10/18/19

Physics 8805: Learning from Data: Lecture 16

Notebook for today:

- Evidence_for_model_EFT_coefficients.ipynb

Questions about mini-project IIa?

- MCMC-parallel-tempering.ipynb

Follow-ups

- Approximate calculations of evidence ratios (easier than absolute evidence!) → 86 - 87
- Class exercise with notebook:
  1. What do you expect to happen taking $\bar{a}$ larger? $5 \to 50 \to 500 \to 5000$
     - wide prior, so like least squares ⇒ peak predicted.
  2. Evaluation of two models using evidence is contingent on the data (so different data can give different conclusions).
     - Compare error of .01 → 0.02 - 0.03 with very high $\bar{a}$
     - What about 0.5?
  3. What does the output of chat tell you? What about correlations?

- Look at Laplace's method and comparison to parameter estimation. 81

- Do 88 on the direct evidence calculation

On board: Put $p(D|M_{k+1},I)/p(D|M_k,I)$ from 85 with $\vec{a}_2 \to a', \vec{a}$ (or $\vec{a}_k$) and then the two limiting cases.

90

10/18/19

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).

- Trotta summary of methods: (possibly old info)

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

2) Nested sampling recasts multidimensional evidence integral into one-dimensional integral, easy to evaluate numerically.
⇒ $\sim 10^5$ 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

- define effective # of parameters. → BDA3 + Trotta

- AIC, BIC, DIC, WAIC ⇒ BDA3 for details.
Summary on next page.

problems with not treating priors

10/18/19

Examples of Information Criteria $\Rightarrow$ computationally much easier

AIC: Akaike Information Criteria

- Essentially frequentist as it relies on the likelihood
- Quantity to calculate:

$$-2\log p(D|\hat{\theta}_{MLE}) + 2k$$

~~# of free parameters~~

likelihood, data, maximum likelihood value of parameters

- Compare the result between models
- Has the ingredients of evidence: improved likelihood is balanced by penalty for additional parameters. No priors
- Not well regarded by Bayesians.

BIC: Bayesian Information Criteria

- Gaussian approximation to Bayesian evidence in limit of large amount of data.
- $BIC = -2\log p(D|\hat{\theta}_{MLE}) + k\ln N$ ← # data points

  # fitted parameters
- Assumes Occam razor penalty is negligible.

DIC: Deviance Information Criteria

- replace $\hat{\theta}_{mle}$ by $\hat{\theta}_{Bayes}$ ← maximum of posterior
- use effective # of parameters

  average θ over posterior

$$p_{DIC} = 2\log p(D|\hat{\theta}_{Bayes}) - E[\log p(D|\theta)]$$

- $DIC = -2\log p(D|\hat{\theta}_{Bayes}) + 2p_{DIC}$

WAIC: Widely Applicable Information Criterior.

- favored by BDA-3 as more fully Bayesian
- given samples $s = 1$ to S

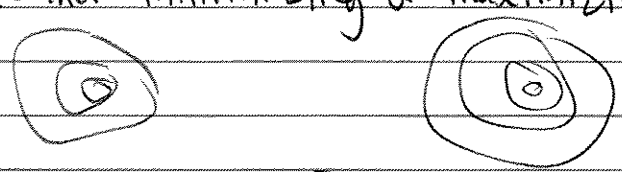
$$\Rightarrow WAIC = 2\sum_{i=1}^{n\,data}\left(\log\frac{1}{S}\sum_{s=1}^{S}p(y_i|\theta^s)\right) - \frac{1}{S}\sum_{s=1}^{S}\log p(D_i|\theta^s)$$

averages over posterior distribution.

10/18/19

Parallel Tempering

- Parallel tempering was particularly introduced to deal with multimodal distributions (see simulation for example)
  - Analogous to the problem of global optimization (either minimizing or maximizing).



  - How do you jump from one region of high posterior density to another when there is a large region of low probability in between?
  - This is a problem for an evidence calculation, because we need to integrate over the entire parameter space.
  - For parameter estimation, it may be sufficient to start walkers in the vicinity of the "best" (highest posterior) part of the posterior.

- Parallel tempering is built on top of an MCMC sampler.
  - The general idea is to simulate N <u>replicas</u> of a system, each at a different temperature.
  - The temperature of a Metropolis-Hastings Markov chain specifies how likely it is to sample from a low-density part of the target distribution.
  - At high temperature, large volumes of phase space are sampled roughly, while low temperature systems have precise sampling in a local region of the parameter space, where they can get stuck in a local energy minimum (meaning a local posterior maximum).
  - Parallel tempering works by letting the systems at different temperatures exchange configurations, which enables the low T system access to a complete set of low-temperature regions.

10/18/19

The temperature comes in analogy to a Boltzmann factor $e^{-E/k_B T} \Rightarrow e^{-\beta T}$ with $\beta \equiv 1/k_B T$.

