Non Informative Prior Distributions

Antonietta Mira
Swiss Finance Institute,
University of Lugano, Switzerland
jointly with
Filippo Macaluso
PhD Student,
University of Lugano, Switzerland

Motivation

An important problem in Bayesian statistics is how to define the prior distribution. Suppose we do not have any prior information about parameter $\theta \in \mathbb{R}$, then we want a prior which minimize its influence on the inference.

First attempt

- Uniform distribution.

Issue

- It not invariant under reparametrization
- It does not integrate to one if $\theta \in \mathbb{R}$ (improper prior)
- In higher dimensions the flat prior becomes really informative, indeed most of the probability mass lies at infinity

Example

- Example 1

Assume we do not have any information on $\theta \in [0,1]$ and thus we put a uniform prior on the interval $[0,1]$

$$\pi(\theta) \propto 1$$

Let now $\phi = \frac{1}{\theta} = k(\theta)$. Since this is a one-to-one transformation of θ we would like that the implied prior on ϕ is also not informative. To the contrary and to our surprise, by the change of variable formula, we obtain:

$$\pi(\phi) = \mid \frac{dk^{-1}(\phi)}{d\phi} \mid = \mid -\frac{1}{\phi^2} \mid = \theta^2$$

Thus the resulting prior on ϕ is not flat anymore.
Recall that:

$$\pi(\phi) = \pi_\theta(k^{-1}(\phi)) \times \mid \frac{dk^{-1}(\phi)}{d\phi} \mid$$

where $k(\cdot)$ is a one-to-one function

- Example 2

In the same setting as before consider a different one-to-one transformation of the original parameter, namely: $\phi = \log \frac{\theta}{1-\theta}$.
Applying the change of variable formula, we get:

$$\pi(\phi) = \frac{e^{\phi}}{(1+e^{\phi})^2}$$

Again the induced prior is not flat anymore: it becomes informative

Next step: we look for a prior which is invariant under reparametrizations

Jeffreys Priors

IDEA: Suppose to have a model for our data that induces the likelihood $L(x|\theta)$ and we want to extract from it a prior on θ.
This approach is not coherent with the *subjective* Bayesian approach since the prior is ,in some sense, derived from the likelihood.
The Jeffreys prior , in the univariate case, is defined as:

$$\pi^J(\theta) = \mathrm{I}(\theta)^{\frac{1}{2}}$$

where I is the Fisher Information defined as:

$$\mathrm{I}(\theta) = -\mathbb{E}_\theta \left[\frac{d^2 \log L(X|\theta)}{d\theta^2} \right]$$

Examples

- Example 1

Let $X_1, \ldots, X_n$ be i.i.d Poisson distributed r.v. with parameter λ. The loglikelihood is:

$$\ell(x|\theta) \propto -n\lambda + \sum_{i=1}^{n} x_i \log \lambda$$

$$\begin{aligned} \mathrm{I}(\lambda) &= -\mathbb{E}_\lambda \left[-\frac{\sum_{i=1}^n x_i}{\lambda^2} \right] = \frac{n}{\lambda} \\ \pi^J(\lambda) &\propto \frac{1}{\sqrt{\lambda}} \end{aligned}$$

an the posterior distribution is

$$\pi(\lambda|x) \propto \pi^J(\lambda) L(x|\lambda) = \frac{1}{\sqrt{\lambda}} e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} = \lambda^{\left(\sum_{i=1}^n x_i - \frac{1}{2}\right)} e^{-n\lambda}$$

thus $\lambda|x \sim Gamma(n, \sum_{i=1}^n x_i + \frac{1}{2})$

We recall that the posterior distribution with the coniugate $Gamma(\alpha, \beta)$ prior is:

$$\lambda | x \sim Gamma(\alpha + n, \sum_{i=1}^{n} x_i + \beta)$$

We now prove (for the unidimensional case) that the Jeffreys priors invariant. Assume $X \sim f(x|\theta)$, $\theta \in \Theta$ and $\phi = k(\theta)$, where k is a differentiable monotone function.
We have that, under the previous transformation, $X \sim f^*(X|k^{-1}(\phi))$ then

