Numerical and Statistical Methods for Finance

Bayesian Inference with R

Pierpaolo De Blasi

University of Torino
email: pierpaolo.deblasi@unito.it
webpage: sites.google.com/a/carloalberto.org/pdeblasi/

Lecture no. 9
21 November 2013

Bayesian learning

Bayes' rule provides a rational method for updating beliefs on the unknown θ in light of new information. The process of inductive learning via Bayes' rule is referred to as *Bayesian inference*.

Given a statistical model $\{(\mathbb{X}, \mathscr{X}, f_\theta) : \theta \in \Theta\}$,

- Describe prior beliefs that θ represents the true parameter value via a distribution $p(\theta)$ over Θ, known as *prior distribution*.
- For $\theta \in \Theta$ and $x \in \mathbb{X}$, the *sampling distribution* $f_\theta(x)$ denotes now the conditional distribution of X given θ:

$$p(x|\theta) := f_\theta(x)$$

- The "inverse" conditional distribution $p(\theta|x)$, known as *posterior distribution*, describes beliefs that θ is the true value, having observed x,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\tilde{\theta})p(\tilde{\theta})\mathrm{d}\tilde{\theta}}$$

Prior distribution

A major difference with the classical approach to statistical inference is that the parameter θ is now a *random variable*.

Bayesian learning begins with a numerical formulation of the *subjective beliefs* about θ. In theory, any distribution $p(\theta)$ over Θ is admissible, the problem is how to elicit prior information in $p(\theta)$.

Typically, one selects a family of distribution indexed by some *hyperparameter* η, then chooses the value of η to reflect prior knowledge on center, dispersion, quantiles, etc... of $p(\theta)$.

Relevant issues

- how much sensitive is inference to the choice of the prior $p(\theta)$?
- How to specify $p(\theta)$ when little prior information on θ is available, i.e. a *noninformative prior*?

Sampling distribution

$X_1, \ldots, X_n$ are *conditionally independent* given θ if

$$p(x_1, \ldots, x_n|\theta) = p_{X_1}(x_1|\theta) \times \cdots \times p_{X_n}(x_n|\theta)$$

When the $p_{X_i}(x_i|\theta)$ are all equal, $X_1, \ldots, X_n$ are *conditionally i.i.d.*:

$$X_1, \ldots, X_n|\theta \sim \text{i.i.d. } p(x|\theta)$$

However, $X_1, \ldots, X_n$ are *marginally* dependent,

$$p(x_1 \ldots, x_n) = \int_\Theta \prod_{i=1}^n p(x_i|\theta)\ p(\theta)\mathrm{d}\theta$$

in fact, they are *exchangeable*, $p(x_1, \ldots, x_n) = p(x_{\pi_1}, \ldots, x_{\pi_n})$, for all permutations $\pi_1, \ldots, \pi_n$ of $1, \ldots, n$.

Such dependence is at the fundation of Bayesian learning: it allows *prediction* of a future value X_{n+1} given the observed ones.

Posterior distribution

The posterior distribution is given, up to a proportionality constant, by the product of the prior distribution and the likelihood,

$$p(\theta|x) \propto \underbrace{p(x|\theta)}_{\text{Likelihood}(\theta)} \times p(\theta)$$

The posterior conveys all information on which we base inference:

- point estimation: $E(\theta|x) = \int_\Theta \theta p(\theta|x)\mathrm{d}\theta$ (*posterior expectation*)
- interval estimate: $(\theta_L, \theta_U) : \int_{\theta_L}^{\theta_U} p(\theta|x)\mathrm{d}\theta = 0.95$ (HPD intervals)
  we do say "the prob that θ belongs to..."
- hypothesis test via *Bayes factor*: for $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ compute $\frac{P(\theta\in\Theta_0|x)}{P(\theta\in\Theta_1|x)} \Big/ \frac{P(\theta\in\Theta_0)}{P(\theta\in\Theta_1)}$ as measure of evidence for H_0.
- prediction via the *posterior predictive density*:

$$X_{n+1}|X_1, \dots X_n \sim p_n(x_{n+1}) = \int_\Theta p(x_{n+1}|\theta)\ p(\theta|x_1, \dots, x_n)\mathrm{d}\theta$$

Posterior approximation

Suppose we cannot draw exact posterior inference since the posterior density $p(\theta|x)$ has not a familiar functional form (or it is not implemented in R).

