A primer on Bayesian Hierarchical modelling

Gianluca Rossi

January 18, 2017

Some art courtesy of the University of Cambridge

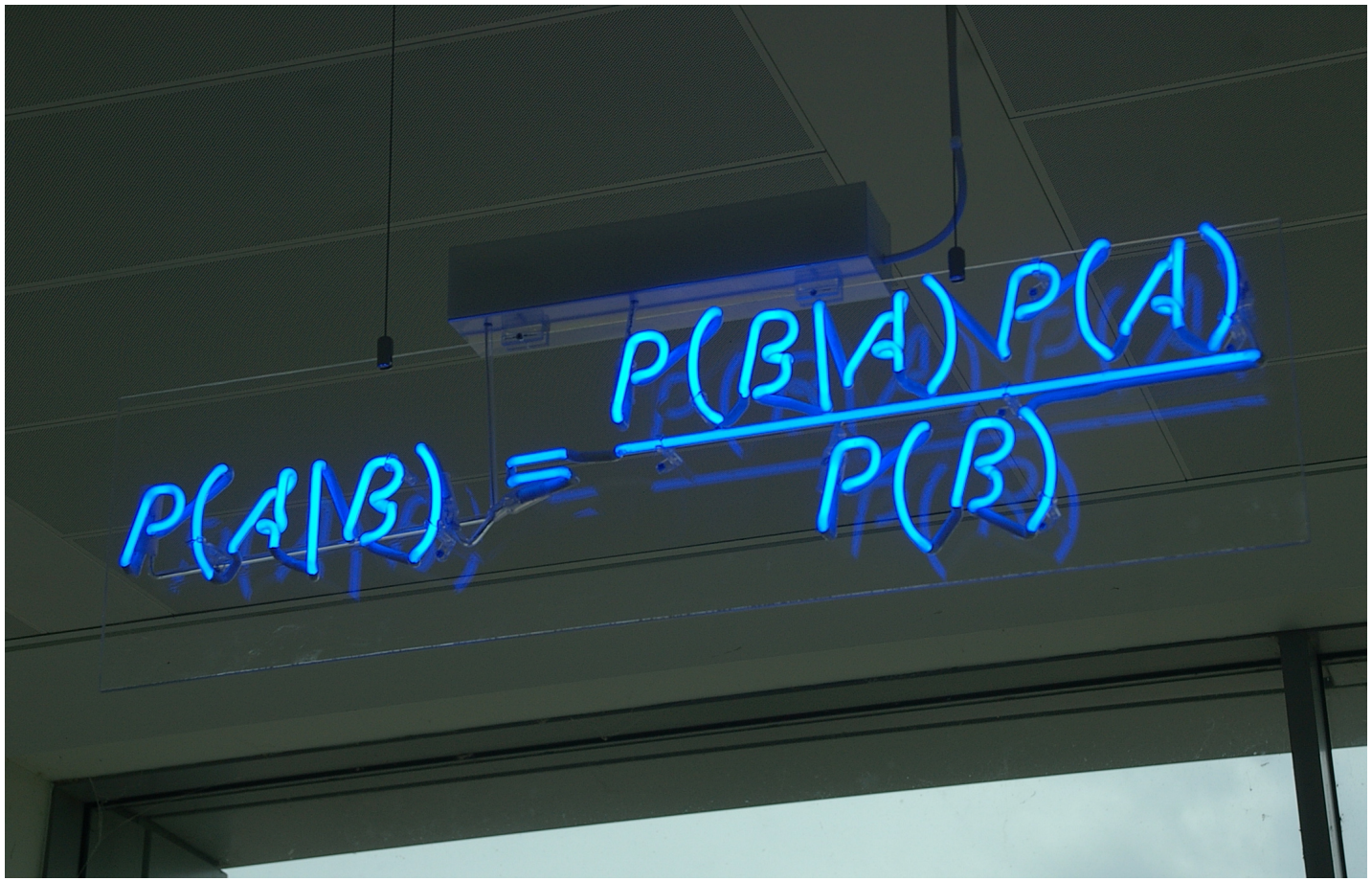

Figure: Bayes' theorem spelt out in blue neon at the offices of Autonomy in Cambridge. (credit: Mattbuck, Wikipedia)