$$\begin{aligned}
\mathrm{I}(\theta) &= \mathbb{E}_\theta \left[\left(\frac{\partial \log L(x|\theta)}{\partial \theta} \right)^2 \right] \\
&= \mathbb{E}_\phi \left[\left(\frac{\partial \log L^*(x|\phi)}{\partial \theta} \right)^2 \right] \\
&= \mathbb{E}_\phi \left[\left(\frac{\partial \log L^*(x|\phi)}{\partial \phi} \frac{d\phi}{d\theta} \right)^2 \right] \\
&= k'(\theta)^2 \mathrm{I}(\phi)
\end{aligned}$$

The proof is based on the differentiation chain rule. From the above identity, taking the square-root on both sides, we get:

$$\pi^J(\phi) \propto \frac{\pi^J(\theta)}{k'(\theta)}$$

Limitation of Jeffreys Prior

This kind of priors do not work well in multidimensional space parameter. In this setting, we have

$$\pi^J(\theta) = |\mathrm{I}(\theta)|^{\frac{1}{2}}$$

where $|\cdot|$ denotes the determinant.

- Example 1

Assume $X \sim N(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)^T$. The Fisher Information matrix is equal to

$$\mathrm{I}(\theta) = -\mathbb{E}_\theta \begin{bmatrix} \frac{1}{\sigma^2} & \frac{2(X-\mu)}{\sigma^2} \\ \frac{2(X-\mu)}{\sigma^2} & \frac{3}{\sigma^2}(X-\mu)^2 - \frac{1}{\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{bmatrix}$$

Then

$$\pi^J(\theta) = |\mathrm{I}(\theta)|^{\frac{1}{2}} = \frac{1}{\sigma^2}$$

The posterior of σ is such that

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2} \sim \chi_n^2$$

It is not acceptable since we would expect to loose one degree of freedom when we estimate μ
Jefreey suggested to assume μ and σ independet aprior and use one dimensional Jeffreys prior for each parameters

- Example 2

We want to estimate $||\theta||^2$ when $X \sim N(\theta, \mathrm{I})$ d-dimensional normal vector. Jeffreys prior of θ is flat and the posterior is a non central chi-squared distribution with d degree of freedom. Then the posterior expected value, given one observation,is

$$\mathbb{E}[||\theta||^2|X] = ||X||^2 + d$$

Note that it is not a good estimate since it adds d whereas we could want to shrink our estimate towards zero. Moreover, the minimum variance frequentist estimate is $||X||^2 - d$

Reference priors

this class of priors has been introduced by Bernardo (1979) and developed in Berger and Bernardo (1989).
Let $X_1, \ldots, X_n$ be i.i.d rvs, with sufficient statistic $T = T(X)$. The Reference prior is defined as a function that maximizes some distance or divergence measure between the prior and posterior, given a set of observations.

- Hellinger distance

$$\int_{\Theta} \left(\sqrt{p(\theta|x)} - \sqrt{p(\theta)} \right)^2 d\theta$$

- Kullback-Leibler divergence

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

Properties of the Kullback-Leibler divergence

- $\mathrm{K}(p(\theta|x), p(\theta)) \geq 0$
- $\mathrm{K}(p(\theta|x), p(\theta)) = 0$ iff $p(\theta|x) = p(\theta)$
- $\mathrm{K}(p(\theta|x), p(\theta)) \neq \mathrm{K}(p(\theta), p(\theta|x))$ i.e. it is not symmetric
- it does not satisfy the triangle inequality

then it is clearly not a norm

The goal is to find a reference prior $\pi(\theta)$ which maximize the Kullback-Leibler divergence averaged over the distribution, $p(t)$, of the sufficient statistic T.
Define the mutual information as:

$$\begin{aligned} I(p(\theta), t) &= \int p(t) \int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta dt \\ &= \int \int p(\theta, t) \log \frac{p(\theta, t)}{p(\theta)p(t)} d\theta dt \end{aligned}$$

Strictly speaking, it is a measure of the amount of information one random variable contains about another

$$\pi(\theta) = arg \max_{p(\theta)} I(p(\theta), t)$$

Since the mutual information is invariant under one-to-one transformations, the resulting reference prior, as defined above, is also invariant.

