Sjournals
Scientific Journals
Scientific Journal of Review (2014) 3(8) 916-925
ISSN 2322-2433
doi: 10.14196/sjr.v3i8.1515

Contents lists available at Sjournals
Scientific Journal of REVIEW
Journal homepage: www.Sjournals.com

Review article

Prior selection: a review

G.H. Gholami[a,*], A. Etemadi[b]
Department of Mathematics, Faculty of science, Urmia University, Urmia, IRAN.
Department of Mathematics, Urmia Branch, Islamic Azad University, Urmia, IRAN.

*Corresponding author; Department of Mathematics, Faculty of science, Urmia University, Urmia, IRAN.

ARTICLE INFO

Article History:
Received 29 July 2014
Accepted 24 August 2014
Available Online 29 August 2014

Keywords:
Prior Distribution
Bayesian Inference
Selection
Drawing

ABSTRACT

The prior distribution is the key to Bayesian inference and its determination is therefore the most important step in drawing this inference. To some extent, it is also the most difficult. In this paper, we'll review different approaches of choosing prior distribution.

Introduction

Specification of the prior distribution of the parameter is obviously a key element of the Bayesian approach, and there are several schools of thought when it comes to assigning priors; these can be loosely categorized into subjective, objective, and empirical or hierarchical prior approaches.

Subjective and objective prior distributions

Subjective prior distributions

This approach seeks to consider the whole available information while making the inference. These information can be either in the form of available experimental data or in the form of the individual professions experience. It's too hard to collect such information through a density function but subjective

prior distributions can be determined in some family of distributions, such as exponential distributions family .These subjective prior distributions can be conjugate priors.

Definition 1

Prior distribution π (θ) is said to be conjugate (or closed under sampling) for f(x|θ) if, the posterior distribution π(θ|x) also belongs to family of π(θ).

Some of the conjugate distributions have been listed in Table 1 and Table 2 [7].

The advantages of applying conjugate priors

There are at least two advantages in using conjugate priors [8]:

- Applying them in Bayesian inference results in the integrals with analytical answers.
- If the computation of posterior distribution be possible, application of this distribution as prior distribution for next inferences would surely result in proper priors.

In general, there are four methods to determine subjective prior distributions [1].

- Relative likelihood approach
- Histogram approach
- Conformity with the given functional form
- Determination of cumulative distribution function

In the following, we'll explain the Relative likelihood approach with an example.

Table 1: Discrete likelihood distributions

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters
Bernoulli	p (probability)	Beta	α , β	$\alpha+\Sigma_{i=1}^{n}x_i, \beta+n-\Sigma_{i=1}^{n}x_i$
Binomial	p (probability)	Beta	α , β	$\alpha+\Sigma_{i=1}^{n}x_i, \beta+\Sigma_{i=1}^{n}N_i-\Sigma_{i=1}^{n}x_i$
Negative Binomial	p (probability)	Beta	α , β	$\alpha+rn, \beta+\Sigma_{i=1}^{n}x_i$
Poisson	λ (rate)	Gamma	k, θ	$k+\Sigma_{i=1}^{n}x_i, \frac{\theta}{n\theta+1}$
Multinomial	$\boldsymbol{p}$ (probability vector)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}+\Sigma_{i=1}^{n}x^{(i)}$
Geometric	p_0 (probability)	Beta	α , β	$\alpha+n, \beta+\Sigma_{i=1}^{n}x_i$

Table 2: Continuous likelihood distributions

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters
Uniform	$U(0,\theta)$	Pareto	x_m, k	$max\{x_{(n)}, x_m\}, k+n$
Exponential	λ (rate)	Gamma	α, β	$\alpha+n, \beta+\Sigma_{i=1}^{n}x_i$
Normal with known σ^2	μ (mean)	Normal	μ_0, σ_0^2	$\left(\frac{\mu_0}{\sigma_0^2}+\frac{\Sigma_{i=1}^{n}x_i}{\sigma^2}\right) \Big/ \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$

Relative likelihood approach

This approach has been often used when parameter space is a subset of real line. In this approach, we compare the likelihood of each point of different parameter space, and then, we directly use these functions for determining subjective prior distributions.