However, suppose we can directly simulate samples from the posterior distribution. Then we can summarize the simulated output via *posterior averaging*. Popular simulation methods are

- *Monte Carlo methods* such as rejection sampling, importance sampling and sampling importance resampling;
- *Gibbs sampler*: in multiparameter models the joint distribution is intractable, however it is easy to sample from the full conditional distribution of each parameter with Monte Carlo methods.

In situations where the full conditional distributions of the parameters do not have a standard form and the Gibbs sampler cannot be easily used, one can resort to *Markov Chain Monte Carlo* (MCMC) methods: construct a Markov chain on Θ with $p(\theta|x)$ as stationary distribution. A general method is the *Metropolis- Hastings algorithm*.

The popularity of Bayesian methods has greatly increased with the development of sampling methods for complex models.

Robust Bayesian estimation of a normal mean

In practice, one may have incomplete prior information about a parameter in the sense that one's beliefs will not entirely define a prior density. There may be a number of different priors that match the given prior information.

In this situation, it is desirable that inference from the posterior is not too much influenced by the exact functional form of the prior. A Bayesian analysis is said to be *robust* to the choice of the prior if the inference is insensitive to different priors that match the user's beliefs.

Suppose we are interested in estimating the true IQ θ for a person. We believe she has average intelligence, median $= 100$, and we are 90% confident that her IQ falls between 80 and 120. By using a symmetric prior, it corresponds to 95th percentile $= 120$.

A normal prior for $p(\theta) \sim \mathrm{N}(\mu, \tau)$ (τ is s.d.) with parameters

$$\mu = 100, \quad \tau = 12.15914$$

match this prior information (see function `normal.select`).

IQ test comprises $n = 4$ scores, y_1, y_2, y_3, y_4. Assuming $y \sim \mathsf{N}(\theta, \sigma)$ for known standard deviation $\sigma = 15$, the observed mean score $\bar{y}$ has distribution

$$\bar{y} \sim \mathsf{N}(\theta, \sigma/\sqrt{n})$$

The normal prior $p(\theta) \sim \mathsf{N}(\mu, \tau)$is conjugate for the normal sampling model: the posterior is also normal

$$p(\theta|\bar{y}) \sim \mathsf{N}(\mu_1, \tau_1)$$

where μ_1 and τ_1 are as follows.
Let P_1 be the posterior precision parameter, $P_1 = 1/\tau_1^2$. Then P_1 is the sum of the data precision n/σ^2 and the prior precision $1/\tau^2$,
$P_1 = n/\sigma^2 + 1/\tau^2$, so that

$$\tau_1 = 1/\sqrt{P_1} = 1\Big/\sqrt{n/\sigma^2 + 1/\tau^2}$$

The posterior mean μ_1 of θ is a weighted average of the sample mean $\bar{y}$ and the prior mean μ where the weights are proportional to the precisions:

$$\mu_1 = \frac{(n/\sigma^2)\bar{y} + (1/\tau^2)\mu}{n/\sigma^2 + 1/\tau^2}$$

We illustrate the posterior calculations for three hypothetical test results, $\bar{y} = 110$, $\bar{y} = 125$ and $\bar{y} = 140$:

```
     ybar      mu1     tau1
[1,]  110 107.2442 6.383469
[2,]  125 118.1105 6.383469
[3,]  140 128.9768 6.383469
```

Student's t prior

Student's t–distribution with location μ, scale τ, and 2 degree of freedom:

$$g_T(\theta|\nu,\mu,\tau) = \frac{1}{\tau} g_T\left(\frac{\theta-\mu}{\tau}\middle|\nu,0,0\right)$$

