Exercise Cauchy Location Posterior

Class: Bayesian Statistics
Instructor: Luiz Max Carvalho
Student: Lucas Machado Moschen

May 22, 2021

Turn in date: until 05/26/2021 at 23:59h Brasilia Time.

Exercise. *Take* $X_i \sim \text{Cauchy}(\theta, 1)$, $i = 1, 2, \ldots, 10$. *In particular, suppose*

$$\boldsymbol{x} = \{-5, -3, 0, 2, 4, 5, 7, 9, 11, 14\}. \tag{1}$$

i) Compute the MLE and l''*;*

ii) Deduce the parameters of the normal approximation to $p(\theta \mid \boldsymbol{x})$*;*

iii) Use an MCMC routine to sample from $p(\theta \mid \boldsymbol{x})$*, obtain a posterior approximation to its density and compare it to the normal approximation;*

iv) Simulate data sets of sizes

$$n = 20, 50, 100, 500, 1000 \text{ and } 10,000$$

and repeat iii.

v) See if you can reduce/increase the discrepancy between the posterior and its approximation by fiddling with the prior (without breaking the regularity assumptions!).

Maximum likelihood estimator

The probability density function of the Cauchy distribution with location parameter θ and scale parameter $\gamma = 1$ with respect to the Lebesgue measure is

$$f(x|\theta) = \frac{1}{\pi[1 + (x - \theta)^2]}. \tag{2}$$

Starting from that, we can define the likelihood for a sample (i.i.d.) with size n

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta) = \left[\pi^n \prod_{i=1}^{n} [1 + (x_i - \theta)^2]\right]^{-1} \tag{3}$$

and the log-likelihood

$$l(\theta|\boldsymbol{x}) = -n\log(\pi) - \sum_{i=1}^{n} \log\left(1 + (x_i - \theta)^2\right). \tag{4}$$

We shall prove this function has an global maximum. To do this, we will first prove that $-l(\theta|\boldsymbol{x})$ is a coercive function, that is, for each sequence $\{\theta_n\}_{n\in\mathbb{N}}$, such that $|\theta_n| \to +\infty$, we have that $-l(\theta|\boldsymbol{x}) \to \infty$. By triangular inequality, for $i = 1, ..., n$,

$$|x_i - \theta| \geq |\theta| - |x_i| \implies 1 + |x_i - \theta|^2 \geq 1 + (|\theta| - |x_i|)^2 \implies \log(1 + |x_i - \theta|^2) \geq \log(1 + (|\theta| - |x_i|)^2)$$

if $|\theta| > |x_i|$. Therefore, if $|\theta_n| \to +\infty$, taking n largely enough, by the relation above,

$$\log(1 + |x_i - \theta|^2) \to +\infty,$$

what proves that $-l(\theta|\boldsymbol{x})$ is coercive. Given that it is continuos, we know there is a global minimum $\bar{\theta}$ (Alexey Izmailov, 2020, Corolary 1.2.8), and $l(\theta|\boldsymbol{x})$ has a global maximum. In order to find the maximum likelihood estimator (MLE) $\hat{\theta}(\boldsymbol{x})$, we find the stationary points for the unrestricted problem:

$$\frac{d}{d\theta} l(\theta|\boldsymbol{x}) = 2\sum_{i=1}^{n} \frac{x_i - \theta}{1 + (x_i - \theta)^2} = 0. \tag{5}$$

A numerical procedure must be used to solve 5. We search the solution in the interval

$$I = (\min\{x_i\}_{i=1}^n, \max\{x_i\}_{i=1}^n),$$

because taking $i = 1, ..., n$,

- if $\theta = \min\{x_i\}_{i=1}^n, x_i - \theta \geq 0$ and $l'(\theta|\boldsymbol{x}) \geq 0$;
- if $\theta = \max\{x_i\}_{i=1}^n, x_i - \theta \leq 0$ and $l'(\theta|\boldsymbol{x}) \leq 0$.