Example 1

Suppose that θ = [0, 1]. Let likelihood for points

$$\theta = 0, 0.2, 0.33, 0.5, 0.66, 0.8, 0.85, 0.9, 1$$

Be relative as below:

$$L(0|x) = L(1|x) = 0$$
$$L(0.2|x) = 0.32L(0.01|x)$$
$$L(0.33|x) = 0.75L(0.01|x)$$
$$L(0.5|x) = 1.26L(0.01|x)$$
$$L(0.66|x) = 1.5L(0.01|x)$$
$$L(0.8|x) = 1.29L(0.01|x)$$
$$L(0.85|x) = 1.09L(0.01|x)$$
$$L(0.9|x) = 0.82L(0.01|x)$$

Considering a constant value for L (0.01|x), for example L (0.01|x) = 1, we can draw corresponding likelihood graph as Figure 2.

Considering more points of parameter space and calculating corresponding likelihood function, likelihood graph will be more similar to curve. Now if we can guess graph's figure or if we can approximate it, then it can be used as prior distribution. In this example, this figure has been approximated as a coefficient of θ^2 (1- θ),

$$\pi(\theta) \propto \theta^2(1-\theta).$$

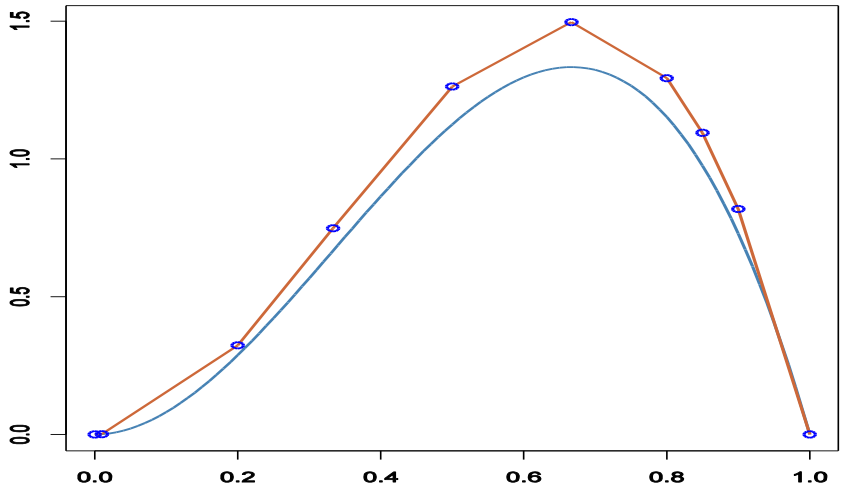

Fig.1. Likelihood graph (broken lines) and fitted prior distribution curve.

Objective prior distributions

The objective approach is opposed to the subjective approach. Instead of trying to formulate a large amount of information in the form of a prior distribution, the objective approach tends to use the less information to obtain the prior. This approach lets the data have the precise role in the posterior distribution. This is often called ``letting the data speak for itself" or ``prior ignorance" .Usually the

objective priors are called non-informative distributions which result in proper posterior distributions.Some of the principles leading us to choose priors based on the objective approach are :

- One is not experienced enough in the studied subject and doesn't want to make the inference in a specific way. In other words, the researcher doesn't want to be biased against a specific area of the parametric space.

- Sometimes, it is hard or even impossible to consider the recommendations of professionals about the structure of the parametric space in form of a prior distribution.

- The researcher wants to minimize the influence of a wrong choice of priors on the inference process.

Different types of non-informative priors

Non-informative prior distributions can be categorized in four groups [1]: Laplace priors, Invariant priors, Jeffrey's' prior and Reference prior distributions.

Laplace prior distribution

non-informative priors were firstly applied by Laplace who issued this example in 1973:

There are n black and white balls in a bag. The first selected ball from the bag is white. What is the probability that the proportion of white balls P is equal to P_0 As Laplace knew nothing else about the proportion of balls except for the first ball being white, he assumed that P in uniformly distributed on $\{\frac{2}{n}, \dots, \frac{n-1}{n}\}$.

This uniform distribution indicates his evasion of weighting to a specific volume of the parameter. Such distributions, which gave the same weight to different spots of a parametric space, are called Laplace priors. There are several objections to Laplace prior distributions, the most important of which is its variance under transformation. This objection would be explained here under:

We assume that P has distribution U (0, 1) which is a Laplace prior. We also know that θ= -ln P ~ exp(1), which is a non-uniform distribution and gives more weight to the volumes tending to zero (it is a discounting function of θ). If the "ignorance" is caused by unawareness about the structure of P parameter space, the existence of this unawareness about any of the functions of P is both logical and natural. This feature is called the principle of invariance under transformation. It's clear that Laplace distribution doesn't follow this principle. Graph shows the density function of both distributions.

