$, otherwise 0.

$$\int d\vec{\theta}\; p(D|\vec{\theta},A,I)\,p(\vec{\theta}|I) \overset{\text{schematic}}{\approx} \frac{1}{(\theta_{max}-\theta_{min})^k}\, e^{-\chi^2(\theta_0)/2}\, \frac{(2\pi)^{k/2}}{\sqrt{\det(\Sigma^{-1})}}$$

Does improvement in likelihood out-weigh the penalty from shrinkage of the parameter phase space?

10/15/19 (84)

Problem: In the toy model example, the evidence is highly sensitive to the prior ⇒ need criterion to determine prior.

Plan for us: Apply an informative prior from EFT naturalness.

⇒ look at Evidence_for_model_EFT_coefficients.ipynb
- Illustrate with toy model as in mini-project I.

- von Neumann quote ⇒ we will constrain use of higher-order parameters to prevent elephant fitting.

- Look at some EFT slides for additional motivation of expansion (request for more physics).
  ⇒
  - Look at classical analog in extra slides.

- Note on "models" EFT is said to be model independent because it uses the most general form consistent with symmetries of underlying physics, ⇒ no extra assumptions.
  - "model" is any theoretical construct for computing an observable.

- Model selection in EFT context could be with models having different dofs (nucleons only, nucleons plus pions, nucleons + Δs + pions) or for different orders in the same EFT (of model problem).
  ⇒ first is a frontier, 2nd demonstrated in paper with model
  ⇒ do this here.

- Look at punch line in Evidence_for_model_EFT_coefficients and then try to explain. Come back to details.

- Class: play with the effects of changing the range of data, the relative error, and the # of points.

10/15/9

Revisit two model discussion

- Models M_1, M_2 with same data set D.

- Evidence: $p(M_1|D,I)$ vs. $p(M_2|D,I)$ no reference to a particular parameter set $\Rightarrow$ comparison between two models, not two fits.
  - Note: In Bayesian model selection, only a comparison makes sense. One does not deal with the hypothesis like: "Model M_2 is correct."
  - Here M_2 is M_1 with one extra order (one more parameter) eventually

- Apply Bayes' Theorem

$$\frac{p(M_2|D,I)}{p(M_1|D,I)} = \frac{p(D|M_2,I)\,p(M_2|I)/p(D|I)}{p(D|M_1,I)\,p(M_1|I)/p(D|I)}$$

(cancel)

Bayes factor

we'll take $=1$ for our example $\Rightarrow$ no a priori preference for best order

(I): $$\frac{p(D|M_2,I)}{p(D|M_1,I)} = \frac{\int d\vec{a}_2\, p(D|\vec{a}_2,M_2,I)\,p(\vec{a}_2|M_2,I)}{\int d\vec{a}_1\, p(D|\vec{a}_1,M_1,I)\,p(\vec{a}_1|M_1,I)}$$

← recall steps
$p(D|M_2,I)$
$\rightarrow p(D,\vec{a}_2|M_2,I)$
$\Rightarrow p(D_2|M_2,\vec{a}_2,I)\,p(a_2,M_2)$

so integration over the entire parameter space.

$\Rightarrow$ Difficult numerically since likelihoods usually peaked but can have long tails that contribute to integral. (cf. Averaging over likelihood vs. finding peak)

- Easiest example $M_1 \rightarrow M_k$, $M_2 \rightarrow M_{k+1}$ } Is going to a higher-order favored by the given data? ← order in FFT expansion

Simplify: M_{k+1} has an additional parameter a' and assume priors Factorize

eg. $e^{-\vec{a}^2/2\sigma^2} \rightarrow e^{-a_0^2/2\sigma^2}\, e^{-a_1^2/2\sigma^2} \cdots e^{-a_k^2/2\sigma^2}$ for Gaussian

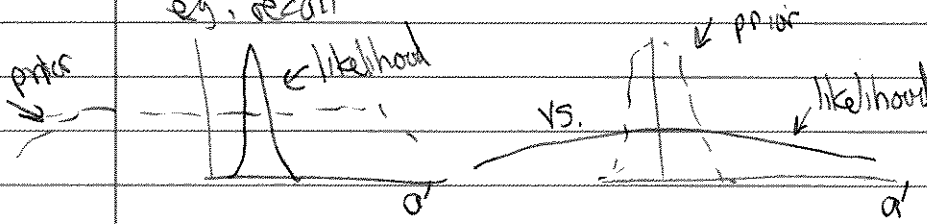
10/15/19

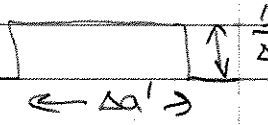
Then $p(\vec{a}_0 | M_{k+1}, I) = p(\vec{a}_1, a' | M_{k+1}, I) = p(\vec{a}_1 | M_{k+1}, I)\, p(a' | M_{k+1}, I)$

Consider cases . . .

i) values of a' that contribute to integrand in numerator of (I) are determined by the likelihood peaked region.

eg. recall



How can we approximate? $p(a' | M_{k+1}, I)$ 

Call the value of the likelihood peak $\hat{a}$ and the width $\delta a'$.

