Data don't speak for themselves
a Bayesian course

JOÃO P. FARIA
Institute of Astrophysics and Space Sciences, Porto
January 18, 2016

Abstract

Bayesian statistics is rising in popularity in the astrophysical literature. It is no longer a debate: "work in Bayesian statistics now focuses on applications, computations, and models. Philosophical debates [...] are fading to the background" [1]. This is happening for two main reasons: faster computers and more complex models. In order to keep up, it is important to understand the fundamentals of Bayesian statistics, but it is as important to know how to deal with data analysis applications. In this course I want to provide a brief introduction to advanced concepts in Bayesian statistics. Emphasis will be on *intuition* and *computation*. No coin tossing, only real applications that relate to our day-to-day problems.

Contents

A (tiny) primer on philosophy

While preparing for this course, I asked people what it was they wanted to learn about. Much to my dismay, no one answered with "the philosophy of Bayesian statistics". Most want to learn about the "theory" and "applications" (read MCMC). They want to know how they can *use* Bayesian statistics in their work.

Fair enough. But bypassing phylosophy means making assumptions. Therefore, I shall introduce here Jaynes' *robot*, an imaginary being whose brain is designed by us, so that it reasons according to certain definite rules [2]. These rules will simply be stated, not derived and not defended. They will be taken as being true. Then, everything that follows logically from them, everything the robot does, will be true.

The robot is objective[1] in its actions, and takes as input *(i)* data and *(ii)* the subjective knowledge we feed into it.

1 The basics of probability theory

Here we present the rules of the game. I follow very closely (too closely) the presentation in [4]. Using Bayesian statistics to analyse data (some might call it doing probabilistic inference) means working with likelihoods, prior probabilities, posterior probabilities and marginalization of nuisance parameters. All this will be explained later but the options we have when working with these mathematical objects are strongly constrained by the rules of probability calculus.

> *On voit, par cet Essai, que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul;*
>
> – Pierre-Simon Laplace

If we have a continuous parameter a, and a probability distribution function $p(a)$ for a, it must obey the normalization condition

$$1 = \int p(a)\, da \tag{1}$$

where the integral is limited by the domain of a.

Often times we will have more than one parameter. Even if we *condition* $p(a)$ on some particular value of another parameter b, that is, we ask for $p(a|b)$ (read "the pdf for a given b"), it must obey the same normalization

$$1 = \int p(a|b)\, da \tag{2}$$

[1] Which is not saying much... [3].

If we have a probability distribution for two things, $p(a, b)$ (read "the pdf for a and b"), you can always factorize it into two distributions, one for a, and one for b given a or the other way around:

$$p(a, b) = p(a)\, p(b|a) \tag{3}$$
$$p(a, b) = p(b)\, p(a|b) \tag{4}$$

These two factorisations together lead[2] to Bayes' theorem:

$$p(a|b) = \frac{p(b|a)\, p(a)}{p(b)}. \tag{5}$$

Conditional probabilities factor just the same as unconditional ones

$$p(a, b|c) = p(a|c)\, p(b|a, c) \tag{6}$$
$$p(a, b|c) = p(b|c)\, p(a|b, c) \tag{7}$$
$$p(a|b, c) = \frac{p(b|a, c)\, p(a|c)}{p(b|c)} \tag{8}$$

where we just carried the condition c through all the terms.

You can integrate out or *marginalize* variables you want to get rid of (or *not* infer) by integrals like

$$p(a|c) = \int p(a, b|c)\, db \tag{9}$$
$$p(a|c) = \int p(a|b, c)\, p(b|c)\, db \tag{10}$$

where the second is a factorized version of the first. Remember that, since b can be a very high-dimensional mathematical object (a set of parameters), integrals like these can be extremely difficult to calculate in practice.

1.1 Some familiarity

Let us write some of the preceding equations with more familiar terms[3]. We have data D and a set of parameters θ that we are interested in learning about. In all our analyses we condition on information $\mathcal{I}$ that we have about the world. Then we can write Eq. (8) as

$$p(\theta|D, \mathcal{I}) = \frac{p(\theta|\mathcal{I})\, p(D|\theta, \mathcal{I})}{p(D|\mathcal{I})} \tag{11}$$

The terms in this equation are usually called

[2] Bayes' theorem is a consequence of the previous equations, it is not assumed as a rule. That's why it's called a theorem, actually.