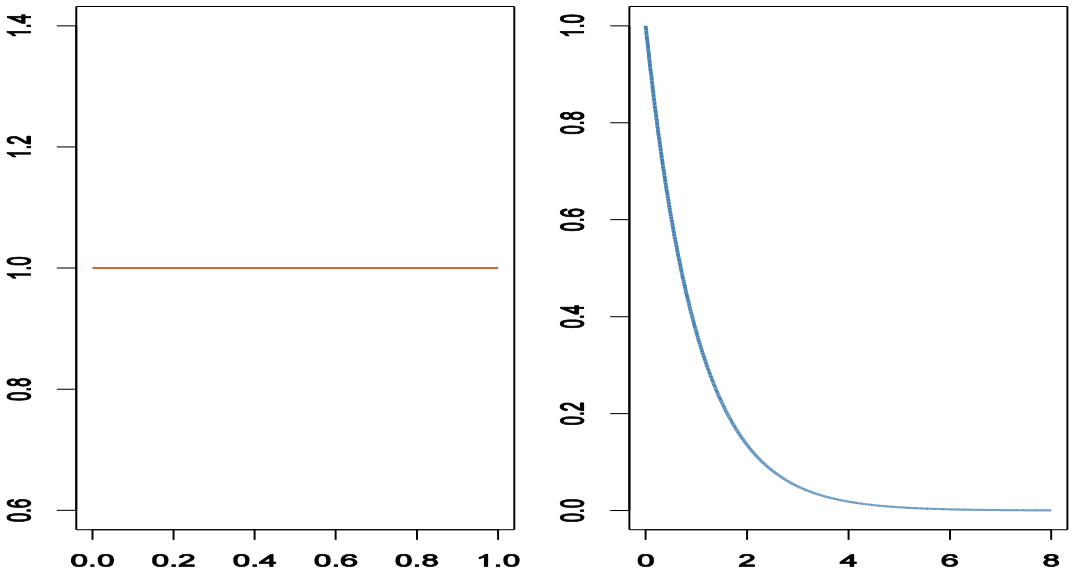

Figure 2: Left : $p \sim U(0,1)$, Right : $\theta = -\ln p \sim exp(1)$

Invariant prior distribution

Invariance properties of distributions are studied according to several groups of transforms. Here, we'll study only Local and Scale transformations.

Definition 2

(Local invariant priors) The family f(.) is local invariant if

$$f_X(x) = f_{X-x_0}(x-x_0)$$

If $f_X(x) = f_{X-x_0}(x-x_0)$ then density function is called local transformation invariant with respect to θ And θ is called local parameter. We say prior distribution π (θ) is local invariant if:

$$\pi(\theta) = \pi(\theta - \theta_0), \ \theta_0$$

Knowing that this transformation must hold with respect to θ0, π(θ) = c is the only solution. Note that

this transformation leads to an improper prior distribution.

Definition 3

(Scale invariant priors) The family f(.) is scale invariant if

$$f_X(x) = \frac{1}{\sigma} f_{\frac{X}{\sigma}}(\frac{x}{\sigma})$$

If $f_X(x) = \frac{1}{\sigma} f\frac{X}{\sigma}(\frac{x}{\sigma})$, then density function is called scale transformation invariant with respect to θ and θ is called scale parameter. We say prior distribution π(θ) is scale invariant if:

$$\pi(\theta) = \frac{1}{c}\pi(\frac{\theta}{c}), \ c > 0$$

Note that the only solution of above equation is $\pi(\theta)=\frac{\alpha}{\theta}$ Where α is a constant value. This transformation also leads to an improper prior distribution. The question is that which group of transformations should be considered. According to the above definitions, we see that there are different ways for defining the invariance properties.

Jeffrey's' prior distribution

Harold Jeffreys who was the British Physician, Mathematician and Statistician presented a prior distribution in 1946 which was based on Fisher Information Matrix. If

$$I(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial \log f(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^2$$

be the Fisher's information matrix, the Jeffreys' prior distribution is:

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}.$$

The symbol E_X Means that the expectation is taken with respect to variables X.