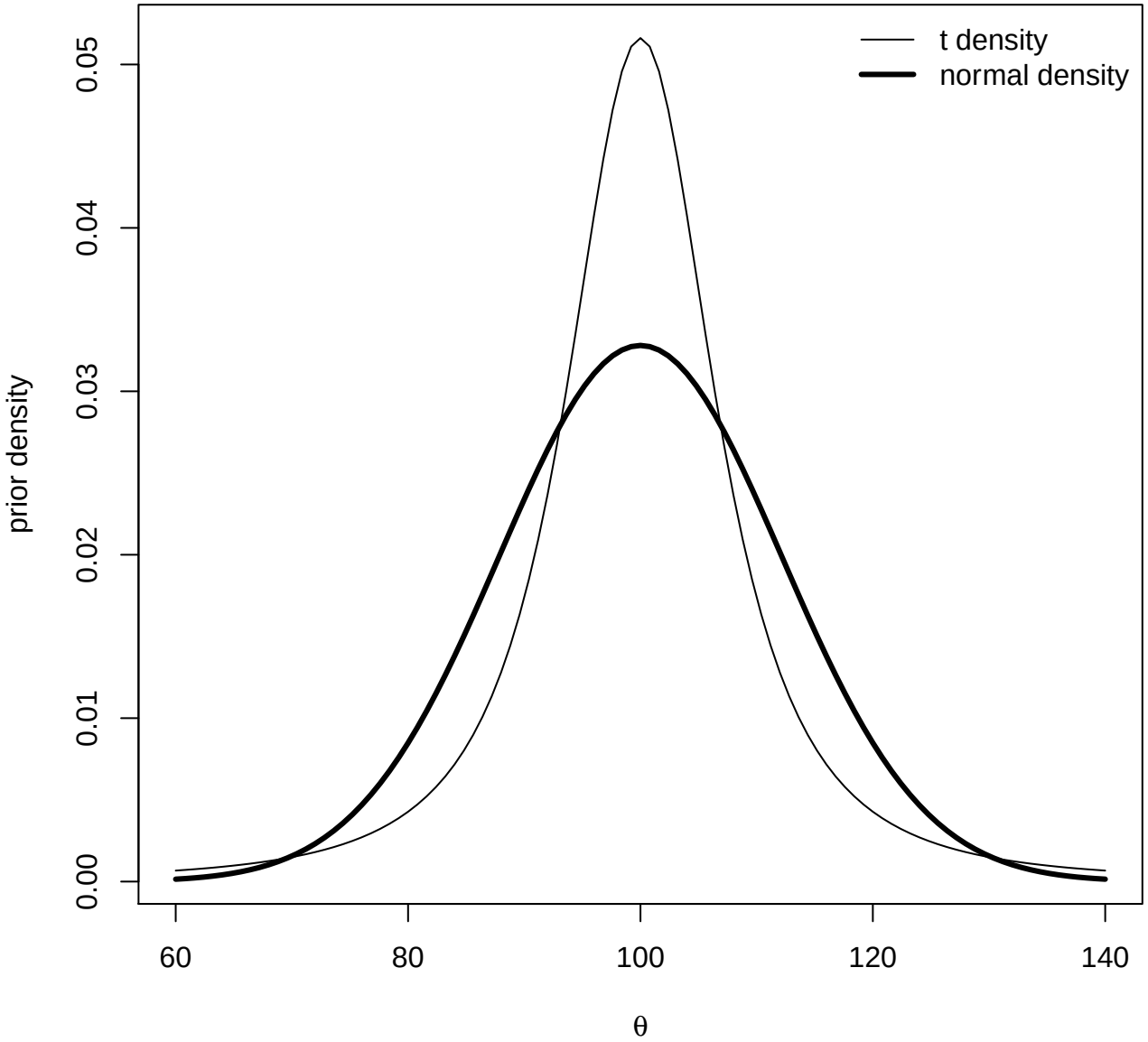
where $g_T(x|\nu,0,0)$ is the t-density with ν degrees of freedom.

Note that $g_T(x|\nu,0,0)$ can be computed with the density function `dt(x,df=2)` in R .