[3] Only if you have heard about Bayesian statistics, otherwise just different terms.

$p(\theta|D,\mathcal{I})$ the posterior distribution

$p(\theta|\mathcal{I})$ the prior distribution

$p(D|\theta,\mathcal{I})$ the likelihood

$p(D|\mathcal{I})$ the evidence, sometimes denoted with a $\mathcal{Z}$.

but calling them this may hide something important (see Section 2.2).

From Eq. (2) we can derive[4] that

$$p(D|\mathcal{I}) = \mathcal{Z} = \int p(\theta|\mathcal{I})\, p(D|\theta,\mathcal{I})\, d\theta \tag{12}$$

Because $\mathcal{Z}$ does not depend on θ, you might see many times Eq. (11) written as

$$p(\theta|D,\mathcal{I}) \propto p(\theta|\mathcal{I})\, p(D|\theta,\mathcal{I}), \tag{13}$$

so **the posterior is proportional to the likelihood times the prior**. Nevertheless, try to stick with Eq. (11); $\mathcal{Z}$ contains in it a big deal of information.

think about it

Eq. (13) means that the likelihood and the prior can be defined up to an arbitrary multiplicative constant (i.e. not a function of θ).

For example, imagine we are interested in comparing two models M_1 and M_2 which have parameters θ_1 and θ_2, respectively. We still only have data D. Bayes' theorem works the same:

$$p(M_i|D,\mathcal{I}) = \frac{p(M_i|\mathcal{I})\, p(D|M_i,\mathcal{I})}{p(D|\mathcal{I})} \tag{14}$$

and the ratio of model probabilities is

$$\frac{p(M_1|D,\mathcal{I})}{p(M_2|D,\mathcal{I})} = \frac{p(M_1|\mathcal{I})\, p(D|M_1,\mathcal{I})}{p(M_2|\mathcal{I})\, p(D|M_2,\mathcal{I})} \tag{15}$$

$$= \frac{p(M_1|\mathcal{I}) \int p(D,\theta_1|M_1,\mathcal{I})\, d\theta_1}{p(M_2|\mathcal{I}) \int p(D,\theta_2|M_2,\mathcal{I})\, d\theta_2} \tag{16}$$

$$= \frac{p(M_1|\mathcal{I}) \int p(\theta_1|M_1,\mathcal{I})\, p(D|\theta_1,M_1,\mathcal{I})\, d\theta_1}{p(M_2|\mathcal{I}) \int p(\theta_2|M_2,\mathcal{I})\, p(D|\theta_2,M_2,\mathcal{I})\, d\theta_2} \quad . \tag{17}$$

See how the evidence (of both models) turns out to be important!
The term

$$\frac{p(M_1|\mathcal{I})}{p(M_2|\mathcal{I})}$$

[4] Multiply both sides of Eq. (8) by $p(D|\mathcal{I})$ and integrate over θ, noting that $p(D|\mathcal{I})$ does not depend on θ and that the posterior obeys Eq. (2).

is the ratio of prior probabilities for the two models. If we believe they are equally likely at the start, then this is equal to 1.

think about it

θ_1 and θ_2 can be any set of parameters and, in particular, they can have different sizes. Say model M_1 has 2 parameters $\theta_1 = (a, b)$ and model M_2 only has one parameter $\theta_2 = (c)$. Then the integrals in Eq. (17) are of different dimensionality. That's fine, probability theory doesn't care. Enough of that "divide by the degrees of freedom" nonsense! [5]

2 Assigning probabilities

Now that we have a set of rules with which we can manipulate our distributions, it is useful to learn how we can assign values to all these terms and actually start calculating things.

