

Lecture 9. Bayesian Inference - updating priors¹

Igor Rychlik

Chalmers

Department of Mathematical Sciences

Probability, Statistics and Risk, MVE300 • Chalmers • May 2011

¹Bayesian statistics is a general methodology to analyse and draw conclusions from data.

$P = P(\text{accidents happen in period } t) = 1 - e^{-\lambda_A P(B) t} \approx \lambda_A P(B) t$,
if probability P is small. Hence Two problems of interest in risk analysis:

- ▶ The first one will deal with the estimation of a probability $p_B = P(B)$, say, of some event B , for example the probability of failure of some system. In figure $B = B_1 \cup B_2$, $B_1 \cap B_2 = \emptyset$
- ▶ The second one is estimation of the probability that at least once an event A occurs in a time period of length t . The problem reduces itself to estimation of the intensity λ_A of A .

The parameters p_B and λ_A are unknown.

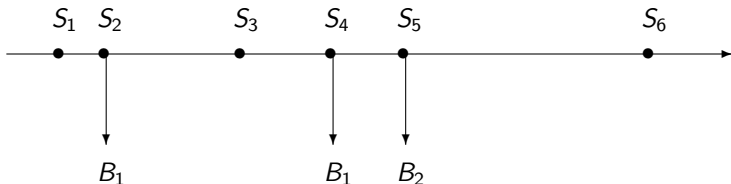


Figure: Events A at times S_i with related scenarios B_j .

Odds for parameters

Let θ denote the unknown value of p_B , λ_A or any other quantity.

Introduce odds q_θ , which for any pair θ_1, θ_2 represents our belief which of θ_1 or θ_2 is more likely to be the unknown value of θ , *i.e.* $q_{\theta_1} : q_{\theta_2}$ are odds for the alternatives $A_1 = “\theta = \theta_1”$ against $A_2 = “\theta = \theta_2”$.

We require that q_θ integrates to one and hence $f(\theta) = q_\theta$ is a probability density function representing our belief about the value of θ . The random variable Θ having the pdf serves as a mathematical model for uncertainty in the value of θ .

Prior odds - posterior odds

Let θ be the unknown parameter ($\theta = p_B$, $\theta = \lambda_A$), while Θ denotes any of the variables P or Λ . Since θ is unknown, it is seen as a value taken by a random variable Θ with pdf $f(\theta)$.

If $f(\theta)$ is chosen on basis of experience without including observations of outcomes of an experiment then the density $f(\theta)$ is called a *prior density* and denoted by $f^{\text{prior}}(\theta)$.

Since our knowledge may change with time (especially if we observe some outcomes of the experiment) influencing our opinions about the values of parameter θ . This leads to new odds - density $f(\theta)$. The modified density $f(\theta)$ will be called the *posterior density* and denoted by $f^{\text{post}}(\theta)$.

The method to update $f(\theta)$ is

$$f^{\text{post}}(\theta) = cL(\theta) f^{\text{prior}}(\theta)$$

How to find likelihood function $L(\theta)$ will be discussed later on.

Predictive probability

Suppose $f(p)$ has been selected and denote by P a random variable having pdf $f(p)$. A plot of $f(p)$ is an illustrative measure of how likely the different values of p_B are.

If only one value of the probability is needed, the Bayesian methodology proposes to use the so-called **predictive probability** which is simply the mean of P :

$$p^{\text{pred}}(B) = E[P] = \int p f(p) dp.$$

The predictive probability measures the likelihood that B occurs in future. It combines two sources of uncertainty: the unpredictability whether B will be true in a future accident and the uncertainty in the value of probability p_B .

Example 6.1

$$P(A \cap B) = P(\text{accidents in period } t) = 1 - e^{-\lambda_A P(B)t} \approx \lambda_A P(B)t,$$

if probability $P(A \cap B)$ is small.

