07 - More priors - 01

Master in Foundations of Data Science
Bayesian Statistics and Probabilistic Programming
Fall 2018-2019

Josep Fortiana

Wednesday, November 07, 2018

Universitat de Barcelona

07 - More priors - 01

Bayesian robustness

Mixture priors: the spinning coin

07 - More priors - 01

Bayesian robustness

Mixture priors: the spinning coin

Concept of robustness

A Bayesian analysis is said to be *robust* to the choice of prior if the inference does not depend on which prior pdf is chosen, from a set of priors compatible with the user's beliefs.

Example: estimating IQ with two priors

Assume the IQ of a person is $X \sim \mathrm{N}(\theta, \sigma^2)$, where $\sigma = 15$ is known.

We want to estimate θ, “the true IQ” of a person whom we believe to be average, hence the prior distribution of θ will have median $m = 100$.

Also, we believe that with 90% probability θ falls between 80 and 120.

Albert, J. (2009), Bayesian computation with R (2nd ed). Springer.

First prior: Normal

Firstly we take the conjugate prior $\theta \sim \mathsf{N}(\mu, \gamma^2)$.

$\mu = m = 100$, and γ is derived from the fact that 80 is the 0.05 quantile of θ or, equivalently, that 120 is the 0.95 percentile.

Since $Z = \dfrac{\theta - \mu}{\gamma} \sim \mathsf{N}(0, 1)$, the 0.05 quantile of Z, equal to `qnorm(0.05)=-1.6449`, must also be equal to $\dfrac{80 - 100}{\gamma}$, hence: $\gamma = 12.1591$.

Second prior: Student t with 2 degrees of freedom

Now we assume that θ is the result of applying a translation and a scale change to a r.v. $Y \sim t(2)$, a standard Student t with 2 degrees of freedom:

$$\theta = m + s \cdot Y, \text{ where } Y \sim t(2),$$

equivalently,

$$Y = \frac{\theta - m}{s} \sim t(2).$$

pdf of the Student t prior

If h_Y is the pdf of the standard r.v. $Y \sim t(2)$, the pdf of $\theta = m + s \cdot Y$ is:

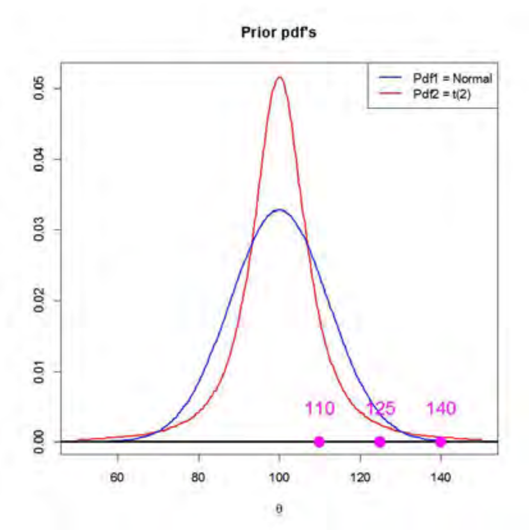
$$h_\theta(u) = \frac{1}{s} \cdot h_Y\left(\frac{u-m}{s}\right),$$

which follows either from the pdf's change of variable formula, or directly, by differentiating the cdf:

$$H_\theta(u) = \mathrm{P}(\theta \leq u) = \mathrm{P}(m + s \cdot Y \leq u) = H_Y\left(\frac{u-m}{s}\right),$$

where H_Y is the cdf of Y.

Prior pdf's


Prior pdf's
Pdf1 = Normal
Pdf2 = t(2)
0.00
0.01
0.02
0.03
0.04
0.05
110
125
140
60
80
100
120
140
θ

Observed data

Our subject takes $n = 4$ IQ tests, obtaining an average score $\bar{x}$.

The we obtain the posterior pdf with both prior pdf's.

We perform our computations for three hypothetical values of $\bar{x}$:

$$\bar{x} = 110, \qquad \bar{x} = 125, \qquad \bar{x} = 140.$$

First posterior

The posterior distribution is $\mathrm{N}(\mu_x, \tau^2)$, where:

$$\mathrm{E}(\theta|x) \;=\; \mu_x \stackrel{\text{def}}{=\!=} \frac{\gamma^2}{\sigma^2/n+\gamma^2}\,\bar{x} + \frac{\sigma^2/n}{\sigma^2/n+\gamma^2}\,\mu$$

$$\mathrm{var}(\theta|x) \;=\; \tau^2 \stackrel{\text{def}}{=\!=} \frac{\sigma^2\,\gamma^2}{\sigma^2+n\,\gamma^2}$$