2.1 The prior

2.2 The likelihood

The likelihood is a prior. I will repeat so you know this is not a typo: the likelihood is a prior. The term $p(D|\theta, I)$ represents your beliefs on what the data will be like, given parameters θ and information I. Therefore, the likelihood encodes prior information. Following [6], the likelihood is *not*

- the process that generated the data
- the pdf that your data kind of looks like when you plot it in a histogram

It is, instead, what we usually call *the model*. And the model is nothing else than a set of assumptions about the data and how they relate to the parameters. **The likelihood provides a (*the*) connection from θ to D.**

3 MCMC

Markov chain Monte Carlo (MCMC) is the workhorse of Bayesian statistics. But how does it work? And what does it actually do? We will write the code for our own MCMC in the practical classes. If that is enough for you to *understand* it, skip the next section. But if you want to know *why* it works, carry on reading.

3.1 A basic MCMC implementation

We will only talk about the Metropolis-Hastings algorithm to do MCMC. It is probably not the simplest, but it is definitely the most often used.
The following is adapted slightly from Wikipedia:

> Let $f(x)$ be a function that is proportional to the desired probability distribution $P(x)$ (a.k.a. the target distribution).
>
> **Initialization:**
> Choose an arbitrary point x_0 to be the first sample, and choose an arbitrary probability density $Q(x|y)$ which suggests a candidate for the next sample value x, given the previous sample value y. A usual choice is to let $Q(x|y)$ be a Gaussian distribution centered at y. The function Q is referred to as the proposal density or jumping distribution.
>
> **For each iteration t:**
>
> - Generate a candidate x' for the next sample by picking from the distribution $Q(x'|x_t)$.
> - Calculate the acceptance ratio $\alpha = f(x')/f(x_t)$. Because f is proportional to P, we have that $\alpha = f(x')/f(x_t) = P(x')/P(x_t)$.
> - If $\alpha \geq 1$, then the candidate is more likely than x_t: automatically accept the candidate by setting $x_{t+1} = x'$.
>   Otherwise, accept the candidate with probability α;
>   if the candidate is rejected, set $x_{t+1} = x_t$, instead.

Let's see this in `Python` code

```python
import numpy as np
from scipy.stats import norm

# number of iterations to run MCMC
niter = 500

# posterior distribution (the distribution we want to sample from)
p = lambda x:norm.pdf(x)

# starting point
x = [0.5]

```

```
for step in range(niter):
    # propose a step
    d = np.random.randn()

    # calculate the ratio of posterior probabilities
    alpha = p(x[-1]+d) / p(x[-1])

    if alpha > 1.:
        x.append(x[-1]+d)      # accept the new step
    else:
        u = np.random.uniform()
        if ratio > u:
            x.append(x[-1]+d)  # accept the new step
        else:
            x.append(x[-1])    # reject the new step (stay where we are)
```

This statement requires citation [**?**]; this one does too [**?**]. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean dictum lacus sem, ut varius ante dignissim ac. Sed a mi quis lectus feugiat aliquam. Nunc sed vulputate velit. Sed commodo metus vel felis semper, quis rutrum odio vulputate. Donec a elit porttitor, facilisis nisl sit amet, dignissim arcu. Vivamus accumsan pellentesque nulla at euismod. Duis porta rutrum sem, eu facilisis mi varius sed. Suspendisse potenti. Mauris rhoncus neque nisi, ut laoreet augue pretium luctus. Vestibulum sit amet luctus sem, luctus ultrices leo. Aenean vitae sem leo.

Nullam semper quam at ante convallis posuere. Ut faucibus tellus ac massa luctus consectetur. Nulla pellentesque tortor et aliquam vehicula. Maecenas imperdiet euismod enim ut pharetra. Suspendisse pulvinar sapien vitae placerat pellentesque. Nulla facilisi. Aenean vitae nunc venenatis, vehicula neque in, congue ligula.

Pellentesque quis neque fringilla, varius ligula quis, malesuada dolor. Aenean malesuada urna porta, condimentum nisl sed, scelerisque nisi. Suspendisse ac orci quis massa porta dignissim. Morbi sollicitudin, felis eget tristique laoreet, ante lacus pretium lacus, nec ornare sem lorem a velit. Pellentesque eu erat congue, ullamcorper ante ut, tristique turpis. Nam sodales mi sed nisl tincidunt vestibulum. Interdum et malesuada fames ac ante ipsum primis in faucibus.