Since our prior belief on the median is 100, we set $\mu = 100$. We find the scale parameter τ such that

$$P(\theta \le 120) = 0.95; \quad P\left(\frac{T-\mu}{\tau} \le \frac{120-\mu}{\tau}\right) = 0.95;$$
$$P\left(T \le \frac{20}{\tau}\right) = 0.95; \quad \frac{20}{\tau} = t_{.05,2}; \quad \tau = 20/t_{.05,2}$$

where T is a standard t random variable with 2 degrees of freedom and $t_{\alpha,2}$ is the quantile of order $1-\alpha$ of T.


t density
normal density
prior density
0.00
0.01
0.02
0.03
0.04
0.05
60
80
100
120
140
θ

We consider the posterior with the Student's t prior for each of the three possible test results. The posterior density is given, up to a proportionality constant, by

$$p(\theta|\bar{y}) \propto \phi(\bar{y}|\theta, \sigma/\sqrt{n}) g_T(\theta|\nu, \mu, \tau)$$

where $\phi(y|\theta, \sigma) \sim \mathrm{N}(\theta, \sigma)$ and $g_T(\theta|\nu, \mu, \tau)$ is the t-density with mean μ, scale parameter τ, and degrees of freedom ν.

Note that it is not possible to derive $p(\theta|\bar{y})$ in explicit form due to the normalizing constant

$$\int_{-\infty}^{\infty} \phi(\bar{y}|\theta, \sigma/\sqrt{n}) g_T(\theta|\nu, \mu, \tau) \mathrm{d}\theta$$

not computable in closed form.

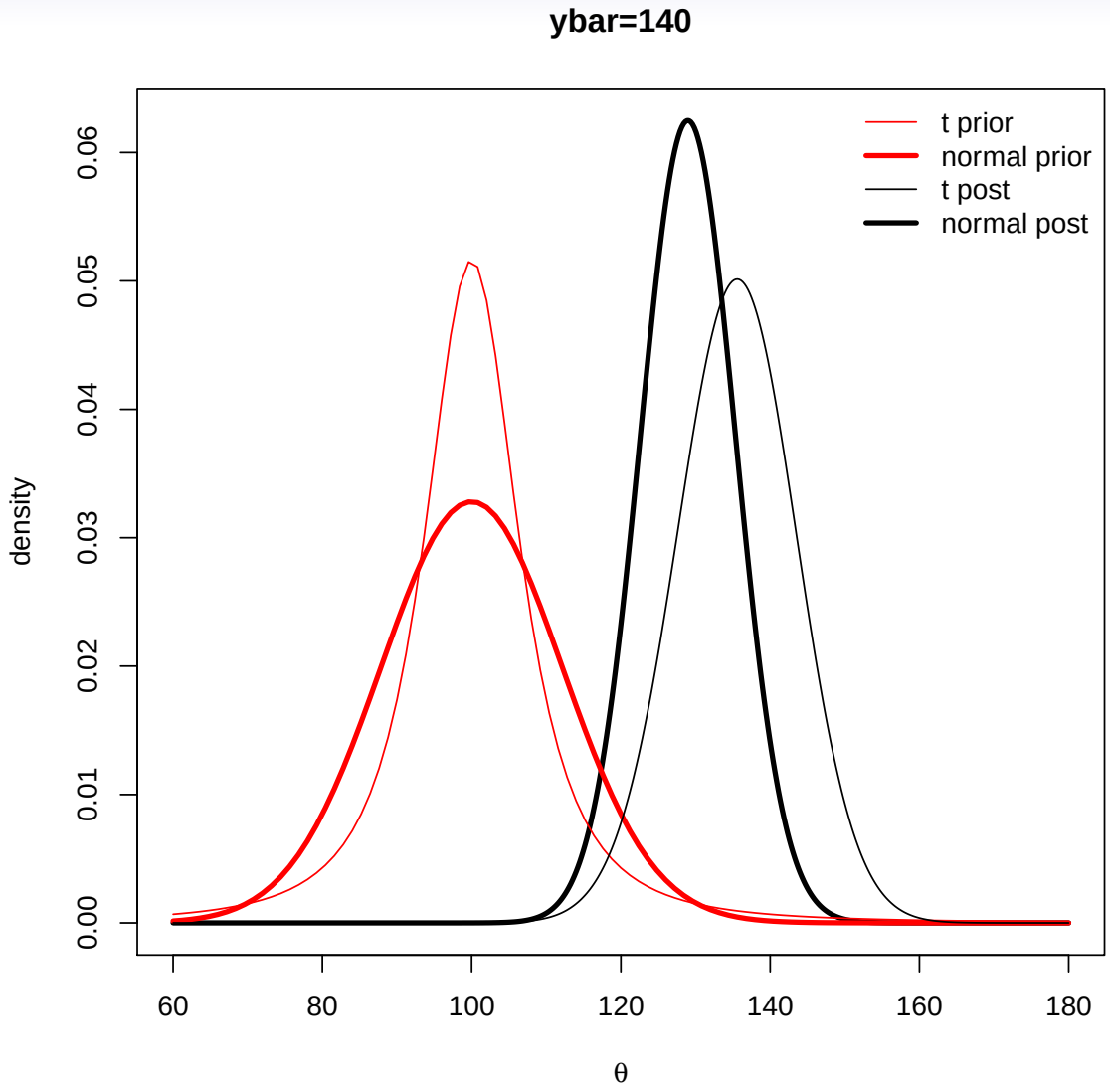
Approximate Bayesian Computation (ABC)