what implies the existence of $\theta^* \in I$ such that $l'(\theta^*|\boldsymbol{x}) = 0$, by the Intermediate Value Theorem (Lima, 1976, Theorem 12, Page 184). We use the Brent's method Brent (2002) to solve the problem and its implementation in SciPy Virtanen (2020). The codes are available at Github[1]Moschen (2021). When $n = 10$ and the data is 1, teh solution is

$$\hat{\theta}(\boldsymbol{x}) \approx 4.531$$

and the function can is drawn in Figure 1

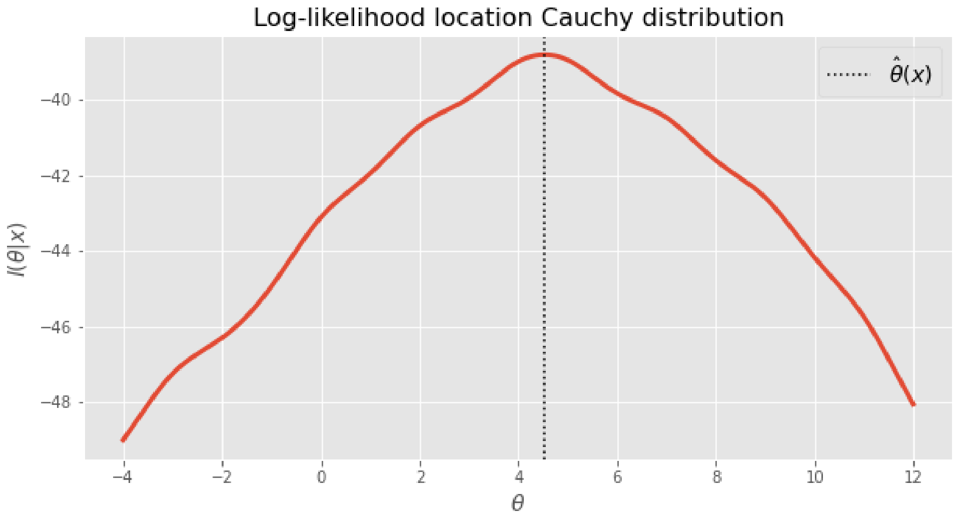

Figure 1: Log-likelihood of the parameter of location from the Cauchy distribution and MLE.

We differentiate once more

$$\frac{d^2}{d\theta^2} l(\theta|\boldsymbol{x}) = 2\sum_{i=1}^{n} \frac{-1 - (x_i - \theta)^2 + 2(x_i - \theta)^2}{(1 + (x_i - \theta)^2)^2} = 2\sum_{i=1}^{n} \frac{(x_i - \theta)^2 - 1}{((x_i - \theta)^2 + 1)^2} \tag{6}$$

and we obtain l''. We can test if the MLE is really the maximizer, that is, $l''\left(\hat{\theta}(\boldsymbol{x})\right) < 0$. In fact,

$$l''\left(\hat{\theta}(\boldsymbol{x})\right) = -1.231 < 0.$$

[1] https://github.com/lucasmoschen/phd-bayesian-statistics

Fisher information

The fisher information for one observation is defined as

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2}l(\theta|x)\right] = -2\,\mathbb{E}_\theta\left[\frac{(x-\theta)^2-1}{((x-\theta)^2+1)^2}\right]. \tag{7}$$

Therefore

$$\begin{aligned} I(\theta) &= 2\,\mathbb{E}_\theta\left[\frac{1}{((x-\theta)^2+1)^2} - \frac{(x-\theta)^2}{((x-\theta)^2+1)^2}\right] \\ &= 2\,\mathbb{E}_\theta\left[\frac{1}{(x-\theta)^2+1} - \frac{2(x-\theta)^2}{((x-\theta)^2+1)^2}\right] \\ &= 2\int_{\mathbb{R}} \frac{1}{\pi((x-\theta)^2+1)^2} - \frac{2(x-\theta)^2}{\pi((x-\theta)^2+1)^3}\,dx, [y = x-\theta, dy = dx] \\ &= \frac{2}{\pi}\int_{\mathbb{R}} \frac{1}{(y^2+1)^2} - \frac{2y^2}{(y^2+1)^3}\,dy. \end{aligned} \tag{8}$$

We have that

$$-\frac{2y^2}{(y^2+1)^3} = \frac{A}{(y^2+1)^3} + \frac{B}{(y^2+1)^2} \implies A + B(y^2+1) = -2y^2 \implies B = -2, A = 2.$$

