Bayesian Inference

Books:

Reference Books:
1. James O. Berger (1985) Springer. (chapter 4).
2. Jose' Bernardo & Adrian Smith, Wiley Bayesian Theory.
3. DeGroot: Optimal statistical decision
4. Ashok Bansal. → Text Book
5. SK. Sinha Bayesian Estimation → Text Book
6. Peter H. Lee → Text Book
7. C.P. Robert → Reference Book.

Syllabus:
1. General Discussion
2. Prior distribution and Elicitation
3. Point Estimation.
4. Interval Estimation.
5. Bayes Testing.
6. Bayes computation.
7. Bayes sufficiency.
8. Bayes Robustness.
9. Empirical Bayes.
10. Hierarchical Bayes.

What is Bayesian Inference:

Statistical Inference

- Classical or Frequentist Inference
- Bayes Inference (Deductive Inference)
- Fiducial Inference

Basis of Bayesian Inference:

$p(\theta|x) \propto L(\underline{x},\theta) \cdot g(\theta)$ → Bayes Theorem

- $p(\theta|x)$: Posterior distribution
- $L(\underline{x},\theta)$: Likelihood function
- $g(\theta)$: prior

Inverse probability (as we go in past to actually calculate probability).

[In Bayes Theorem, we actually calculate the conditional probability of past given present values].

If consider all the continuous variables, then

$$p(\theta|x) = \frac{L(x;\theta)\cdot g(\theta)}{\int L(x;\theta)\, g(\theta)\, d\theta}$$

If we consider all the discrete variables, then

$$p(\theta|x) = \frac{L(x;\theta)\cdot g(\theta)}{\sum_{\theta} L(x;\theta) g(\theta)}$$

Classical Inference, was first developed by R.A. Fisher (1910-1950).

In classical definition of probability

$$P(A) = m/n$$

This definition has plenty of drawbacks.
Hence come the statistical definition of probability.
Where we keep on increasing the experiment.
Then the ratio converges to a constant value and that is the probability.

Then, the definition becomes

$$P(A) = \lim_{n\to\infty} (m/n).$$

This definition has no limitations.

This definition is called frequentist inference.
This inference is also called Inductive Inference.

Thomas Bayes, was behind the entire Bayes Paradigm (1702-1761).

[He wrote a letter to Richard Price about the inverse probability (one-half page) and the letter was published later in 1763.

Around 1920, Ramsey, Good, Savage, Lindley published some foundational papers and R.A. Fisher was completely against those papers. Then in 1953 again this came into picture.

* Other than classical ~~inferences~~ probability, and statistical probability there is some other probability that is subjective probability.
  - Based on belief, previous experimentations.
  - Non-repeatable.

The concept of prior probability is based on subjective probability.

(*) Likelihood function says information about θ, based on random sample.

The likelihood function is based on objective probability.

Both subjective and objective are combined to provide an updated version. This is the posterior probability.

$x|\theta \rightarrow$ distribution of x given parameter θ. (θ fixed value / unknown constant).

$\theta|x \rightarrow \theta$ random variable

(θ is treated as random variable, though it is fixed. But whatever θ be, we should have proper justification.

θ is update belief.

(*) When we use objective information in prior, we call it Objective Bayesian Inference.

Classical Inference

$x \sim f(x|\theta)$

$x_1, x_2, \ldots, x_m$

$L(x;\theta)$.

Inductive ~~Subjective~~ Inference

Bayesian Inference

$x \sim f(x|\theta)$.

$\theta \sim g(\theta)$.

$x_1, x_2, \ldots, x_m$

$L(x;\theta)$.

$p(\theta|x) = L(x;\theta) \cdot g(\theta)$

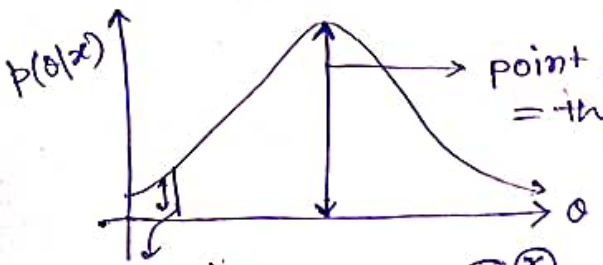
↳ Deductive Inference

Advantages of Bayesian Inference

1. Easy Interpretation:

We consider $p(\theta|x)$.