Posterior parameters

Posterior expectations:

Observed $\bar{x}$	110	125	140
μ_x	107.24	118.11	128.98

Posterior variance and standard deviation (the same for the three cases):

$$\tau^2 = 40.7471, \qquad \tau = 6.3833.$$

Second posterior

Since the prior pdf $h_\theta(\,\cdot\,)$ is not a conjugate prior for a normal mean, we evaluate the joint pdf:

$$f(\bar{x},\theta) = f(\bar{x}|\theta) \times h_\theta(\theta),$$

on a grid, N points on an interval $I = (m-\Delta, m+\Delta)$, where Δ is chosen so that $f \approx 0$ whithout it and N is large enough for the desired precision.

Second posterior

Normalizing the resulting vector to sum 1 we have the pmf of a discretization of the posterior distribution.

With this pmf we evaluate whatever posterior properties we need.

Posterior parameters

Expectations, m, standard deviations s:

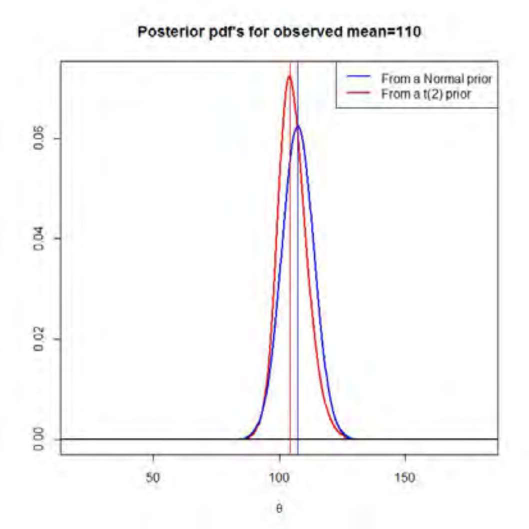
Observed $\bar{x}$	110	125	140
m	105.292	118.084	135.413
s	5.842	7.885	7.973

Discussion

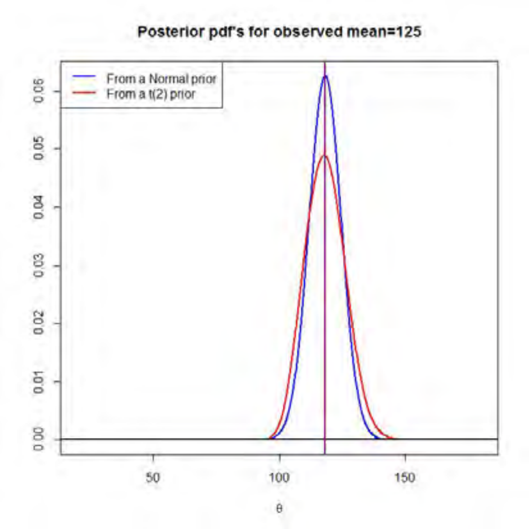
When a normal prior is used, the posterior will always be a compromise between the prior information and the observed data, even when the observed data conflicts with prior beliefs.

In contrast, when a t prior is used, when the likelihood falls in the flat-tailed portion of the prior and the posterior will resemble the likelihood function.

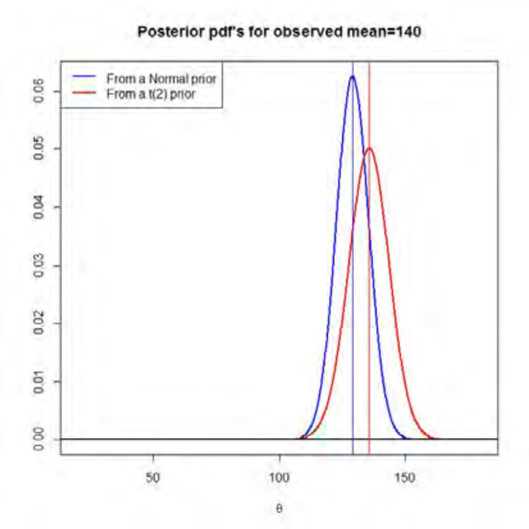
Posterior pdf's



Posterior pdf's



Posterior pdf's



07 - More priors - 01

Bayesian robustness

Mixture priors: the spinning coin

Persi Diaconis

Stanford stat and math professor Persi Diaconis, formerly a professional magician, famously determined how many times a deck of cards must be shuffled in order to give a mathematically random result (seven). He's also dabbled in coin games.

BTW: Persi Diaconis on randomness

See his 2013 video talk:

The Search for Randomness

The spinning coin

Fact: if a coin is spinned on its edge instead of being flipped, proportion of heads or tails is not around 50% but rather such values as 25% or 75% are obtained.