Then,

$$\begin{aligned} I(\theta) &= \frac{2}{\pi}\int_{\mathbb{R}} \frac{1}{(y^2+1)^2} - \frac{2y^2}{(y^2+1)^3}\,dy, \\ &= \frac{2}{\pi}\int_{\mathbb{R}} \frac{1}{(y^2+1)^2} - \frac{2}{(y^2+1)^2} + \frac{2}{(y^2+1)^3}\,dy, \\ &= \frac{2}{\pi}\int_{\mathbb{R}} \frac{-1}{(y^2+1)^2} + \frac{2}{(y^2+1)^3}\,dy, \\ &= \frac{2}{\pi}(-S_2 + 2S_3), \end{aligned} \tag{9}$$

where

$$S_k = \int_{\mathbb{R}} \frac{1}{(y^2+1)^k}\,dy.$$

Integrating by parts, we obtain

$$\int_{\mathbb{R}} \frac{1}{(y^2+1)^k}\,dy = \frac{y}{(y^2+1)^k}\bigg|_{-\infty}^{\infty} + \int_{\mathbb{R}} \frac{2ky^2}{(y^2+1)^{k+1}}\,dy = \int_{\mathbb{R}} \frac{2k}{(y^2+1)^k} - \frac{2k}{(y^2+1)^{k+1}}\,dy,$$

since the first expression is 0 whenever $k > 1/2$ and the second is calculated in the same way as before. We conclude that

$$S_k = 2kS_k - 2kS_{k+1} \implies S_{k+1} = \frac{2k-1}{2k}S_k,$$

where $k \geq 1$ and $S_1 = \arctan(y)\big|_{-\infty}^{\infty} = \pi$. Following this, we obtain that

$$S_2 = \frac{1}{2}\pi \text{ and } S_3 = \frac{3}{4}\frac{1}{2}\pi = \frac{3}{8}\pi.$$

The Fisher Information for one observation for the location parameter of the Cauchy distribution is

$$I(\theta) = \frac{2}{\pi}\left(-\frac{1}{2}\pi + \frac{3}{4}\pi\right) = \frac{1}{2}. \tag{10}$$

For n observations,

$$I_n(\theta) = -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2}l(\theta|\boldsymbol{x})\right] = -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2}\sum_{i=1}^n l(\theta|x_i)\right] = \sum_{i=1}^n -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2}l(\theta|x_i)\right] = nI(\theta) = \frac{n}{2}. \tag{11}$$

General Regularity Conditions

Here we shall verify if there is mathematical support for the convergence to the Normal distribution of the posteriori. In order to do that, we shall first verify the General Regularity Conditions (Schervish, 1996, Page 436). The first four points are straightforward:

1. The parameter space is $\Omega = \mathbb{R}$, then it is finite.
2. Since Ω is an open set, its interior is itself. Therefore, $\theta \in \Omega = \text{int } \Omega$.
3. We will elicit the priors in the next section, with support on Ω and continuos everywhere.
4. The log-likelihood, equation 4, is infinitely differentiable in $\mathbb{R}$ because it is a composition of infinitely differentiable functions and it is defined everywhere. The second derivative is explicit in equation 6 and it is clearly continuos.

Although in this example the MLE is unique and the unique local maximum, this is not always the case. For instance, consider an example from (Young and Smith, 2005, Page 122) where we have X_1 and X_2 iid distributed from $Cauchy(\theta, 1)$. Let $X_1 = a$ and $X_2 = b$. Therefore, a local maximum is solution to the equation

$$\frac{a-\theta}{1+(a-\theta)^2} + \frac{b-\theta}{1+(b-\theta)^2} = 0,$$

as we indicated in equation 5. Simplifying the system and factoring, we obtain

$$(a+b-2\theta)(\theta^2 - (a+b)\theta + 1 + ab) = 0.$$

If $\Delta = (a+b)^2 - 4(1+ab) = (a-b)^2 - 4 > 0$, that is, $|X_1 - X_2| > 2$, we have three solutions,

$$\theta_m = \frac{X_1 + X_2}{2}$$

and the solutions of the polynomial of second order. The bad news is that, if θ_1 and θ_2 are the solutions,