So two different widths: before data $\Delta a'$ (prior) and after $\delta a'$ (likelihood)

$$\Rightarrow \frac{p(D|M_{k+1}, I)}{p(D|M_k)} \doteq \frac{\delta a'}{\Delta a'} \frac{\int d\vec{a}_1\, p(D|\vec{a}_1, \hat{a}', M_{k+1}, I) \times p(\vec{a}_1|M_{k+1}, I)}{\int d\vec{a}_1\, p(D|\vec{a}_1, M_k, I)\, p(\vec{a}_1|M_k, I)}$$

(from integrating over a'; peak value; $\rightarrow$ posterior; from $p(a'|M_{k+1}, I)$)

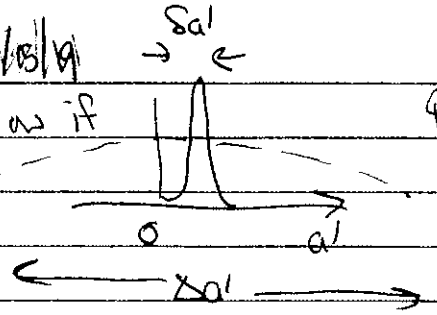
$\Rightarrow$ The ratio of the integrals is the gain in the likelihood from an extra parameter with value $\hat{a}$ (cf. $M_{k+1}(\hat{a}'=0) = M_k$)

- But also "Occam factor" or "Occam penalty" $\frac{\delta a'}{\Delta a'}$

$\Rightarrow$ How much parameter space collapses in face of data, we thought initially that a' could be anywhere in $\Delta a'$, but find after data it is only in $\delta a'$. What a waste (less predictive) if $\delta a' \ll \Delta a'$.

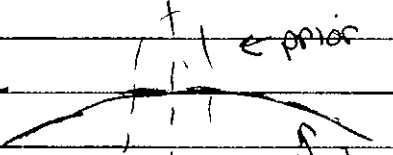
- These factors play off each other: if we add a parameter to a nested model, we expect to gain because $\hat{a}'$ is more informative (it could be $a'=0$ instead)

10/18/19



Now if then $a'=0$ likelihood is $\ll a'=\hat{a}'$ likelihood
$\Rightarrow$ evidence ratio $\gg 1$ and exclusion of this parameter is highly favored.

Unless you put flat prior from near $-\infty$ to near $+\infty$.
But we have a naturalness prior, so $\Delta a'$ restricted.



Now suppose Turn analysis on its head

$$\frac{p(D|M_{k+1},I)}{p(D|M_k,I)} \hat{=} \frac{\int d\vec{a}_k\, p(D|\vec{a}_k,\hat{a}',M_{k+1},I)\, p(\vec{a}_k|M_{k+1},I)\, \delta a' \cdot \frac{1}{p(a'|M_{k+1},I)}}{\int d\vec{a}_k\, p(D|\vec{a}_k,M_k,I)\, p(\vec{a}_k|M_k,I)}$$

← replace dependence on a' because weak; ← normalization

normalization integral, dominated by prior so $\hat{a}'\approx 0$ can be used

But M_{k+1} with $\hat{a}'=0$ is $M_k \Rightarrow$ Bayes ratio $\to 1$ (not decrease)
Same argument for $k+1 \to k+2 \to \ldots$
$\Rightarrow$ we have saturation of a_k's

Summary: naturalness prior cuts down on wasted space in the parameter phase space that might be ruled out by data.

This EFT is a simpler model (in the model selection sense) than the same functional form with unconstrained or only weakly constrained LECs.

Seen in Fig. 8 $\Rightarrow$ see notebook.
Ratio is about 5, so quadratic is moderately more favorable (compare logs)
$\Rightarrow$ returning the prior

Predict: Based on your experience, how does this behavior change if we have more data (higher energy) or more certain data? So depends on data.

10/15/19

Return to notebook to look at calculation of evidence with linear algebra.

- Integrals to calculate are Gaussians in multiple variables: $\vec{a} = (a_0, \ldots, a_k)$ plus $\bar{a}$.

- We can write them with matrices.

E.g. see Lecture 10 and 11 notes,

$$\chi^2 = (Y - A\theta)^T \Sigma^{-1} (Y - A\theta)$$

N data points

where $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ (Data), $A = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^k \\ 1 & x_2 & x_2^2 & \ldots \\ \vdots & & & \\ 1 & x_N & x_N^2 & x_N^k \end{bmatrix}$ (k columns, N rows), $\theta = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix}$, $\Sigma_{N\times N} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \\ & \ddots & \\ & & \sigma_N^2 \end{pmatrix}$

we've taken these = 0 mostly

χ^2_{MLE} when $\hat{\theta} = (A^T \Sigma^{-1} A)^{-1} (A^T \Sigma^{-1} Y)$ (from Lecture 11 notes)

Here we have a couple of options:

i) Use $\int e^{-\frac{1}{2}x^T A x + B^T x} \, dx = \sqrt{\det(2\pi A^{-1})} \, e^{\frac{1}{2} B^T A^{-1} B}$ (completing the square) (different A here, sorry!)

generic square matrix A and vector B

ii) Use conjugacy. See "conjugate prior" entry in Wikipedia.

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{p(D)}$$

if $p(\theta)$ Gaussian (μ_0, σ_0) and $p(D|\theta)$ Gaussian (μ, σ), so is $p(\theta|D)$

$$\tilde{\mu} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}} \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{N} y_i}{\sigma^2} \right) \qquad \tilde{\sigma}^2 = \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \right)^{-1}$$

(check $N \to \infty$, μ_0 and σ_0 are irrelevant?)

$\Rightarrow$ Generalized formula for multivariate Gaussians