Since this posterior density does not have convenient functional form, we summarize it by using a direct "prior times likelihood" approach.

Essentially we are approximating the continuous posterior density by a discrete distribution on a grid of θ values. We then use this discrete distribution to compute the posterior mean and posterior standard deviation.

	Normal			Student's t		
	ybar	mu1	tau1	ybar	mu1t	tau1t
[1,]	110	107.2442	6.383469	110	105.2921	5.841676
[2,]	125	118.1105	6.383469	125	118.0841	7.885174
[3,]	140	128.9768	6.383469	140	135.4134	7.973498

There are substantial differences in the posterior moments using the two priors when the observed mean score is inconsistent with the prior mean, as with "extreme" test result $\bar{y} = 140$.



Posterior densities for the two priors when $\bar{y} = 140$.

When a normal prior is used, the posterior will always be a compromise between the prior information and the observed data, even when the data result conflicts with one's prior beliefs about the location of IQ.

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

In this case, the inference about the mean is robust to the choice of prior (normal or t) when the observed mean IQ score is consistent with the prior beliefs. But in the case where an extreme IQ score is observed, we see that the inference is not robust to the choice of prior density.

Rejection Sampling

Note that the posterior with the Student's t prior is only available up to the normalizing constant: $p(\theta|\bar{y}) \propto \phi(\bar{y}|\theta, \sigma/\sqrt{n}) g_T(\theta|\nu, \mu, \tau)$ that is it is not known in explicit form. We use rejection sampling.

Suppose we have a method for generating a random variable having density $g(\theta)$. We can use this as the basis for generating from $p(\theta|\bar{y})$ by generating θ^* from $g(\theta)$ and then accepting this generated value with a probability proportional to $p(\theta^*|\bar{y})/g(\theta^*)$.

Let c be a constant such that

$$p(\theta|\bar{y})/g(\theta) \leq c \quad \text{for all } \theta \tag{1}$$

Step 1. Generate θ^* from $g(\theta)$.

Step 2. Generate a random number U.

Step 3. If $U \leq p(\theta^*|\bar{y})/cg(\theta^*)$, set $\theta = \theta^*$. Otherwise return to Step 1.