Note that we are taking the *arg* max w.r.t $p(\theta)$, then to solve this optimization problem we need calculus of variations.

Calculus of variations: an hint

Calculus of variations is the infinite dimensional counterpart when we are dealing with optimization problem. In this case, we maximize of minimize *functionals*. These mathematical objects are mapping from a set of functions, say Ω, to the real numbers.

For example, integrals and norms are functionals:

- $f \longmapsto A[f] = \int_\Omega H(f(x))dx$
- $f \longmapsto \left(\int |f|^p dx\right)^{\frac{1}{p}}$ (L^p norm)

How to solve this optimization problem?

Solution of the problem is given by a function s.t. the functional derivatives is equal to zero, where the functional derivatives of $A[x_0]$ is defined as:

$$\lim_{\epsilon \to 0} \frac{A[x_0 + \epsilon] - A[x_0]}{\epsilon}$$

To get this point we need to solve the associated *Euler-Lagrange equation*

Recalling the total derivative formula, let $L = L(x, \epsilon y, \epsilon z)$ then

$$\frac{dL}{d\epsilon} = \frac{\partial L}{\partial y}\frac{dy}{d\epsilon} + \frac{\partial L}{\partial z}\frac{dz}{d\epsilon} = \frac{\partial L}{\partial y}y + \frac{\partial L}{\partial z}z$$

EULER-LAGRANGE EQUATION

Let $f_\epsilon = f(x) + \epsilon\eta(x)$ be the variation of the function f due to $\epsilon\eta(x)$ where ϵ is small and $\eta(x)$ is a differentiable function such that $\eta(a) = \eta(b) = 0$, namely, it is zero at the boundary . Let $L = L\left(x, f_\epsilon = f + \epsilon\eta, f'_\epsilon = \frac{df}{dx} + \frac{d\eta}{dx}\epsilon\right)$ and the functional

$$A = \int_a^b L\left(x, f_\epsilon = f + \epsilon\eta, f'_\epsilon = \frac{df}{dx} + \frac{d\eta}{dx}\epsilon\right) dx$$

then

$$\begin{aligned}
\frac{dA}{d\epsilon} &= \frac{d}{d\epsilon}\int_a^b L\left(x, f_\epsilon = f + \epsilon\eta, f'_\epsilon = \frac{df}{dx} + \frac{d\eta}{dx}\epsilon\right) dx \\
&= \int_a^b \frac{d}{d\epsilon}\left[L\left(x, f_\epsilon = f + \epsilon\eta, f'_\epsilon = \frac{df}{dx} + \frac{d\eta}{dx}\epsilon\right)\right] dx
\end{aligned}$$

The total derivative of $\frac{dL}{d\epsilon}$ is equal to

$$\begin{aligned}
\frac{dL}{d\epsilon} &= \frac{\partial L}{\partial x}\frac{dx}{d\epsilon} + \frac{\partial L}{\partial f_\epsilon}\frac{df_\epsilon}{d\epsilon} + \frac{\partial L}{\partial f'_\epsilon}\frac{df'_\epsilon}{d\epsilon} \\
&= \frac{\partial L}{\partial f_\epsilon}\frac{df_\epsilon}{d\epsilon} + \frac{\partial L}{\partial f'_\epsilon}\frac{df'_\epsilon}{d\epsilon} \\
&= \frac{\partial L}{\partial f_\epsilon}\eta(x) + \frac{\partial L}{\partial f'_\epsilon}\eta'(x)
\end{aligned}$$

thus

$$\frac{dA}{d\epsilon} = \int_a^b \left[\frac{\partial L}{\partial f_\epsilon}\eta(x) + \frac{\partial L}{\partial f'_\epsilon}\eta'(x)\right] dx$$

