6/18/19

Revisit model-selection_I.ipynb

- Modified version model-selection_I_rjf.ipynb

- Recall the model and aspects we can adjust
  - $y_i = x_i \sin x_i + \varepsilon_i$

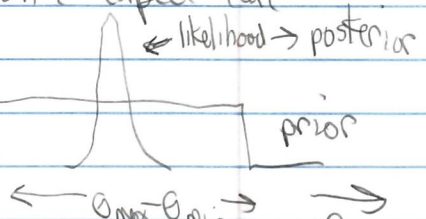
⇒ num_data — no. of data points
x_min, x_max — x_min ≤ x ≤ x_max
sigma — standard deviation for noise at each point

- Look at least squares fit for standard case and note in figures the transition from underfit to good fit to overfit
  - Note the comparison of exact and fit lines
    ⇒ where does it stop changing ⇒ what we want a signal for

- Examine the "anomalous" behavior seen in the exercise session
  - eg. with a small x interval, the evidence apparently favors high orders that are manifestly overfitting.

eg. x_min=0, x_max=1

- Step back a moment and recall why we didn't expect that to happen: the Occam penalty
  - We'll see this in another context, but basic idea is that it is a ratio of volumes



⇒ for each parameter

prior $p(\theta_i | I) = \begin{cases} \frac{1}{\theta_{max} - \theta_{min}} & \text{for } \theta_{min} \le \theta_i \le \theta_{max} \\ 0 & \text{otherwise} \end{cases}$

posterior $p(\theta | D, I) = \frac{1}{(\theta_{max} - \theta_{min})^k} \frac{1}{(\sqrt{2\pi}\sigma_0)^N} e^{-\chi^2/2}$

Laplace's method expand about θ^*

$\Rightarrow \int d^k\theta \; p(\theta | D, I) \doteq \frac{1}{(\theta_{max} - \theta_{min})^k} e^{-\chi^2(\theta^*)/2} \frac{\sqrt{(2\pi)^k}}{\sqrt{\det(\Sigma^{-1})}}$

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

6/18/19

* But what should the width of the prior be??
  - If we take it much larger than the default [-10,10], we recover the desired behavior with the peak moving to small order as the range of data decreases or the size of the error increases.

num_data	x_max	sig0	theta_max	peak	
20	3.	0.1	10 or 100	3	
20	1.	0.1	100	≥8	
"	"	"	1000	1	!
20	3.	0.5	10	2	clear overfitting at higher orders
"	"	"	100	1	
20.	3.	1.0	100	0	
"	"	"	10	2	
10	3.	0.1	10	4 (but also 8)	
"	"	"	100	3	
"	"	"	1000	3	
"	"	"	10000	2-3	

- So we can recover reasonable results, but this seems totally adhoc if we insist on non-informative prior. But it is not here!
  - Is there a criterion for choosing the prior range?

- Minka says to use a unit-invariant prior with a parameter α determined from the data.
  - See reference and application to choosing an underlying polynomial.

- Trotta discusses some possibilities in "Bayes in the Sky" eg. use something called the Fisher information matrix F.

Laplace approximation $\Rightarrow$ $p(D|m) = \mathcal{L}_{max} \frac{|\bar{E}|^{-1/2}}{|P|^{-1/2}} e^{-\frac{1}{2}(\theta^{*T}\Sigma^{-1}\theta^{*} - \bar{\theta}^{T}F\bar{\theta})}$ $\leftarrow$ MLE

with $F = \Sigma^{-1} + P$ and $\bar{\theta} = F^{-1}\Sigma^{-1}\theta^{*}$. $\Rightarrow$ gives volume factor à la Occam

6/18/19

Bottom line: In this example, the evidence is highly sensitive to the prior ⇒ need criterion to determine prior.

Plan for us: Look at 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).
  ⇒ EFT_slides_II.pdf
  - Do classical analogy: multipole moments fist slide 9/49 → 24/49

- 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.

6/18/19

Revisit two model discussion that Christian did
- 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 ⇒ 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_1 is correct."
  - Here M_2 is M_1 with one extra order (one more parameter) eventually

- Apply Bayes' Theorem

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

(cancel)

Bayes factor

we'll take = 1 for our example ⇒ 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 $p(D|M_2,I) \to p(D,\vec{a}_2|M_2,I) \to p(D_2|M_2,\vec{a}_2,I)\,p(\vec{a}_2,M_2|I)$

so integration over the entire parameter space.
⇒ Difficult numerically since likelihoods usually peaked but can have long tails that contribute to integral

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

Simplify: M_{k+1} has an additional parameter $\bar{a}$ and assume priors flat in