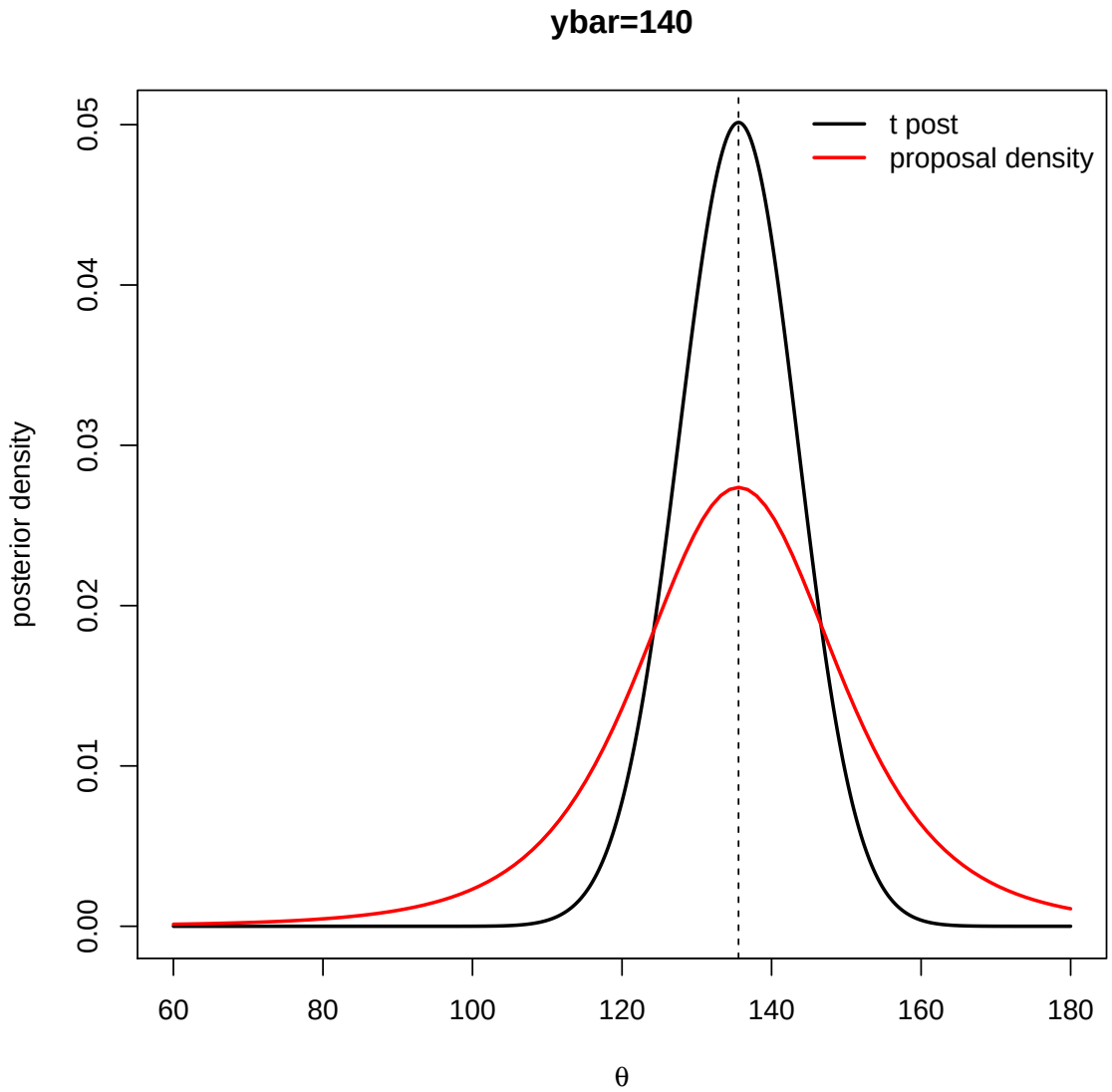
(i) *The random variable θ generated by the rejection method has density $p(\theta|\bar{y})$.*