Jeffreys' prior distribution is invariant under reparametrization of θ. This is the main property of that. If ϕ be a function of θ, we have

$$\begin{aligned}
\pi(\phi) &= \pi(\theta) \mid \frac{d\theta}{d\phi} \mid \\
&\propto \sqrt{\det I(\theta)} \mid \frac{d\theta}{d\phi} \mid \\
&= \sqrt{\det \mathbb{E}_X \left[\frac{dL(x|\theta)}{d\theta}\right]^2 (\frac{d\theta}{d\phi})^2} \\
&= \sqrt{\det \mathbb{E}_X \left[\frac{dL(x|\theta)}{d\theta}\frac{d\theta}{d\phi}\right]^2} \\
&= \sqrt{\det \mathbb{E}_X \left[\frac{dL_\theta(x|\theta)}{d\phi}\right]^2} \\
&= \sqrt{\det \mathbb{E}_X \left[\frac{dL_\phi(x|\phi)}{d\phi}\right]^2} = \sqrt{\det I(\phi)}.
\end{aligned}$$

$$\pi(\mu|\sigma) \propto \sqrt{I(\mu|\sigma)}$$

$$\begin{aligned}
\Rightarrow I(\mu|\sigma) &= \mathbb{E}_X \left[(\frac{\partial}{\partial\mu} \log f(X|\mu,\sigma))^2\right] \\
&= -\mathbb{E}_X \left[\frac{\partial^2}{\partial\mu^2} \log f(X|\mu,\sigma)\right]
\end{aligned}$$

In the following we explain obtaining Jeffreys' prior distribution with an example.

Example 2

Determine Jeffreys' prior distribution for N(μ,α).

$$f(X|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$
$$\Rightarrow \log f(X|\mu,\sigma) = -\frac{1}{2\sigma^2}(x-\mu)^2 - \log(\sqrt{2\pi\sigma^2})$$
$$\Rightarrow \frac{\partial}{\partial\mu}\log f(X|\mu,\sigma) = \frac{(x-\mu)}{\sigma^2}$$
$$\Rightarrow \frac{\partial^2}{\partial\mu^2}\log f(X|\mu,\sigma) = -\frac{1}{\sigma^2}$$

We have

$$I(\mu|\sigma) = -\mathbb{E}_X\left[\frac{\partial^2}{\partial\mu^2}\log f(X|\mu,\sigma)\right] = \frac{1}{\sigma^2}$$

Therefore

$$\pi(\mu|\sigma) \propto \sqrt{I(\mu|\sigma)} = \sqrt{\frac{1}{\sigma^2}}$$
$$\Rightarrow \pi(\mu|\sigma) \propto \frac{1}{\sigma}$$

Being σ constant, one can say $\pi(\mu|\sigma) \propto c$. It means that this distribution is improper. Now, we calculate Jeffreys' prior distribution for σ.

$$\pi(\sigma|\mu) \propto \sqrt{I(\sigma|\mu)}$$
$$\Rightarrow I(\sigma|\mu) = \mathbb{E}_X\left[\left(\frac{\partial}{\partial\sigma}\log f(X|\sigma,\mu)\right)^2\right] = -\mathbb{E}_X\left[\frac{\partial^2}{\partial\sigma^2}\log f(X|\sigma,\mu)\right]$$

We also have

$$\log f(X|\sigma,\mu) = -\frac{(x-\mu)^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}$$
$$\Rightarrow \frac{\partial}{\partial\sigma}\log f(X|\sigma,\mu) = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}$$
$$\Rightarrow \frac{\partial^2}{\partial\sigma^2}\log f(X|\sigma,\mu) = -\frac{3(x-\mu)^2}{\sigma^4} + \frac{1}{\sigma^2}$$
$$= -\frac{3}{\sigma^2}\left(\frac{x-\mu}{\sigma}\right)^2 + \frac{1}{\sigma^2}$$

So

$$I(\sigma|\mu) = \mathbb{E}_X\left[\frac{3}{\sigma^2}\left(\frac{x-\mu}{\sigma}\right)^2 - \frac{1}{\sigma^2}\right]$$
$$= \frac{3}{\sigma^2}\mathbb{E}_X\left[\left(\frac{x-\mu}{\sigma}\right)^2\right] - \frac{1}{\sigma^2}$$
$$= \frac{3}{\sigma^2} - \frac{1}{\sigma^2} = \frac{2}{\sigma^2}$$

Therefore

$$\pi(\sigma|\mu) \propto \sqrt{I(\sigma|\mu)} = \sqrt{\frac{2}{\sigma^2}} = \frac{1}{\sigma}$$
$$\Rightarrow \pi(\sigma|\mu) \propto \frac{1}{\sigma}$$