Persi Diaconis on the spinning coin

According to Diaconis, “the reasons for the bias are not hard to infer. The shape of the edge will be a strong determining factor – indeed, magicians have coins that are slightly shaved; the eye cannot detect the shaving, but the spun coin always comes up heads”.

A prior for the spinning coin problem

For n tosses of a spinning coin, the number x of heads up is a $\mathrm{B}(n, \theta)$, and θ's prior pdf will typically be a bimodal function (presenting two local maxima).

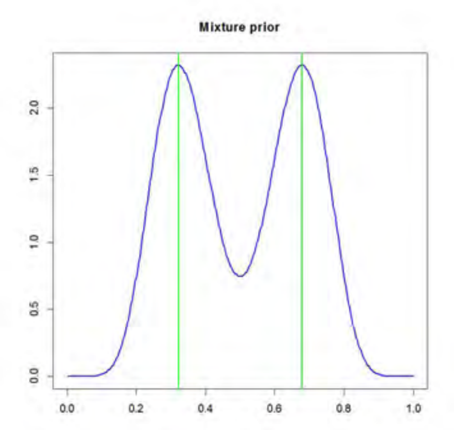
Hence it cannot be modelled with a $\mathsf{Beta}(\alpha, \beta)$, which has a single mode at:

$$\frac{\alpha - 1}{\alpha + \beta - 2}$$

Diaconis, Persi and Donald Ylvisaker (1985) *Quantifying prior opinion.* In: J.M. Bernardo et al (eds), Bayesian Statistics 2, Elsevier Science Publishers

A possible prior

$$0.50\ \mathrm{Beta}(10, 20) + 0.50\ \mathrm{Beta}(20, 10).$$



Interpretation of a mixture prior

The mixture prior can be thought of as a weighted combination of "beta populations", the weights γ_i measuring the prior degree of belief that the actual coin was chosen from the i-th population.

Simulation of a mixture prior

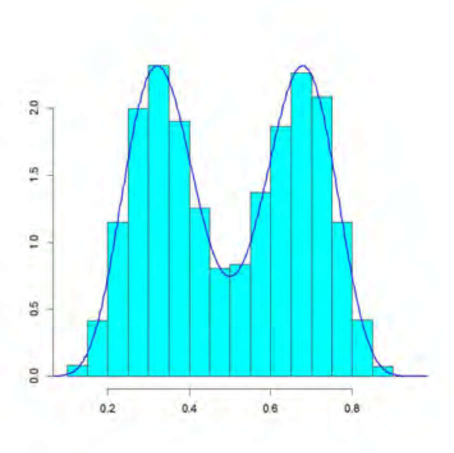
To generate a sequence of random numbers from a mixture:

$$\gamma \cdot h_1 + (1-\gamma) \cdot h_2, \quad \gamma \in (0,1),$$

we generate a sequence of realizations of $I \sim \mathrm{Ber}(\gamma)$ and:

- For each entry equal to 1, a realization of h_1,
- For each entry equal to 0, a realization of h_2.

Simulation of a mixture prior



Bayesian modelling with a mixture prior

Assume the prior pdf for θ is:

$$h(\theta) = \gamma \cdot h_1(\theta) + (1-\gamma) \cdot h_2(\theta),$$

and the likelihood is: $f(x|\theta)$. Then, the joint pdf is:

$$f(x,\theta) = \gamma \cdot f_1(x,\theta) + (1-\gamma) \cdot f_2(x,\theta).$$

where:

$$f_i(x,\theta) = f(x|\theta) \cdot h_i(\theta), \qquad i = 1, 2.$$

Prior predictive pdf from a mixture prior

Integrating out θ, the marginal for x:

$$f(x) = \gamma \cdot f_1(x) + (1-\gamma) \cdot f_2(x),$$

where:

$$f_i(x) = \int f(x|\theta) \cdot h_i(\theta)\, d\theta, \qquad i = 1, 2.$$

Computing the posterior pdf from a mixture prior

From Bayes' formula:

$$h(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{\gamma \cdot f_1(x,\theta) + (1-\gamma) \cdot f_2(x,\theta)}{\gamma \cdot f_1(x) + (1-\gamma) \cdot f_2(x)}.$$

We will use the obvious notation:

$$h_i(\theta|x) = \frac{f_i(x,\theta)}{f_i(x)} \qquad i = 1, 2.$$

Posterior pdf from a mixture prior

The posterior pdf is:

$$h(\theta|x) = \widehat{\gamma}(x) \cdot h_1(\theta|x) + (1 - \widehat{\gamma}(x)) \cdot h_2(\theta|x),$$

where the *posterior mixture weights* are:

$$\widehat{\gamma}(x) = \frac{\gamma \cdot f_1(x)}{\gamma \cdot f_1(x) + (1 - \gamma) \cdot f_2(x)}$$

and $1 - \widehat{\gamma}(x)$.