The predictive probabilities

$$\begin{aligned} P^{\text{pred}}(A) &= E[P(A)] = \int (1 - \exp(-\lambda t)) f_{\Lambda}(\lambda) d\lambda \\ &\approx \int t \lambda f_{\Lambda}(\lambda) d\lambda = tE[\Lambda].^2 \end{aligned}$$

$$\begin{aligned} P^{\text{pred}}(A \cap B) &= \int (1 - \exp(-\rho \lambda t)) f_{\Lambda}(\lambda) f_P(\rho) d\lambda d\rho \\ &\approx \int t \rho \lambda f_{\Lambda}(\lambda) f_P(\rho) d\lambda d\rho = tE[\Lambda]E[P]. \end{aligned}$$

Example 6.2

²For small x , $1 - \exp(-x) \approx x$.

Credibility intervals:

- ▶ In the Bayesian approach the lack of knowledge of parameter value θ is described using the probability densities $f(\theta)$ (odds). Random variable Θ having the pdf $f(\theta)$ models our knowledge about θ .
- ▶ The initial knowledge is described using $f^{\text{prior}}(\theta)$ density and as the data are gathered it is updated

$$f^{\text{post}}(\theta) = c L(\theta) f^{\text{prior}}(\theta).$$

- ▶ The pdf $f^{\text{post}}(\theta)$ summarizes our knowledge about θ . However if one value of for the parameter is needed then

$$\theta^{\text{predictive}} = E[\Theta] = \int \theta f^{\text{post}}(\theta) d\theta.$$

- ▶ If one wishes to describe the variability of θ by means of an interval then the so called **credibility interval** can be computed

$$[\theta_{1-\alpha/2}^{\text{post}}, \theta_{\alpha/2}^{\text{post}}]$$

Gamma-priors:

Conjugated priors are families of pdf for Θ which are particularly convenient for recursive updating procedures, *i.e.* when new observations arrive at different time instants. We will use three families of conjugated priors:

Gamma pdf:

$\Theta \in \text{Gamma}(a, b)$, $a, b > 0$, if

$$f(\theta) = c \theta^{a-1} e^{-b\theta}, \quad \theta \geq 0, \quad c = \frac{b^a}{\Gamma(a)}.$$

The expectation, variance and coefficient of variation for $\Theta \in \text{Gamma}(a, b)$ are given by

$$E[\Theta] = \frac{a}{b}, \quad V[\Theta] = \frac{a}{b^2}, \quad R[\Theta] = \frac{1}{\sqrt{a}}.$$

Updating Gamma priors:

The Gamma priors are conjugated priors for the problem of estimating the intensity in a Poisson stream of events A . If one has observed that in time \tilde{t} there were k events reported and if the prior density $f^{\text{prior}}(\theta) \in \text{Gamma}(a, b)$, then

$$f^{\text{post}}(\theta) \in \text{Gamma}(\tilde{a}, \tilde{b}), \quad \tilde{a} = a + k, \quad \tilde{b} = b + \tilde{t}.$$

Further, the predictive probability of at least one event A during a period of length t is given by

$$p^{\text{pred}}(A) \approx tE[\Theta] = t \frac{\tilde{a}}{\tilde{b}}$$

In Example 6.2 the $f^{\text{prior}}(\theta)$ was exponential with mean $1/30$ [days⁻¹]. This is Gamma(1,30) pdf. Suppose that in 10 days we have not observed any accidents then posteriori density $f^{\text{post}}(\theta)$ is Gamma(1,40). Hence

$$p^{\text{pred}}(A) \approx \frac{t}{40}.$$

Conjugated Beta-priors:

Beta probability-density function (pdf):

$\Theta \in \text{Beta}(a, b)$, $a, b > 0$, if

$$f(\theta) = c \theta^{a-1} (1 - \theta)^{b-1}, \quad 0 \leq \theta \leq 1, \quad c = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}.$$

The expectation and variance of $\Theta \in \text{Beta}(a, b)$ are given by

$$E[\Theta] = p, \quad V[\Theta] = \frac{p(1-p)}{a+b+1},$$

where $p = a/(a + b)$. Furthermore, the coefficient of variation

$$R(\Theta) = \frac{1}{\sqrt{a+b+1}} \sqrt{\frac{1-p}{p}}.$$

Updating Beta-priors:

The Beta priors are conjugated priors for the problem of estimating the probability $p_B = P(B)$.

