Learning from data: Why Bayes is Better

Christian Forssén

Department of Physics, Chalmers University of Technology, Sweden

Sep 23, 2019

1 Why Bayes is Better (I)

1.1 Quotes from one pioneering and one renaissance Bayesian authority

Laplace:

> *Probability theory is nothing but common sense reduced to calculation."*

Sivia

> *Bayesian inference probabilities are a measure of our state of knowledge about nature, not a measure of nature itself."*

1.2 Advantages of the Bayesian approach

1. Provides an elegantly simple and rational approach for answering, in an optimal way, any scientific question for a given state of information. This contrasts to the recipe or cookbook approach of conventional statistical analysis. The procedure is well-defined:
   - Clearly state your question and prior information.
   - Apply the sum and product rules. The starting point is always Bayes' theorem.
2. For some problems, a Bayesian analysis may simply lead to a familiar statistic. Even in this situation it often provides a powerful new insight concerning the interpretation of the statistic.

3. Incorporates relevant prior (e.g., known signal model or known theory model expansion) information through Bayes' theorem. This is one of the great strengths of Bayesian analysis.

    - For data with a small signal-to-noise ratio, a Bayesian analysis can frequently yield many orders of magnitude improvement in model parameter estimation, through the incorporation of relevant prior information about the signal model.

4. Provides a way of eliminating nuisance parameters through marginalization. For some problems, the marginalization can be performed analytically, permitting certain calculations to become computationally tractable.

5. Provides a way for incorporating the effects of systematic errors arising from both the measurement operation and theoretical model predictions.

6. Calculates probability of hypothesis directly: $p(H_i|D, I)$.

7. Provides a more powerful way of assessing competing theories at the forefront of science by automatically quantifying Occam's razor.

The Bayesian quantitative Occam's razor can also save a lot of time that might otherwise be spent chasing noise artifacts that masquerade as possible detections of real phenomena.

Occam's razor. Occam's razor is a principle attributed to the medieval philosopher William of Occam (or Ockham). The principle states that one should not make more assumptions than the minimum needed. It underlies all scientific modeling and theory building. It cautions us to choose from a set of otherwise equivalent models of a given phenomenon the simplest one. In any given model, Occam's razor helps us to shave offthose variables that are not really needed to explain the phenomenon. It was previously thought to be only a qualitative principle.

1.3 Nuisance parameters (I)

See demonstration notebook: A Bayesian Billiard game

1.4 Nuisance parameters (II): marginal distributions

Assume that we have a model with two parameters, $\theta 0, \theta_1$, although only one of them (say θ_1) is of physical relevance (the other one is a nuisance parameter). Through a Bayesian data analysis we have the joint, posterior pdf

$$p(\theta_0, \theta_1|D, I).$$

The marginal posterior pdf $p(\theta_1|D, I)$ is obtained via marginalization

$$p(\theta_1|D, I) = \int p(\theta_0, \theta_1|D, I)d\theta_0.$$

Figur 1: Did the Leprechaun drink your wine, or is there a simpler explanation?

Assume that we have N samples from the joint pdf. This might be the Markov Chain from an MCMC sampler: $\{(\theta_0, \theta_1)_i\}_{i=0}^{N-1}$. Then the marginal distribution of θ_1 will be given by the same chain by simply ignoring the θ_0 column, i.e., $\{\theta_{1,i}\}_{i=0}^{N-1}$.

See the interactive demos created by Chi Feng for an illustration of this: The Markov-chain Monte Carlo Interactive Gallery.

1.5 Error propagation (I): marginalization

The Bayesian approach offers a straight-forward approach for dealing with (known) systematic uncertainties; namely by marginalization. Let us demonstrate this with an example

Inferring galactic distances with an imprecise knowledge of the Hubble constant The Hubble constant acts as a galactic ruler as it is used to measure astronomical distances according to $v = H_0 x$. An error in this ruler will therefore correspond to a systematic uncertainty in such measurements.

Here we use marginalization to obtain the desired posterior pdf $p(x|D, I)$ from the joint distribution of $p(x, H_0|D, I)$

$$p(x|D, I) = \int_{-\infty}^{\infty} dH_0 p(x, H_0|D, I).$$

Using Bayes' rule: $p(x, H_0|D, I) \propto p(D|x, H_0, I)p(x, H_0|I)$, the product rule: $p(x, H_0|I) = p(H_0|x, I)p(x|I)$, and the fact that H_0 is independent of x: $p(H_0|x, I) = p(H_0|I)$, we find that

$$p(x|D, I) \propto p(x|I) \int dH_0 p(H_0|I)p(D|x, H_0, I),$$

which means that we have expressed the quantity that we want (the posterior of x) in terms of quantities that we know.

Assume that the pdf $p(H_0|I)$ is known via its N samples $\{H_i\}_{i=0}^{N-1}$ generated by the MCMC sampler.

This means that we can approximate

$$p(x|D, I) \propto \int dH_0 p(H_0|I)p(D|x, H_0, I) \approx \frac{1}{N} \sum_{i=1}^{N} p(D|x, H_i, I)$$

where we have used a uniform prior for the distance $p(x|I) \propto 1$.

1.6 Error propagation (II): prior information

Example 3.6.2 in Sivia.

- Consider a Bragg peak amplitude that is proportional to the square of a complex structure function: $A = f^2$.
- The amplitude is measured with an uncertainty $A = A_0 \pm \sigma_A$ from a least-squares fit to experimental data.
- What is $f = f_0 \pm \sigma_f$?

See notes and demonstration notebook.