↳ Basis of Bayesian information



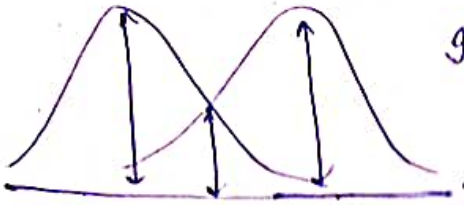
point estimate = the point (value of θ) with most belief (probability)

posterior probability at different values of θ

(*)(*) If loss function = Squared error loss function, the posterior mean = point estimate

Why squared error loss function?

→ Because of overestimation and under-estimation.



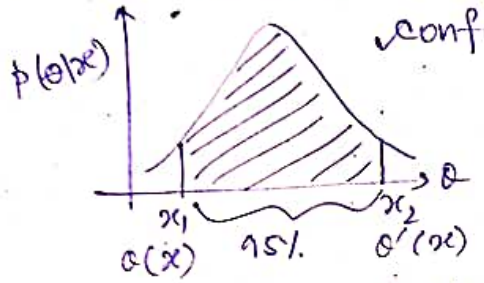
If distribution is skewed, what is the solution?

→ Then posterior mean (that is point estimate) is not appropriate

(*) Sometimes, overestimation greater more loss.
sometimes, underestimation →

→ Overestimation, underestimation equally vulnerable, while estimating the average.

While in Interval Estimation,



confidence interval using posterior distribution.

Other cases: $H_0 : \theta \leq \theta_0$
$H_1 : \theta > \theta_0$

$p(H_0|x) \gtrless p(H_1|x)$

depending on the probabilities
The one with more probability is always preferred.

Mean, variance in Bayesian Paradigm:

(*) NO samples exist
(*) M-Sample mean, variance does not exist.
(*) Concept of type-I error is not valid here

Bayesian Inference is just a probabilistic comparision.

$$\left.\begin{aligned} P[T_1(x) < \theta < T_2(x)] &= 0.95 \\ P[\theta(x) < \theta < \theta'(x)] &= 0.95 \end{aligned}\right\}$$

We are 95% confident that the random interval $(T_1(x), T_2(x))$ contains θ.

Interpretation of $P[\theta(x) < \theta < \theta'(x)] = 0.95$

P [The random θ is going to be covered in these two fixed numbers $\theta(x)$ and $\theta'(x)$] = 0.95

Then $P[1.71 < \theta < 3.12] = 0.95$

$P[1.70 < \theta < 3.20] = 0.95$ → This interval may or may not contain θ.

1.21, 2.34, 3.12, 2.75, 3.15.

we can not say whether this samples belongs to 5% or 95%.

θ was Random, and θ is still random in Bayesian Inference. Hence Bayesian Inference have easy Interpretation

2. Bayes Inference obeys the likelihood principle

likelihood principle:

When we consider two sets samples and calculate their likelihoods and draw inferences separately, then their likelihoods are proportional to each other Bayesian Inference and not in classical inference.

Consider, 12 tosses where 9 heads and 3 tails, to and the hypothesis is $H_0 : \theta = 1/2$ vs $H_1 : \theta > 1/2$

Here $n = 12$ fixed.

We consider x = number of success in n trials.

$$L_1(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

$$= \binom{12}{9} \theta^9 (1-\theta)^3.$$

Suppose, the tossing is done until 3rd tail appears

X = no. of heads required to complete the experiment.