$$\frac{\theta_1 + \theta_2}{2} = \frac{X_1 + X_2}{2} = \theta_m$$

and

$$l''(\theta_m | X_1, X_2) = \frac{1}{2}\frac{(X_1 - X_2)^2 - 4}{((X_1 - \theta_m)^2 + 1)^2} + \frac{1}{2}\frac{(X_1 - X_2)^2 - 4}{((X_2 - \theta_m)^2 + 1)^2} > 0$$

what implies we have a local minimum here. By the expression's symmetry 4, we conclude that we have two possible values for the MLE with positive probability. According to Young and Smith (2005), this problem appears with positive probability when n tends to infinity. This problem complicates the demonstration of consistency of the MLE, but Bai and Fu (1987) proves that $\hat{\theta}(\boldsymbol{x})$ converges to the true value θ_0 exponentially. By the continuity $l''(\cdot|x)$, the consistency of the MLE and the Continuos mapping theorem, we know that

$$l''(\hat{\theta}(\boldsymbol{x})|\boldsymbol{x}) \xrightarrow{p} l''(\theta|\boldsymbol{x}).$$

and

$$l''^{-1}(\hat{\theta}(\boldsymbol{x})|\boldsymbol{x}) \xrightarrow{p} l''^{-1}(\theta|\boldsymbol{x}).$$

However, we were not able to prove that

$$-l''^{-1}(\hat{\theta}(\boldsymbol{x})|\boldsymbol{x}) \xrightarrow{p} 0.$$

We believe this is the case because $-\mathbb{E}[l''(\theta|\boldsymbol{x})] = \frac{n}{2} \to \infty$. We could not also prove the Conditions 6 and 7. The condition 6 can be seen similar to the consistency, so we also believe it is true.

Normal approximation

Following (Schervish, 1996, Theorem 7.89), let

$$\Sigma_n = -l''^{-1}(\hat{\theta}(\boldsymbol{x})|\boldsymbol{x})$$

and the posterior density of $\Psi_n = \Sigma_n^{-1/2}(\theta - \hat{\theta}(\boldsymbol{x}))$ converges (in probability uniformly on compact sets) to the standard Normal density. In other words,

$$\frac{\theta - \hat{\theta}(\boldsymbol{x})}{\Sigma_n^{1/2}} \sim \mathcal{N}(0,1)$$

what implies that the parameters of the normal approximation are

$$\mu = \hat{\theta}(\boldsymbol{x}) \approx 4.531 \tag{12}$$

and

$$\sigma^2 = \Sigma_n \approx 1/1.231 \approx 0.812. \tag{13}$$

Eliciting the priors

In this section, we elicit two different priors to compare the results in the next sections.

Non-informative prior

We use the Jeffreys' Prior which has the property of invariance to reparametrization. Although this prior does not satisfy the likelihood principle, it is good when no subjective information is given before the data. We have already calculated the Fisher information before, then the prior is

$$\pi(\theta) \propto I^{1/2}(\theta) \propto 1. \tag{14}$$

We obtain a improper prior. Therefore, we need to prove the posterior is proper almost surely.

The posterior will be

$$p(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta) \propto \prod_{i=1}^{n} \frac{1}{1+(x_i-\theta)^2}$$

Observe that, $\forall \theta \in \mathbb{R}$,

$$0 < \frac{1}{1+(x_i-\theta)^2} \leq 1.$$

Then,

$$\prod_{i=1}^{n} \frac{1}{1+(x_i-\theta)^2} \leq \frac{1}{1+(x_1-\theta)^2}$$

and

$$\int_\Omega \prod_{i=1}^{n} \frac{1}{1+(x_i-\theta)^2}\, d\theta \leq \int_\Omega \frac{1}{1+(x_1-\theta)^2}\, d\theta = \pi,$$

what proves the posteriori is proper almost surely, even though its expression is really difficult.

Informative prior

We already know that $\theta \in \mathbb{R}$ and we need a distribution over this space. There are infinitely many options. We will consider two approaches.