To simplify, we assume $\epsilon = 0$, then $f_\epsilon = f$ and $f'_\epsilon = f'$, so the last equation becomes:

$$\begin{aligned} &= \int_a^b \left[\frac{\partial L}{\partial f}\eta + \frac{\partial L}{\partial f'}\eta'\right] dx \\ &= \int_a^b \frac{\partial L}{\partial f}\eta dx + \int_a^b \frac{\partial L}{\partial f'}\eta' dx \\ &= \int_a^b \frac{\partial L}{\partial f}\eta dx + \left[\frac{\partial L}{\partial f'}\eta\right]_a^b - \int_a^b \eta \frac{d}{dx}\frac{\partial L}{\partial f'} dx \end{aligned}$$

Since $\left[\frac{\partial L}{\partial f'}\eta'\right]_a^b = 0$ for the boundary property of $\eta(x)$, then $\frac{dA}{d\epsilon} = 0$ if

$$\begin{aligned} \int_a^b \eta(x)\left[\frac{\partial L}{\partial f} - \frac{d}{dx}\frac{\partial L}{\partial f'}\right] dx &= 0 \\ \iff \frac{\partial L}{\partial f} - \frac{d}{dx}\frac{\partial L}{\partial f'} &= 0 \end{aligned}$$

The last result is given by the *fundamental lemma of calculus of variations*

Reference Analysis: Motivation

The declared objective of reference Bayesian analysis is to specify a prior distribution such that, even for moderate sample sizes, the information provided by the data should dominate the prior information because of the vague nature of the prior knowledge. (Bernardo and Ramon, 1998)

- the amount of information expected from an experiment depends on the available prior knowledge: the more prior information available, the less information may be expected to be obtained from the data
- An infinitely large experiment would provide all missing information, hence it is, theoretically, possible to obtain a measure of the amount of missing information as a limit of a functional of the prior distribution
- Then, we can define vague prior knowledge as that with the largest missing information

Univariate Reference Distributions

Recalling the mutual information:

$$I(p(\theta), t) = \int p(t) \int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta dt$$

this quantity can be interpreted as the average (expected information) of the K-L divergence w.r.t the marginal distribution of $p(t)$ of T
It is the amount of information about θ which t may be expected to provide

ISSUE

- Intractability

SOLUTION

- Asymptotics

The idea is not to observe a single experiment which gives rise to the sufficient statistics t, but we image to repeat m *independent* experiments which give rise to the following vector of sufficient statistics:
$z^m = (t^1, \ldots\ldots, t^m)$.
Note that the experiments are conditional on θ which remains the same throughout this hypothetical exercise. Then the corresponding expected infomation from these hypothetical experiments is

$$I(p(\theta), z^m) = \int p(z^m) \int p(\theta|z^m) \log \frac{p(\theta|z^m)}{p(\theta)} d\theta dz^m$$

and the expected value of the perfect information on θ is

$$\lim_{m \to \infty} I(\theta, z^m)$$

Note that we need to prove that such limit exists and this quantity measures the *missing information* on θ as a function of the prior $p(\theta)$

In this setting the reference prior $\pi(\theta)$ is the prior that maximize the missing information functional and given t the posterior is simply

$$\pi(\theta|t) \propto p(t|\theta)\pi(\theta)$$

ISSUE

- the limit is infinite (it is not if θ can only take a finite range of values)
- intractability

SOLUTION

- Derive a sequence of π_m which maximize $I(\cdot)$ and then $\pi = \lim_{m\to\infty} \pi_m$

The mutual information can be rewritten as:

$$I^m(p(\theta), z^m) = \int_\Theta p(\theta) \log \frac{f_m(\theta)}{p(\theta)} d\theta$$

where

$$f_m(\theta) = \exp\left(\int p(z^m|\theta) \log p(\theta|z^m) dz^m \right)$$

and

$$p(\theta|z^m) = \prod_{j=1}^{m} p(t^j|\theta) p(\theta)$$

is the posterior for θ as z^m has been observed. We also have this constraint $\forall p(\theta)$

$$\int_\Theta p(\theta) d\theta = 1$$

then the m-prior distribution $\pi_m(\theta)$ that maximize the above quantity is the maximum of the functional

$$A[p(\theta)] = \int_\Theta p(\theta) \log \frac{f_m(\theta)}{p(\theta)} d\theta + \lambda \left[\int_\Theta p(\theta) d\theta - 1 \right]$$

as we have seen in the previous slides we need to find $p(\theta)$ s.t.