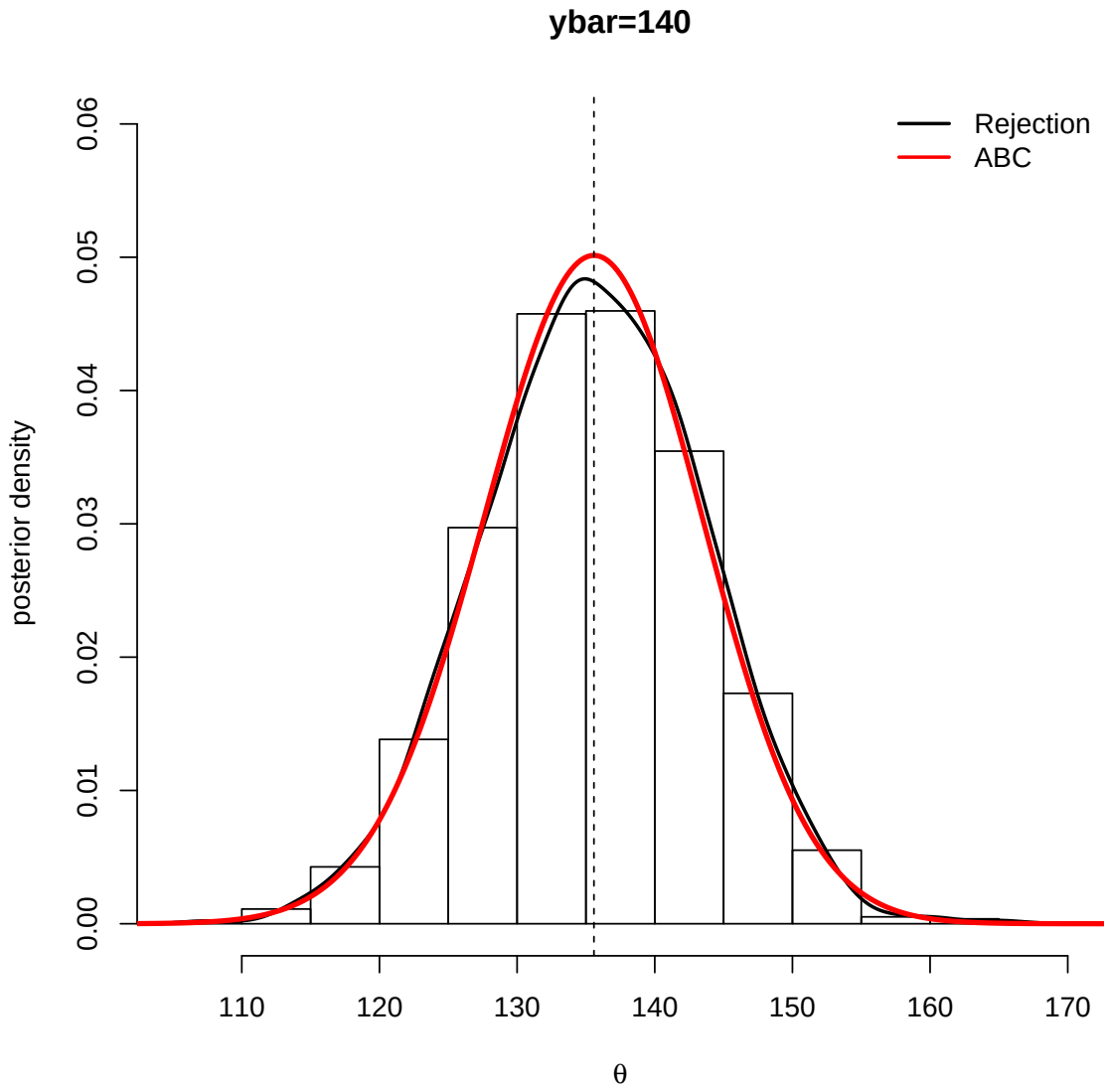
(ii) *The number of iterations of the algorithm that are needed is a geometric random variable with mean c.*

Implementation issues:

- Since the posterior takes values on $\mathbb{R}$, choose $g(\theta)$ with tails heavier than $p(\theta|\bar{y})$ so that c in (1) exists *(why?)*
- No need of the normalizing constant of $p(\theta|\bar{y})$ when computing $\theta_m := \arg\max_\theta p(\theta|\bar{y})/g(\theta)$. Choose g unimodal with mode equal to posterior mode so that θ_m coincides with the mode
- With $c = p(\theta_m|\bar{y})/g(\theta_m)$, no need of the normalizing constant for computing the acceptance probability:

$$\frac{p(\theta^*|\bar{y})}{cg(\theta^*)} = \frac{p(\theta^*|\bar{y})}{g(\theta^*)} \Big/ \frac{p(\theta_m|\bar{y})}{g(\theta_m)} = \frac{p(\theta^*|\bar{y})}{p(\theta_m|\bar{y})} \Big/ \frac{g(\theta^*)}{g(\theta_m)}$$


ybar=140
t post
proposal density
posterior density
0.00
0.01
0.02
0.03
0.04
0.05
60
80
100
120
140
160
180
θ


ybar=140
Rejection
ABC
posterior density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
110
120
130
140
150
160
170
θ

Resources

- BOOKS
  - Owen J., Maillardet R. and Robinson A. (2009). *Introduction to Scientific Programming and Simulation Using R.* Chapman & Hall/CRC.
  - Ross, S. (2006). *Simulation. 4th edn.* Academic Press.
  - Albert, J. (2008). *Bayesian Computation with R (2nd ed).* Springer.
- WEB
  - R software: `http://www.r-project.org/`
  - Owen, et al. (2009): `http://www.ms.unimelb.edu.au/spuRs/`
  - Albert (2008): `http://bayes.bgsu.edu/bcwr/`