Therefore $X \sim$ Negative Binomial $(r = 3, \theta)$

$$L_2(\theta) = \binom{r+x-1}{x} \theta^x (1-\theta)^r.$$

$$= \binom{11}{9} \theta^9 (1-\theta)^3.$$

Then $P(X \geq 9 \mid \theta = \frac{1}{2})$ for $L_1(\theta)$

$$= \sum_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = 0.075$$

Considering Negative Binomial sampling plan

$$P(X \geq 9 \mid \theta = \tfrac{1}{2}) = \sum_{j=9}^{\infty} \binom{2+x}{j} \theta^j (1-\theta)^3.$$

$$= 0.0325.$$

Both likelihood functions are proportional to each other.

P-value for $L_1(\theta)$ is > 0.05

p-value for $L_2(\theta)$ is < 0.05

Both provides different inferences, one rejects and another accepts.

Hence Bayesian Paradigm obeys likelihood Principle, and classical does not.

3. Bayesian Inference does not lead to ~~absolute~~ absurd result.

$X \sim \text{Poisson}(\lambda)$.

$\theta = e^{-2\lambda}$ $\forall \theta < 1$.

The UMVUE of θ is $(-1)^x$.

This is an absurd estimator, as if $x = \text{odd}$ then θ will take negative values.

There is no such absurd estimators in Bayesian Paradigm.

1. Classical inference can be considered as a special case of Bayesian Inference. i.e. $g(\theta) \propto$ a constant.

$$p(\theta|x) = \frac{L(\underline{x}, \theta)\, g(\theta)}{\int L(\underline{x}, \theta)\, g(\theta)\, d\theta}$$

$p(\theta|x) \propto L(\underline{x}, \theta)$. if $g(\theta) \propto$ a constant.

2. $\underline{x} : x_1, x_2, \ldots, x_n$

One another observation x_{n+1} is added.

In classical inference MLE can not be updated for the new observation.

In Bayesian we can do that.

In Bayesian we calculate new posterior distribution as

$$p(\theta|\underline{x}, x_{n+1}) = \frac{L(x_{n+1}, \theta) \cdot p(\theta|\underline{x})}{\int L(x_{n+1}, \theta) \cdot p(\theta|\underline{x})}$$

The previous posterior is acting as prior to the new posterior distribution.

This is a formula for learning.
In Bayesian Inference, we learn by the formula itself.

3. Subjective Probability and Prior Elicitation :

The theory of subjective probability has been created to enable one to talk about probabilities when frequentist view point does not apply (some even argue that frequentist viewpoint never applies, it is being impossible to have an iid sequence of infinite repeatations of any scenario except in a certain imaginary sense). The main idea of subjective probability is to let the probability of event reflect the personal belief in the chance of occurance of the event.

For example, you may have a personal feeling as to the chance that unemployment rate, say θ will be between 3% and 4% even though no frequentist probability can be assign to the event. There is of course nothing terribly suprising about this. It is common to think of personal probability all the time when betting the outcome of a game, when evaluating the chance of a rain and in many other situations. Generally, frequentist probability is calculated by simply calculated on the basis of relative probabilities, where as subjective probability on the basis of introspection.

The simplest way of determining the subjective probability is by comparing the ~~probabilities~~ events determining relative likelihood.

Say for example, it is desired to find probability of E. simply compare E with E^c.

If E is felt to be twice as likely to occur as E^c then clearly P(E) = 2/3 and $P(E^c)$ = 1/3.

* The subjective probabilities should be considered in such a way that $P(A)+P(A^c)=1$.
* If we are consistent while constructing the subjective probabilities then the subjective probabilities must follow 3 Axioms of probability.

For example,

P(Occuring) and P(Not occuring) are calculated by subjective probabilities but both are based on different situations.

This is rather loosely stated but corresponds to the intuitive manner in which we think probabilities. As with Utility theory, a formal set of axioms can be considered under which subjective probabilities can be considered to exist and will behave in the fashion of usual probabilities.

An alternative characterization of subjective probability can be acheive through consideration of betting. In betting scenario one determines P(E) by imaging being involved in a Gamble where in z will be lost if E occurs and 1-z will be gained if E^c occurs. We consider $0 \le z \le 1$. The idea is then to choose is z. so, that the gamble is fair.