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

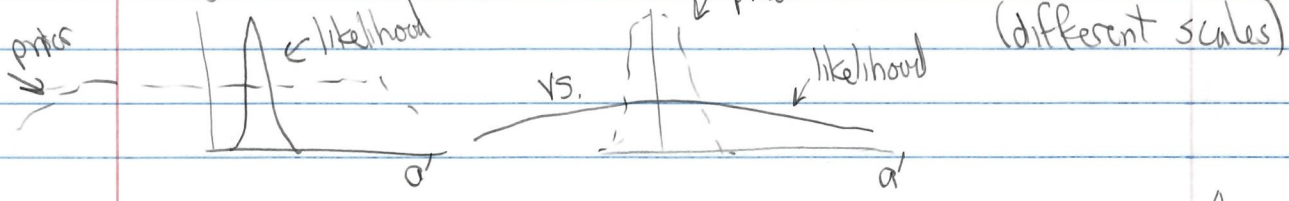
6/18/19

Then $p(\vec{a}_2 | m_{k+1}, I) = p(\vec{a}_1, a' | m_{k+1}, I) = p(\vec{a}_1 | m_{k+1}, I)\, p(a' | m_{k+1}, I)$

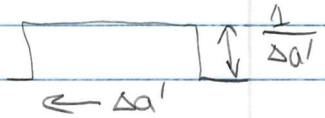
Consider cases...

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 → posterior)

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

($\delta a'$: from integrating over a'; $\hat{a}'$: peak value; denominator: from $p(a'|m_{k+1}, I)$)

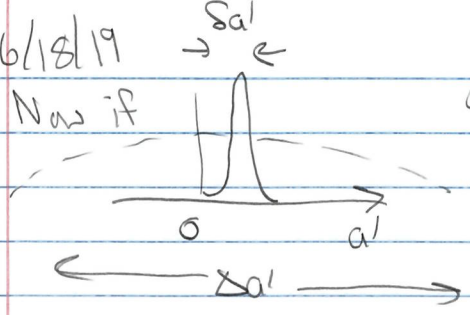
$\Rightarrow$ The ratio of the integrals is the gain in the likelihood from an extra parameter with value $\hat{a}$ (cf. $m_{k+1}(\hat{a}'=0) = m_k$)

- 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 $\delta a' \ll \Delta a'$.

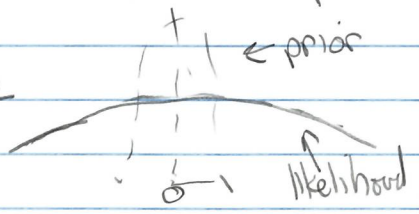
- These factors play off each other: if we add a parameter to a nested model, we expect to gain because $\hat{a}'$ is more information (it could be $a'=0$ instead)

6/18/19



Now if then $a'=0$ likelihood is $\ll a'=\hat{a}'$ likelihood
$\Rightarrow$ evidence ratio $\gg 1$ and inclusion 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.

$$\frac{p(D|m_{k+1},I)}{p(D|m_k,I)} \cong \frac{\int d\vec{a}_k\, p(D|\vec{a}_k,\hat{a}',m_{k+1},I)\,p(\vec{a}_k|m_{k+1},I)\int da'\,p(a'|m_{k+1},I)}{\int d\vec{a}_k\, p(D|\vec{a}_k,m_k,I)\,p(\vec{a}_k|m_k,I)}$$

← replace dependence on a' because weak; ↑ normalization $=1$

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

- 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.

Thus 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.

6/18/19

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 Th1a-7 to Th1a-9

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

N data points

k

where $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ Data

$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & \\ \vdots & & & & \\ 1 & x_N & x_N^2 & \cdots & x_N^k \end{bmatrix}$ N rows

$\theta = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix}$

$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & & \\ & \ddots & \ddots & \\ & \ddots & \ddots & \\ & & & \sigma_N^2 \end{pmatrix}$ $N \times N$

we've taken these $= 0$ mostly

χ^2_{MLE} when $\hat{\theta} = (A^T \Sigma^{-1} A)^{-1} (A^T \Sigma^{-1} Y)$ see Th1a-9 for one demonstration

Here we have a couple of options:

i) Use $\int e^{-\frac{1}{2}x^T A x + B^T x} \, dx = \overbrace{\sqrt{\det(2\pi A^{-1})}}^{\text{complete the square}} \, e^{\frac{1}{2} B^T A^{-1} B}$ (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 with (μ_0, σ_0) and $p(D|\theta)$ Gaussian in (μ, σ), so is $p(\theta|D)$

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

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

$\Rightarrow$ Generalized formula for multivariate Gaussians

6/18/14

Computational possibilities for evidence

Many possible challenges
- likelihood sharply peaked in prior range, but could have long tails with significant contribution to integrals
- likelihood could be multimodal
- posterior may only be significant on thin "sheets" in parameter space (cf. sampling visualization from last week)

- Trotta summary of methods: (old info)

1) Thermodynamic integration → simulated annealing
computational cost depends heavily on dimensionality & parameter space and on details of likelihood function
Cosmological applications → up to 10^7 likelihood evaluations (10^2 times MCMC-based parameter estimation).
⇒ emulator (Thursday)

2) Nested sampling recasts multidimensional evidence integral into one-dimensional integral, easy to evaluate numerically.
⇒ ~10^5 likelihood evaluation.
- multinest is more efficient still.

3) Approximations to the Bayes factor
- If models are nested: ask whether new parameter is supported by data
- Laplace approximation may be good (as we've used) but be careful of priors

problems with not treating priors

- define effective # of parameters. ⇒ BDA3 + Trotta

- AIC, BIC, DIC, WAIC ⇒ BDA3