The exact formulation

Assuming we are working in a discrete space, the Bayes' Theorem can be written as:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{1}$$

$$= \frac{P(D|\theta)P(\theta)}{\sum_{\theta^\star} P(D|\theta^\star)P(\theta^\star)} \tag{2}$$

$$\propto P(D|\theta)P(\theta) \tag{3}$$

The three components of the equation are:

- $P(\theta|D)$, the posterior
- $P(D|\theta)$, the likelihood function
- $P(D)$, the normalising function

The likelihood function

The likelihood is a function of the parameters of a statistical model given data. In other words, the likelihood describes the probability of observing the data given the parameters of the statistical model.

The prior distribution

The prior is a probabilistic formulation of our beliefs before collecting data.

- Priors can be weak, moderate or strong, depending on how concentrated the probability density is compared to the problem space
- In hierarchical models, priors can be implicit, thus determined by higher level's parameters
- It's important to investigate priors when working with complex models because these could strongly impact the final results (this is not *p-hacking* or *The Garden of Forking Paths*)

The posterior distribution

The posterior distribution is the updated belief after collecting new data.

- Could have the same form of the prior (*conjungate* property), but this is not very interesting when dealing with complex models or using Markov Chain Monte Carlo processes (MCMC)

The main reasons behind the success of Bayesian Inference are:

- It allows to take into account prior knowledge, when this is relevant
  - Speed-up convergence
  - Attribute non-zero probability to yet unobserved events (contrary to frequentist approach)
- Move beyond single point estimation
- Estimate probability in a joint parameter space

Let's assume we are trying to estimate the batting performance of baseball players. We are interested on predicting the Batting Average (AVG) for 947 players, based on the 2012 MLB season statistics.

Our model is:

$$z_s \sim Binomial(N_s, \theta_s) \quad \text{where } s \in [1, S] \tag{4}$$

$$\theta_s \sim Beta(\alpha, \beta) \quad \text{where} \quad \alpha = 1 \text{ and } \beta = 1 \tag{5}$$

- The likelihood is a Binomial function
- The prior is a Beta probability distribution (Uniform)
- We interested in computing the posterior distribution for the Batting Average (also abbreviated AVG, θ) of $S = 947$ players

Since Binomial likelihood and Beta prior are conjungate, we can derive a solution analytically, which is very convenient to understand how everything fits within the Bayes' Theorem framework.

Omitting the indices for clarity, we have that:

$$P(\theta|z,N) = \frac{P(z,N|\theta)P(\theta)}{P(z,N)} \tag{6}$$

$$= \frac{\left(\frac{N!}{z!(N-z)!}\right)\theta^z(1-\theta)^{(N-z)}\frac{\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}}{B(\alpha,\beta)}}{P(z,N)} \tag{7}$$

$$\propto \frac{\theta^z(1-\theta)^{(N-z)}\frac{\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}}{B(\alpha,\beta)}}{P(z,N)} \tag{8}$$

$$\propto \frac{\theta^{((z+\alpha)-1)}(1-\theta)^{((N-z+\beta)-1)}}{B(\alpha,\beta)P(z,N)} \tag{9}$$

$$= \frac{\theta^{((z+\alpha)-1)}(1-\theta)^{((N-z+\beta)-1)}}{B(z+\alpha,N-z+\beta)} \tag{10}$$

Notes:

- (7) Plugging in the definition of Binomial and Beta distributions
- For (10) to be a probability distribution, as it must be, the denominator must be the normalising factor for the corresponding Beta distribution, which is $B(z+\alpha, N-z+\beta)$

See Stan model in *baseball.ipynb* notebook

A fundamental question, often underestimated, when doing Bayesian Data Analysis is:

Data sensitivity

How does Bayes' Theorem weight more recent data?