Section Name

Cras gravida, est vel interdum euismod, tortor mi lobortis mi, quis adipiscing elit lacus ut orci. Phasellus nec fringilla nisi, ut vestibulum neque. Aenean non risus eu nunc accumsan condimentum at sed ipsum. Aliquam fringilla non diam sed varius. Suspendisse tellus felis, hendrerit non bibendum ut, adipiscing vitae diam. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla lobortis purus eget nisl scelerisque, commodo rhoncus lacus porta. Vestibulum vitae turpis tincidunt, varius dolor in, dictum lectus. Aenean ac ornare augue, ac facilisis purus. Sed leo lorem, molestie sit amet fermentum id, suscipit ut sem. Vestibulum orci arcu, vehicula sed tortor id, ornare dapibus lorem. Praesent aliquet iaculis lacus nec fermentum. Morbi eleifend blandit dolor, pharetra hendrerit neque ornare vel. Nulla ornare, nisl eget imperdiet ornare, libero enim interdum mi, ut lobortis quam velit bibendum nibh.

Figure 1: Fish

Morbi tempor congue porta. Proin semper, leo vitae faucibus dictum, metus mauris lacinia lorem, ac congue leo felis eu turpis. Sed nec nunc pellentesque, gravida eros at, porttitor ipsum. Praesent consequat urna a lacus lobortis ultrices eget ac metus. In tempus hendrerit rhoncus. Mauris dignissim turpis id sollicitudin lacinia. Praesent libero tellus, fringilla nec ullamcorper at, ultrices id nulla. Phasellus placerat a tellus a malesuada.

Conclusion

Fusce in nibh augue. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. In dictum accumsan sapien, ut hendrerit nisi. Phasellus ut nulla mauris. Phasellus sagittis nec odio sed posuere. Vestibulum porttitor dolor quis suscipit bibendum. Mauris risus lectus, cursus vitae hendrerit posuere, congue ac est. Suspendisse commodo eu eros non cursus. Mauris ultrices venenatis dolor, sed aliquet odio tempor pellentesque. Duis ultricies, mauris id lobortis vulputate, tellus turpis eleifend elit, in gravida leo tortor ultricies est. Maecenas vitae ipsum at dui sodales condimentum a quis dui. Nam mi sapien, lobortis ac blandit eget, dignissim quis nunc.

1. First numbered list item

2. Second numbered list item

Donec luctus tincidunt mauris, non ultrices ligula aliquam id. Sed varius, magna a faucibus congue, arcu tellus pellentesque nisl, vel laoreet magna eros et magna. Vivamus lobortis elit eu dignissim ultrices. Fusce erat nulla, ornare at dolor quis, rhoncus venenatis velit. Donec sed elit mi. Sed semper tellus a convallis viverra. Maecenas mi lorem, placerat sit amet sem quis, adipiscing tincidunt turpis. Cras a urna et tellus dictum eleifend. Fusce dignissim lectus risus, in bibendum tortor lacinia interdum.

References

1. Andrew Gelman. *Bayesian data analysis*. Chapman & Hall/CRC texts in statistical science. CRC Press, Boca Raton, third edition edition, 2014.

2. E. T. Jaynes and G. Larry Bretthorst. *Probability theory: the logic of science*. Cambridge University Press, Cambridge, UK ; New York, NY, 2003.

3. Andrew Gelman and Christian Hennig. Beyond subjective and objective in statistics. *ArXiv150805453 Stat*, August 2015.

4. David W. Hogg. Data analysis recipes: Probability calculus for inference. *ArXiv e-prints*, 1205:4446, May 2012.

5. Rene Andrae, Tim Schulze-Hartung, and Peter Melchior. Dos and don'ts of reduced chi-squared. *ArXiv10123754 Astro-Ph Physicsphysics Stat*, December 2010.

6. B. J. Brewer. The prior isn't the only prior, 2013.