Maximum entropy prior

Prior to the data, we have no knowledge about the position of the parameter. In this approach, we define

$$\mathbb{E}^\pi[\theta] = 0 \text{ and } \operatorname{Var}^\pi[\theta] = \sigma^2$$

and we use the Lebesgue measure as the reference measure $\pi_0(\theta)$. As calculated in (Robert, 2007, Example 3.2.4), the prior will have Normal distribution with parameters 0 and σ^2. We will compare small values (strong prior) of σ^2 to large values (weak prior). In particular, we will compare the values

$$\sigma^2 = 1, 100 \text{ and } 10000.$$

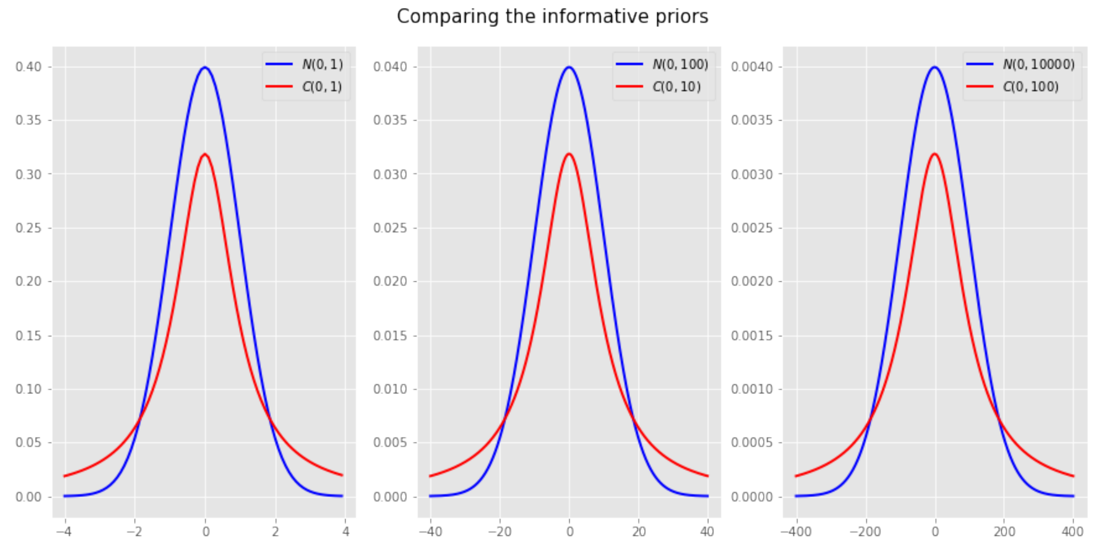

Figure 2: Comparing the distributions defined above.In blue, we see the Normal distribution, and in red we see the Cauchy distribution.

Heavy tail prior

The Normal distribution has a light tail and decreases exponentially to both sides. It shall be interesting to analyse when the tail is heavier. We choose the Cauchy distribution with location parameter 0 and scale parameter γ. Here the mean and variance are undefined, so it's harder to compare them with the Normal distribution with moment matching. However γ is a scale parameter as $\sigma = \sqrt{\sigma^2}$ in the Normal distribution. Therefore, we will compare the values

$$\gamma = 1, 10 \text{ and } 100$$

with the same interpretation as the Normal distribution. Figure 2 presents the comparison of the priors.

Posterior distribution in the suggested case

Considering the data given and the priors elicited previously, we can obtain the posterior distribution through a MCMC routine. In this case, we use PyMC3 library (Salvatier et al. (2016)) in the Python language. We sampled from this procedure 5000 draws and 1000 tune iterations. After simulating with each prior, we get $\Pr(\theta \geq 5 | X = \boldsymbol{x})$ and compare with the one given by Schervish (1996). The results are in the Table 1. We can see that the most similar occurs when the prior is the lebesgue measure in the real line and, for that reason, we can reproduce a similar figure.

Prior	$U(-\infty, +\infty)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 100)$	$\mathcal{N}(0, 10000)$	$C(0, 1)$	$C(0, 10)$	$C(0, 100)$
Probability	0.3569	0.0	0.3251	0.3493	0.1437	0.3022	0.3485

Table 1: Comparative among priors and its posteriors probabilities inferred.