For the Diaconis spinning coin experiment

He reports $x = 3$ heads out of $n = 10$ spins.

The (partial) posterior pdf's are:

$$\begin{aligned} h_1 &\sim \mathrm{Beta}(\alpha_1 + x, \beta_1 + n - x) \\ &= \mathrm{Beta}(10 + 3, 20 + 10 - 3) = \mathrm{Beta}(13, 27), \\ h_2 &\sim \mathrm{Beta}(\alpha_2 + x, \beta_2 + n - x) \\ &= \mathrm{Beta}(20 + 3, 10 + 10 - 3) = \mathrm{Beta}(23, 17). \end{aligned}$$

For the Diaconis spinning coin experiment

The (partial) prior predictive pdf's are:

$$
\begin{aligned}
f_1 &\sim \text{Beta-Binom}(n, \alpha_1, \beta_1) = \text{Beta-Binom}(10, 10, 20),\\
f_2 &\sim \text{Beta-Binom}(n, \alpha_2, \beta_2) = \text{Beta-Binom}(10, 20, 10),
\end{aligned}
$$

which, for the observed $x = 3$, give:

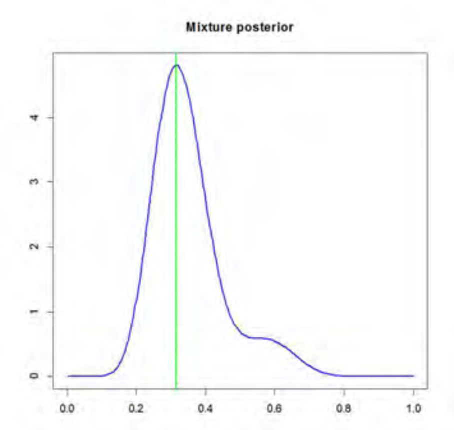
$$
f_1(x) = 0.2276, \qquad f_2(x) = 0.0277.
$$

Posterior weights

$$\widehat{\gamma}(x) = \frac{\gamma \cdot f_1(x)}{\gamma \cdot f_1(x) + (1-\gamma) \cdot f_2(x)}$$
$$= 0.8915$$

Posterior pdf

$$0.8915\ \mathrm{Beta}(13,27) + 0.1085\ \mathrm{Beta}(23,17).$$



MAP estimator of θ

The posterior distribution attains its maximum value where $\mathrm{Beta}(13, 27)$ does:

$$\hat{\theta}_{MAP} = \frac{13-1}{13+27-2} = 0.3158,$$

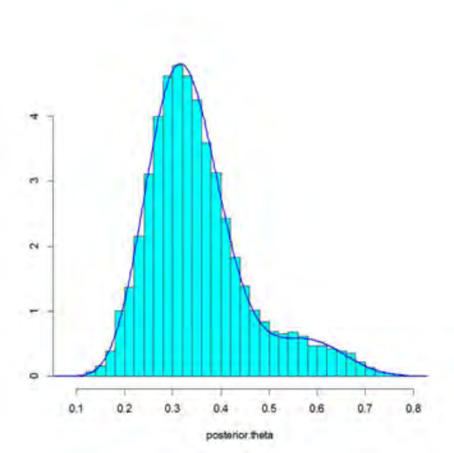
which may be compared with the classical ML estimator:

$$\hat{\theta}_{ML} = \frac{x}{n} = 0.4286,$$

and the lower mode of the prior distribution:

$$m_1 = \frac{10-1}{10+20-2} = 0.3214.$$

Simulating the posterior pdf



Credible interval with a given probability

The easiest way to obtain an interval $[a, b]$ of θ values such that its posterior probability is a given value, e.g., $\pi = 0.95$ is to use a random sample from the posterior distribution.

Taking it quantile-symmetrical, i.e., such that both tails have probability $(1 - \pi)/2 = 0.025$, the interval is:

$$(0.1932, 0.6400).$$

A more elaborate mixture prior

On reflection, it was decided that tails had come up more often than heads in the past; further some coins seemed likely to be symmetric.

A final approximation to the prior was taken as:

$$0.50 \text{ Beta}(10, 20) + 0.20 \text{ Beta}(15, 15) + 0.30 \text{ Beta}(20, 10).$$

Exercise

Obtain the posterior pdf for this prior, a mixture of three Beta distributions.