[Fair game,
When a game is played for a longer time, the overall utility = 0.
Gain = Loss = 0].

Resulting in $U(-z)P(E) + U(1-z)\cdot P(E^c)$ as Expected utility.

Expected utility = 0

$\Rightarrow U(-z)P(E) + U(1-z)P(E^c) = 0.$

[Utility = − Expectation].

This can be written as,

$$\Rightarrow u(-z)P(E) + u(1-z)\{1-P(E)\} = 0$$

$$\Rightarrow P(E)\{u(-z) - u(1-z)\} + u(1-z) = 0$$

$$\Rightarrow P(E) = \frac{u(1-z)}{u(1-z) - u(-z)} \simeq \frac{1-z}{1-z+z} \simeq 1-z$$

[Utility is and subjectived probability are Inter-related. If utility is challenging then subjective probability is equally chalenging and vice-versa]

One might object that this betting mechanism is circular since utility function was constructed by considering probability aspects. And now we are trying to determine P(E) from the knowledge of utility. So, the dilema can be resolved by noting that in the construction of utility function, any available probability mechanism can be used. (say a random number table). There is no need to involve probabilities of E.

The betting and scoring rule scenarios are very useful in providing additional evidence for the use of subjective probability in quantifying uncertainty. We do not however view them as good operational devices for determining subjective probability. They can be of view in ensuring that a probability elicitor is careful or honest but the mechanism do not appear intuitively accessible.

Practical Difficultiy of subjective Probability Elicitation

1. care must be taken to ensure consistency. for example it is unusual to conclude an irrational

$$P(A) = 1/3 \;,\; P(B) = 1/3 \;,\; P(A \text{ or } B \text{ or both}) = 3/4.$$

conclusion for some one trained in probability.

2. Everyone has great deal of trouble in accurately specifying small probabilities.

Savage (1971)
Kadane & others (1980).

[Usually, we are not able to consider small changes in probability of numbers. like (5.0 - 6.0) → 0.90, what is P(Height exactly 5.89).].

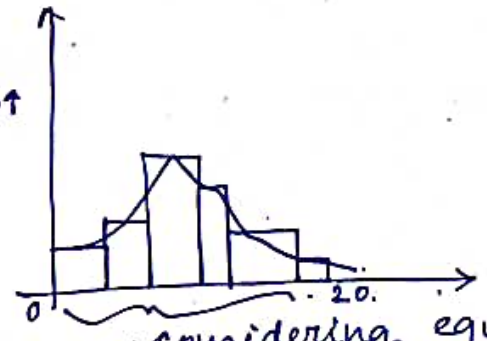
Prior Elicitation:

(Anthony Haegen — Prior Elicitation University of Schafield)

This can be said as subjective determination of prior.

$\pi(\theta)$ is to be determined subjectively where $\theta \in (H)$. If (H) is discrete, the problem is simply to determine subjective probability of each element of E. When (H) is discrete and infinite then the ~~probability~~ problem is considerably more difficult.

The first approach is Histogram Approach. When (H) is an invertal on real line, then the most obvious approach is histogram Approach. The approach is divide (H) into interwals, determine subjective probability of each interwal then plot a probability Histogram. for this histogram a smooth density of $\pi(\theta)$ can be sketched.



considering equal probabilities in the interwals.

Problem: 1. How many interwals.
2. Difficult to get proper density function in that particular interwal.
3. Difficulty to get probability in tails.

No clear cut rule establishes how many intervals, what sizes of intervals etc. should be used in the histogram approach. For some problems, only the very crude histograms and priors will be needed, (especially when samples/observations are very large). While for others highly detailed versions may be required, the exact needs can be determined by robustness consideration.

Two others difficulties are —
priors so obtained is somewhat difficult to work with and the second ones has no tails.