Thereby, we can compare the posterior obtained by numerical integration with the theoretical Normal distribution with mean and variance specified in equations 12 and 13. The comparison result can be seen in Figure 3 and in Table 3. The summary given by PyMC3 library is in Table 2. We also plot the other posteriors generated by the other priors in Figure 4.

	Parameter				MCSE		ESS				
	Mean	Sd	HDI_3%	HDI_97%	Mean	Sd	Mean	Sd	Bulk	Tail	$\hat{R}$
θ	4.583	1.498	1.608	7.508	0.024	0.017	3831.0	3728.0	3926.0	4447.0	1.0

Table 2: Summary MCMC routine and some statistics.

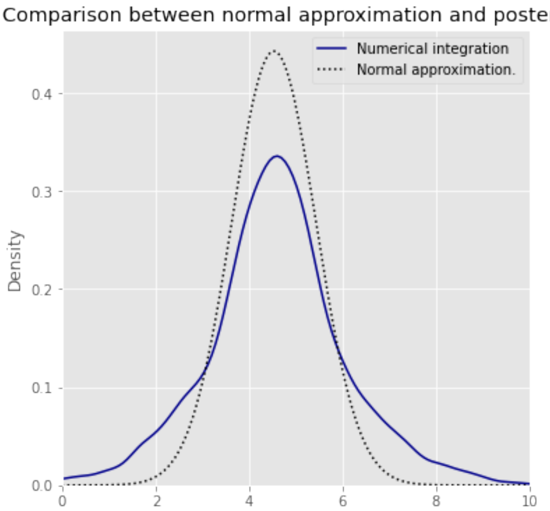

Figure 3: Posterior density for θ in Cauchy example and normal approximation.

	Normal approx.	**Posterior MCMC**
median	4.53136	4.57545
2.5th per	2.76499	1.55876
25th per	3.92349	3.76738
75th per	5.13923	5.38569
0.975th per	6.29773	7.79505
P(X>=mean(x))	0.557943	0.5598

Table 3: Comparison quartiles and probabilities from the distributions.

We observe that strong priors have results very different when compared to the normal approximation, since more data is needed to convince the distribution to move. In particular, we when the prior is $\mathcal{N}(0, 1)$, we are claiming that $\Pr(\theta \geq 2) < 0.03$. However, the difference to the normal approximation does not indicate a problem, because the sample can be too small yet, especially when the prior is too strong. When the prior is weaker, we see less difference in the distribution choice.

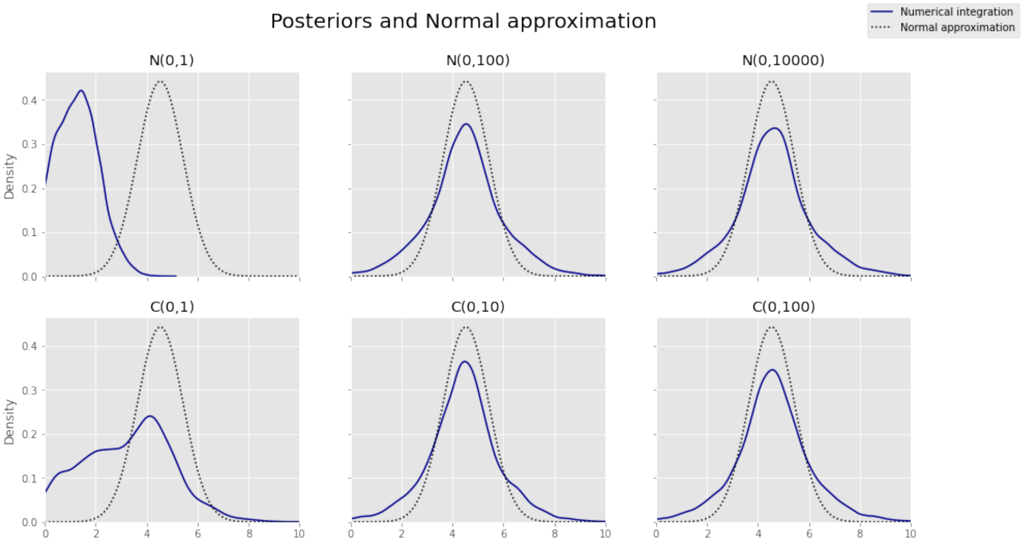

Figure 4: Posterior density for θ in Cauchy example and normal approximation for other priors.