- Instead of E, we have $-\log p_\beta(x)$, so $p_\beta(x) = e^{-\beta(-\log p(x))} = (p(x))^\beta$

The emcee documentation writes this as

$$\pi_{\underset{1/\beta}{T}} = [l(x)]^{1/T}\, p(x)$$

but we'll write it a bit more conventionally. We want the posterior

$$p(\vec{\theta}\,|\,D, I) = \underset{\text{normalization}}{C}\; \underset{\text{likelihood}}{p(D\,|\,\vec{\theta}, I)}\; \underset{\text{prior}}{p(\theta\,|\,I)}$$

and generalize to the temperature-dependent posterior p_β:

$$p_\beta(\vec{\theta}\,|\,D, I) \propto p(D\,|\,\vec{\theta}, I)^\beta\, p(\theta\,|\,I) \qquad 0 \le \beta \le 1$$

or

$$\log p_\beta(\theta\,|\,D, I) = C + \beta \log p(D\,|\,\vec{\theta}, I) + \log p(\theta\,|\,I)$$

So the desired distribution is $\beta = 1$, and $\beta = 0$ is the prior.

- At large temperature, $\beta = 0$, and we sample the prior, which should encompass the full space.
- As we approach $\beta = 1$ we focus increasingly on where the likelihood is large.

- We will use $\beta \in \{1, \beta_2, \ldots, \beta_N\}$ running in parallel, but swapping members of the chains in such a way that detailed balance is preserved.

10/18/19

- The general sampling strategy is:
  - at intervals, pick a pair of adjacent chains at random: β_i and β_{i+1}
  - propose a swap of their current positions at this time t, namely exchange
    $\vec{\theta}_{t,i}$ and $\vec{\theta}_{t,i+1}$
  - accept this proposal with probability (note there is no $(i+1)$s)

$$r = \min\left\{1, \frac{p_\beta(\vec{\theta}_{t,i+1}|D,\beta_i,I)\, p_\beta(\vec{\theta}_{t,i}|D,\beta_{i+1},I)}{p_\beta(\vec{\theta}_{t,i}|D,\beta_i,I)\, p_\beta(\vec{\theta}_{t,i+1}|D,\beta_{i+1},I)}\right.$$

  - This will preserve detailed balance

- To specify for the sampler:
  - n_s: propose a swap every n_s iterations (which is implemented by drawing a random number $U_1 \sim \text{Uniform}[0,1]$ every iteration and proposing a swap if $U_1 \leq 1/n_s$.
  - The length and spacing of the temperature ladder.

- To calculate the evidence, we can use thermodynamic integration (see Goggans and Chi, AIP Conf. Proc. 707, 59 (2004)).

- Define temperature dependent evidence: $Z(\beta) \equiv \int d\vec{\theta}\, p(D|\vec{\theta},I)^\beta\, p(\vec{\theta}|I)$ so we want $Z(1)$.

  $p(D|\vec{\theta},I)^\beta = e^{\beta \ln p(D|\vec{\theta},I)}$

- $Z(\beta)$ satisfies a differential equation: $\frac{d \ln Z(\beta)}{d\beta} = \frac{1}{Z(\beta)} \int d\vec{\theta}\, (\ln p(D|\vec{\theta},I))\, p(D|\vec{\theta},I)^\beta\, p(\vec{\theta}|I)$

$$\equiv \langle \ln p(D|\vec{\theta},I)\rangle_\beta$$

So we can integrate over β:

$$\ln Z(1) = \cancel{\ln Z(0)}^{\,1,\ \text{normalized}} + \int_0^1 d\beta\, \langle \ln p(D|\vec{\theta},I)\rangle_\beta$$

average of the log likelihood at temperature β.

$\Rightarrow$ estimate from emcee samples by computing the average of $p(D|\vec{\theta},I)$ within each chain and then evaluating the integral from a quadrature formula (e.g. Simpson's rule).

10/18/19

An example of parallel tempering using Emcee (2.2.1 or before) is given in MCMC-parallel-tempering.ipynb.

Comments:

- First we set up a bi-modal distribution (it was originally intended to ~~be~~ a surprise, so the code was hidden). Just two Gaussians with different amplitudes.

- A first sampling try with an ordinary Metropolis-Hastings (mH) sampler fails by only finding one modes.

- But then if one checks two chains, separate modes are found and the chains do not mix. With many walkers one might find multiple modes, but the relative normalizations would not work because each walker would not explore the space.

- The PTSampler is no longer in the Emcee distribution, but is in the ptemcee as a "drop-in replacement".
  - The setup includes a temperature grid chosen so that the integration via quadrature over temperature for the evidence has a finer grid at low temperatures for greater accuracy.
  - Note the range at different temperatures and how the multimodal structure emerges.