Starting from the Bayes' Theorem equation we can solve this problem very easily:

$$P(\theta|D', D) = \frac{P(D', D|\theta)P(\theta)}{\sum_{\theta^\star} P(D', D|\theta^\star)P(\theta^\star)} \tag{11}$$

$$= \frac{P(D'|\theta)P(D|\theta)P(\theta)}{\sum_{\theta^\star} P(D'|\theta^\star)P(D|\theta^\star)P(\theta^\star)} \tag{12}$$

$$= \frac{P(D|\theta)P(D'|\theta)P(\theta)}{\sum_{\theta^\star} P(D|\theta^\star)P(D'|\theta^\star)P(\theta^\star)} \tag{13}$$

$$= P(\theta|D, D') \tag{14}$$

It's important to understand that in the Bayes' Theorem the posterior distribution is invariant from the ordering of the data. More recent data is weighted the same as old data. This is because Bayes' Theorem assumes a *stationary scenario*.
To give more weight to more recent data we need to modify the likelihood function to generate non independent data.

Hierarchical modelling means expressing the Bayes' Theorem as dependencies between parameters instead of an expression about the joint parameter space.

But what does this mean?

Let's improve on our previous example. We are still dealing with baseball data and want to estimate players' performance.
This time however we want to have different priors based on the role of the player (*pitcher*, *1st base*, etc. . .). In this formulation, performance of players with a similar role will influence the final prediction for the player.

Differently from the previous model specification we will re-parametrise the Beta prior to use mode (ω) and concentration (κ). In this way we can give hyper-priors to the parameters of the Beta prior.

$$z_s \sim Binomial(N_s, \theta_s) \quad \text{where } s \in [1, S] \tag{15}$$

$$\theta_s \sim Beta(\omega_c(\kappa_c - 2) + 1, (1 - \omega_c)(\kappa_c - 2) + 1) \tag{16}$$

$$\omega_c \sim \dots \tag{17}$$

$$\kappa_c \sim \dots \tag{18}$$

Now, let's plug this into the Bayes' Theorem framework. Again, to simplify the notation we are going to omit the indices.

$$P(\theta, \omega, \kappa|D) = \frac{P(D|\theta, \omega, \kappa)P(\theta, \omega, \kappa)}{P(D)} \tag{19}$$

$$= \frac{P(D|\theta)P(\theta, \omega, \kappa)}{P(D)} \tag{20}$$

$$= \frac{P(D|\theta)P(\theta|\omega, \kappa)P(\omega, \kappa)}{P(D)} \tag{21}$$

We made the following steps:

- (20) because likelihood doesn't depend on ω and κ
- (21) because $P(\theta|\omega, \kappa) = \frac{P(\theta,\omega,\kappa)}{P(\omega,\kappa)}$, thus $P(\theta, \omega, \kappa) = P(\theta|\omega, \kappa)P(\omega, \kappa)$

When dealing with a continuous problem space (21) becomes:

$$P(\theta_1, \dots m|D) = \frac{P(D|\theta_1, \dots m)P(\theta_1, \dots m)}{P(D)} \tag{22}$$

$$= \frac{P(D|\theta_1, \dots m)P(\theta_1, \dots m)}{\sum_m \int \mathrm{d}\theta_m P(D|\theta_1, \dots m)P(\theta_1, \dots m)} \tag{23}$$

$$= \frac{\prod P_m(D|\theta_m, m)P_m(\theta_m|m)P(m)}{\sum_m \int \mathrm{d}\theta_m \prod_m P_m(D|\theta_m, m)P_m(\theta_m|m)P(m)} \tag{24}$$

See Stan model in *baseball.ipynb* notebook