Reference prior distributions

Reference prior, introduced by Bernardo in 1979, is a typical class of non-informative priors which was developed by Berger and Bernardo in 1989 . The method through which the Reference prior distribution can be obtained is known as Berger-Bernardo. In one dimensional case, this method results in a Jeffreys' prior and in multidimensional case, this method has more several priorities compared to that of Jeffrey's. In such case, π(θ) is defined as a function which maximizes the missing information and it is known as Kullback-Leibler divergence.

$$D_{KL}\left(\pi(\theta|x), \pi(\theta)\right) = \int \pi(\theta|x)\log\frac{\pi(\theta|x)}{\pi(\theta)}\mathrm{d}\theta$$

In most cases, this distance leads to non-informative prior distributions. The aim is to maximize the above integral. In fact, this method maximizes the expected posterior information about X when π(θ) is the prior distribution. In 1992, Berger and Bernardo have proved that those priors maximizing the expected value are generally discrete distributions .

Empirical and hierarchical bayes

In previous sections, we discussed the methods through which prior distributions were determined. Assume that we have considered a parametric family of distributions as priors. Hence the parameter (hyperparameter in the whole model) of these distributions is unknown, determination of the prior is still incomplete. Hierarchical and empirical Bayesian approaches present a method to determine such hyperparameters .

Hierarchical bayes

Assume that hyperparameters of prior distribution have a distribution with different parameters. These new parameters have also a distribution with another parameters. Such models are called hierarchical models. Theoretically, the number of these levels of such models can be unlimited; But practically, large number of such levels causes the model be very complicated.

When these levels increase, the importance of hyperparameters decreases. Therefore, we can take the hyperparameters as a constant value in a specific level of modelling. Then, we can lessen their (hyperparameters) effects by integrating over hyperparameters and obtain a prior distribution without any parameters. For instance, consider a model with two levels of parameters:

$$x \sim f(x|\theta)\ ,\ \theta \sim \pi(\theta|\eta)\ ,\ \eta \sim \pi(\eta|a)$$

where a is a constant value. Therefore,

$$\pi(\theta) = \int \pi(\theta, \eta)\, d\eta = \int \pi(\eta|a)\pi(\theta|\eta)\, d\eta$$

which means that prior distribution θ is only a function of a and other constant values.

Empirical bayes

The empirical Bayesian method gives us a function of observations and hyperparameters by integrating over posterior distribution of all parameters and hyperparameters being proportional to the parameter. This function can be called the likelihood function of hyperparameters. Then, it estimates the hyperparamteres using the maximum likelihood method (which is a non-Bayesian method!). For example, suppose that

$$x \sim f(x|\theta)\ ,\ \theta \sim \pi(\theta|\eta)$$

$$L(\eta|x) = f(x|\eta) = \int f(x, \theta|\eta)\, d\theta = \int f(x|\theta)\pi(\theta|\eta)\, d\theta$$

$$\hat{\eta}_{ML} = \arg \max_{\eta \in \boldsymbol{H}} L(\eta|x)$$

Then, Bayesian inference can be continued using these estimated hyperparameters. Some of the Bayesians don't take this method as Bayesian since the data are used to determine the hyperparameteres.

Outline

We have elucidated different apparoaches toward choosing the prior distribution in this article. The point here to be accounted is that there is no definite way for choosing the prior among different approaches. Therefore, to select a prior distribution, the researcher can choose a specific approach for a definite problem according to its advantages and disadvantages.

References

1. Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, Second Edition.
2. Berger, J.O., Bernardo, J.M., 1989. Estimating a Product of Means: Bayesian analysis with reference Priors. J. Amer. Statist. Assoc., 84,200-207.
3. Berger, J.O., Bernardo, J.M., 1992. On the Development of the Reference Prior Method. Bayesian Statistics, Oxford University Press London.
4. Bernardo, J.M., 1979. Reference Posterior Distribution for Bayesian Inference (with Discussion). J. Roy. Statist. Soc., B41,113, 113-147,
5. Casella, G., Berger, J.O., 2002. Statistical Inference". Second Edition, Duxbury Adv. Ser.
6. Christian, P.R., 2001. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Sec.Edit., Springer.
7. Fink, D., 1997. A Compendium of Conjugate Priors". The Magazine of Western History, Publisher: Citeseer, Pages: 1-4.
8. Gholami, G.H., 2008. Change-point Problems in Regression: A Bayesian Approach.Ph.D. Thesis.
9. Ho, P.D., 2009. A First Course in Bayesian Statistical Methods. Springer.
10. Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. London., A186, 453, 453-461.
11. Jeffreys, H., 1961. Theory of Probability. Third Edition, Oxford University Press, London.