Let $\theta = p_B$. If one has observed that in n trials (results of experiments), the statement B was true k times and if the prior density $f^{\text{prior}}(\theta) \in \text{Beta}(a, b)$ then

$$f^{\text{post}}(\theta) \in \text{Beta}(\tilde{a}, \tilde{b}), \quad \tilde{a} = a + k, \quad \tilde{b} = b + n - k.$$

$$p^{\text{pred}}(B) = \int_0^1 \theta f^{\text{post}}(\theta) d\theta = \frac{\tilde{a}}{\tilde{a} + \tilde{b}}.$$

Consider example of treatment of waste water. Let p be the probability that water is sufficiently cleaned after a week of treatment. If we have no knowledge about p we could use the uniform priors. It is easy to see that it is $\text{Beta}(1,1)$ pdf.

Suppose that 3 times water was well cleaned and 2 times not. This information gives the posterior density $\text{Beta}(4,3)$ and the predictive probability that water is cleaned in one week is $4/7$.

Conjugated Dirichlet-priors:

Dirichlet's pdf:

$\Theta = (\Theta_1, \Theta_2) \in \text{Dirichlet}(\mathbf{a})$, $\mathbf{a} = (a_1, a_2, a_3)$, $a_i > 0$, if

$$f(\theta_1, \theta_2) = c \theta_1^{a_1-1} \theta_2^{a_2-1} (1 - \theta_1 - \theta_2)^{a_3-1}, \quad \theta_i > 0, \theta_1 + \theta_2 < 1,$$

where $c = \frac{\Gamma(a_1+a_2+a_3)}{\Gamma(a_1)\Gamma(a_2)\Gamma(a_3)}$. Let $a_0 = a_1 + a_2 + a_3$; then

$$E[\Theta_i] = \frac{a_i}{a_0}, \quad V[\Theta_i] = \frac{a_i(a_0 - a_i)}{a_0^2(a_0 + 1)}, \quad i = 1, 2.$$

Furthermore the marginal probabilities are Beta distributed, *viz.*

$$\Theta_i \in \text{Beta}(a_i, a_0 - a_i), \quad i = 1, 2.$$

Updating Dirichlet's priors.

The Dirichlet priors are conjugated priors for the problem of estimating the probabilities $p_i = P(B_i)$, $i = 1, 2, 3$, B_i are disjoint, $p_1 + p_2 + p_3 = 1$.

Let $\theta_i = p_i$. If one has observed that the statement B_i was true k_i times in n trials and the prior density $f^{\text{prior}}(\theta_1, \theta_2) \in \text{Dirichlet}(\mathbf{a})$,

$$f^{\text{post}}(\theta_1, \theta_2) \in \text{Dirichlet}(\tilde{\mathbf{a}}), \quad \tilde{\mathbf{a}} = (a_1 + k_1, a_2 + k_2, a_3 + k_3),$$

where $k_3 = n - k_1 - k_2$. Further

$$p^{\text{pred}}(B_i) = E[\Theta_i] = \frac{\tilde{a}_i}{\tilde{a}_1 + \tilde{a}_2 + \tilde{a}_3}.$$

Let B_1 ="player A wins", B_2 ="player B wins" (there is possibility of draw). If we do not know strength of players we could use uniform priors which corresponds to Dirichlet(1,1,1) pdf. Now we observed that in two matches A won twice, hence the posterior density is Dirichlet(3,1,1) and the predictive probability that A wins the next match is then 3/5.

Posterior pdf for large number of observations.

If $f^{\text{prior}}(\theta_0) > 0$ then $\Theta \in \text{AsN}(\theta^*, (\sigma_{\mathcal{E}}^*)^2)$ as $n \rightarrow \infty$, where θ^* is the ML estimate of θ_0 and $\sigma_{\mathcal{E}}^* = 1/\sqrt{-\ddot{l}(\theta^*)}$.