Sampling and asymptotic behavior

We generate samples from Cauchy distribution with location parameter 5 and scale parameter 1 with sizes specified previously. After that, the parameters of the Normal distribution are calculated using the principles highted in MLE and normal approximation sections. They can be visualized in Table 4. We see that there is a consistency in the MLE observed because it seems to converge to 5, what is a good news. The scale parameter is vanishing apparently.

Parameter/n	**20**	**50**	**100**	**500**	**1000**	**10000**
Location	5,4	4,67	5,05	4,98	5,01	5
Scale	0,38	0,21	0,14	0,06	0,04	0,01

Table 4: Parameters of the normal approximation calculated with precision two.

From these values, we can compare the distribution of

$$\Sigma_n^{-1/2}(\theta - \hat{\theta}(\boldsymbol{x}))$$

given $X = \boldsymbol{x}$ to the Normal distribution with mean 0 and variance 1. We do not compare the posterior distribution of θ directly, because we know the posterior tightens on the true value. We use the Flat prior because it seems to be the chosen by Schervish in the example. We will test the other prior distributions posteriorly. The results are showed in Figure 5 and they appear to be very interesting. Even when $n = 20$, we obtain a good approximation.

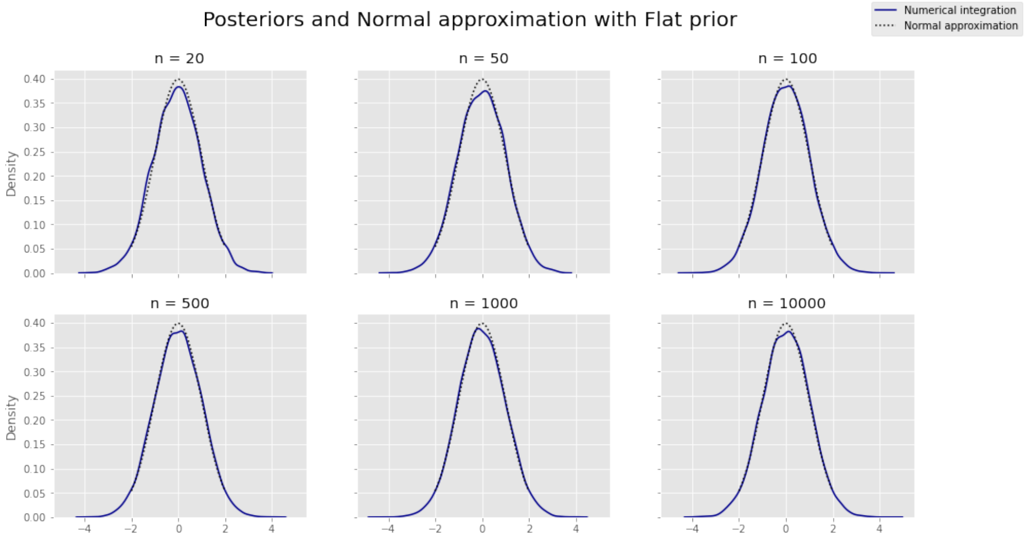

Figure 5: Posterior density for θ in Cauchy example and normal approximation for larger samples.

Bibliography

Alexey Izmailov, M. S. (2020). *Otimização: Condições de otilmalidade, elementos de análise convexa e de dualidade, vol. 1*. Coleção matemática universitária.

Bai, Z. D. and Fu, J. C. (1987). On the maximum-likelihood estimator for the location parameter of a cauchy distribution. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 15(2):137–146.

Brent, R. (2002). Algorithms for minimization without derivatives. *Englewood Cliffs, Prentice Hall*, 19.

Lima, E. L. (1976). *Curso de análise volume 1*. IMPA.

Moschen, L. M. (2021). Github's repository. https://github.com/lucasmoschen/phd-bayesian-statistics.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Salvatier, J., Wiecki, T., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*.

Schervish, M. (1996). *Theory of Statistics*. Springer Series in Statistics. Springer New York.

Virtanen, P. e. a. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.