$$\frac{\partial}{\partial \epsilon} A[p(\theta) + \epsilon \eta(\theta)]|_{\epsilon=0} = 0$$

and after some computation we get

$$\int_\Theta \eta(\theta) \left[\log f_m(\theta) - \log p(\theta) + \lambda \right] d\theta = 0$$

$\Longrightarrow$ by the fundamental lemma of calculus of variations

$$\log f_m(\theta) - \log p(\theta) + \lambda = 0$$

so

$$p(\theta) \propto \ f_m(\theta)$$

Remarks

- This approach provides an *implicit* solution for the prior which maximize the mutual information $I^m(\cdot)$. Indeed, f_m depends on the prior through the posterior
- if $m \to \infty$ we find an approximation $p^*(\theta|t^m)$ of the posterior which is independent of the prior $p(\theta)$. Then under some regularity conditions

$$p_m^*(\theta) = \exp\left(\int p(t^m|\theta) \log p^*(\theta|t^m) dt^m\right)$$

where $p^*(\theta|t^m) = \frac{\prod_{j=1}^m p(t^j|\theta)}{\int_\Theta \prod_{j=1}^m p(t^j|\theta) d\theta}$ and then by Bayes Theorem

$$\pi_m(\theta|t) \propto p(t|\theta) p_m^*(\theta)$$

- the reference posterior is given by the log-divergence limit of $\pi_m(\theta|t)$
- any positive function $\pi(\theta)$ s.t $\forall \theta \in \Theta$ and some $c(t) > 0$

$$\pi(\theta|t) = c(t) p(t|\theta) \pi(\theta)$$

is called reference prior for θ

- Invariance to transformations
- Independence of sample size
- Compatibility with sufficient statistics
- In the explicit form of the reference prior the reference is given by

$$\pi(\theta) = c \lim_{m \to \infty} \frac{f_m^*(\theta)}{f_m^*(\theta_0)}$$

for some $c > 0$ and $\theta_0 \in \Theta$

Suppose to observe a sequence $x_1, \ldots, x_n$, where the rv $X_i \sim U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathbb{R}$. Let $p(\theta)$ be the prior for θ.

It is well known that $t_n = [x_{\min}^{(n)}, x_{\max}^{(n)}]$ is the sufficient statistic for θ. Clearly $p(\theta|x) = p(\theta|t_n) \propto p(\theta)$ and $\theta \in [x_{\max} - \frac{1}{2}, x_{\min} - \frac{1}{2}]$

$$\begin{aligned} f_m^*(\theta) &= \exp\left(\int p(t^m|\theta) \log p^*(\theta|t^m) dt^m\right) \\ &= \exp\left\{\mathbb{E}_\theta\left[-\log\left(1 - \left(x_{\max}^{(mn)} - x_{\min}^{(mn)}\right)\right)\right]\right\} \end{aligned}$$

Noting that for large m the the expectation on the RHS can be approximated by

$$\log\left(1-\left(\mathbb{E}\left[x_{\max}^{(mn)}\right]-\mathbb{E}\left[x_{\min}^{(mn)}\right]\right)\right)$$

and that the distribution of $w = x_{\max}^{(mn)} - \theta - \frac{1}{2}$ and $u = x_{\min}^{(mn)} - \theta - \frac{1}{2}$ are, respectively, $Be(w|mn,1)$ and $Be(u|mn,1)$ then the previous quantity is equal to

$$-\log\left(1-\frac{mn}{mn+1}+\frac{1}{mn+1}\right)=\log\left(\frac{mn+1}{2}\right)$$

Thus $f_{mn}^{*} = \frac{mn+1}{2}$ and the explicit reference prior is

$$\pi(\theta) = c \lim_{m\to\infty} \frac{(mn+1)/2}{(mn+1)/2} = c$$

Moreover, the reference posterior $\pi(\theta|x) \propto c$

Applications

- Applications where we want to estimate the ratio $\frac{\theta_i}{\theta_j}$ of parameters of multinomial distribution.
- Insurance application: interest in assessing how many times more likely risk i is than risk j
- political application: ratio of the percentages of votes that candidates i and j may be expected to obtain