It means that

$$f^{\text{post}}(\theta) \approx c \exp\left(\frac{1}{2}\ddot{l}(\theta^*)(\theta - \theta^*)^2\right) = c \exp\left(-\frac{1}{2}\left((\theta - \theta^*)^2/(\sigma_{\mathcal{E}}^*)^2\right)\right).$$

Sketch of proof:

$$l(\theta) \approx l(\theta^*) + \dot{l}(\theta^*)(\theta - \theta^*) + \frac{1}{2}\ddot{l}(\theta^*)(\theta - \theta^*)^2.$$

Now likelihood function $L(\theta) = e^{l(\theta)}$ and $\dot{l}(\theta^*) = 0$, thus

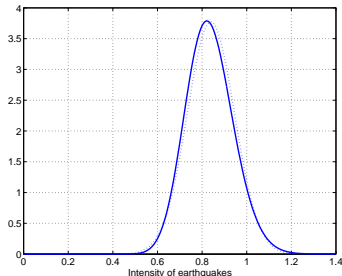
$$\begin{aligned} L(\theta) &\approx \exp\left(l(\theta^*) + \dot{l}(\theta^*)(\theta - \theta^*) + \frac{1}{2}\ddot{l}(\theta^*)(\theta - \theta^*)^2\right) \\ &= c \exp\left(\frac{1}{2}\ddot{l}(\theta^*)(\theta - \theta^*)^2\right). \end{aligned}$$

Example earthquake data:

We have demonstrated that time between earthquakes is $\text{Exp}(a)$. Here it is more convenient to use parameter $\theta = 1/a$, i.e. the intensity of earthquakes. The ML estimate $\theta^* = 1/\bar{x}$ and $\ddot{l}(\theta) = -n/\theta^2$. Since $\bar{x} = 437.2$ days we have that $\theta^* = 364/437.2 = 0.8395 \text{ years}^{-1}$, while

$$(\sigma_{\mathcal{E}}^*)^2 = \frac{(\theta^*)^2}{n} = 0.0112.$$

Consequently $\Theta^* \approx N(0.8395, 0.0112)$. This can be used to give approx. confidence interval for θ or $p = P(T > 4.1) = \exp(-4.1\theta)$.



Let use non-informative priors $f^{\text{prior}}(\theta) = 1/\theta$ then the gamma posterior density has parameters $a = 62$ and $b = (437.2/365) \cdot 62 = 74.26$;
 $f^{\text{post}}(\theta) \in \text{Gamma}(62, 74.26)$ (solid line):
Asymptotic normal posterior pdf $N(0.8395, 0.0112)$ (dotted line).

Transport of nuclear fuel waste

Spent nuclear fuel is transported by railroad. From historical data, one knows that there were 4 000 transports without a single release of radioactive material. Since fuel waste is highly dangerous, one has discussed the possibility of constructing a special (very safe and expensive) train to transport the spent fuel.

One problem was the definition of an acceptable risk p^{acc} for an accident, i.e. one wishes the probability of an accident θ , say, to be smaller than p^{acc} . Since θ is unknown and uncertainty of its value is modelled by a random variable Θ the issue is to check, on basis of available data and experience, whether the predictive probability $P(\Theta < p^{\text{acc}})$ is high.

A number between 10^{-8} and 10^{-10} was first proposed for p^{acc} , i.e. the average waiting time for an accident is 10^8 to 10^{10} transports. In such a scale the experienced 4000 safe transports looks clearly negligible and hence the conclusion was: if one wishes to transport the waste with the required reliability, one needs to develop transport systems with maximum reliability.

How the information about 4 000 problem free transports affects our beliefs about risk for accidents. Suppose that accidents happen independently with probability θ . Then³

$$P(\text{"No accidents for 4 000 transports"} \mid \Theta = \theta) = (1 - \theta)^{4000} \approx e^{-4000\theta},$$

and the posterior density $f^{\text{post}}(\theta) = cf^{\text{prior}}(\theta)e^{-4000\theta}$ will be close to zero for any reasonable choice of the prior density and $\theta > 10^{-3}$. This agrees with the conclusion of Kaplan and Garrick that the information of 4 000 release-free transport is quite informative:

"The experience of 4 000 release-free shipments is not sufficient to distinguish between release frequencies of 10^{-5} or less. However, it is sufficient to substantially reduce our belief that the frequency is on the order of 10^{-4} and virtually demolish any belief that the frequency could be 10^{-3} or greater".

If we assume that the required safety is $p = 10^{-8}$, then the information of 4 000 accident-free transports is insignificant; on the other hand, the required safety may never be checked.

³Here we use that for small θ , $e^{-\theta} \approx 1 - \theta$. In addition $\lim_{n \rightarrow \infty} \left(1 - \frac{a}{n}\right)^n = e^{-a}$.