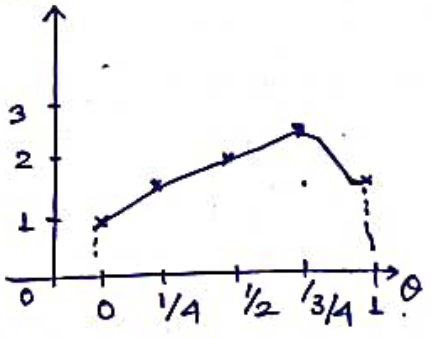
2. Relative Likelihood Approach:

[Tail probabilities are always difficult to get. Hence mostly we draw densities based on their central values]

Suppose Θ is a subset of real line. The procedure consists simply comparing. The intuitive comprises to directly from this determinations. say $\Theta = [0,1]$. and directly sketching prior densities Begin by determining the relative likelihoods of most likely and least likely parameter points.

let $\theta = 3/4$, most likely.
$\theta = 0$, least likely

suppose $\theta = 3/4$ is estimated to be three times as likely to be true value of θ. as true value of $\theta = 0$. determine the relative likelihoods of three other points say $1/4$, $1/2$ and 1. For simplicity compare all this points with $\theta = 0$. It is felt that $\theta = 1/2$ and $\theta = 1$ are twice as likely as $\theta = 0$. While $\theta = 1/4$ is 1.5 times as likely as $\theta = 0$.



Assign a prior belief to $\theta = 0$. More points can be included if finer sketch is desired.

The prior density so obtained may not be proper. but is not really desired.

Suppose a consider constant c is exist which make $c\pi(\theta)$ a proper.

The Bayes action is same whether $c\pi(\theta)$ or $\pi(\theta)$ is the prior.

So, c is not important.

So, any a minimizing $\int L(\theta,a)\, c\pi(\theta)\, d\theta$ will minimise the other $c\int L(\theta,a)\pi(\theta)\, d\theta$.

When Θ is unbounded, since the relative likelihood determination can only be done in a finite region, one must think what to do outside the region

Two particular problems—

1. Shape of the density outside the regions, should it decrease, θ^{-1}, θ^{-2} or $e^{-\theta}$.

2. Need to normalise density:

Carefully determined central density, correctly proportional to the outside the scale of density. This can be done by simply determining the prior probabilities of central and outside regions and making sure that the estimated prior densities provides corresponding masses two regions.

3. Matching a given functional form

This is the most used and unused method to get a prior distribution. The idea is to simply assume that $\pi(\theta)$ is of a given functional form and then to choose the density of this form that closely matches the prior beliefs. Once a density is matched the task is to subjectively determine the prior parameters to exactly specify the density.

Say for example $B(\alpha,\beta)$ is the functional form. One can then estimate, (for example, say the prior mean μ and the prior variance σ^2

$$\mu = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Unfortunately, the estimation of prior moments is often difficult, especially when the tails of the density have drastic effects on the moments. For example, if the tails behave like $K\theta^{-2}$ on $(0, \infty)$, then the density has no moment. If K is small enough, this tail will have almost insignificant probability. Since it is the probability, that can be more reasonably specified subjectively, a tail of so small probability cannot realistically be known.

Nevertheless, its influence on the moment can be great, atleast when dealing with unbounded parameters. For bounded parameters the things are comparatively easier.

⊛ A better method of getting prior parameter is to subjectively estimate several fractiles, and then choose the parameter of the given functional form to obtain density matching this fractiles as closely as possible. It is precisely the estimation of probabilities of Variance reasons.

Say for example $\Theta = (-\infty, \infty)$ and suppose the prior is normal and then it is subjectively determined, then median of prior is '0' and 1/4 and 3/4 fractiles are -1 and 1. Since for normal distribution it is symmetrical ly distributed about '0', so $\mu = 0$.

$$P\left[Z < \frac{-1}{(2.19)^2} \right] = 1/4., \text{ where } Z \sim N(0,1).$$

$N \sim (0, 2.19)$

Alternatively, if it is assumed the prior is $C(0,1)$, one finds that cauchy(0,1) is the appropriate choice.

Because the median is 0, so

$$\int_{-\infty}^{-1} \frac{1}{\pi} \frac{1}{(1+\theta^2)} d\theta = \frac{1}{4}.$$

Two observations —

1. For a given assumed functional form only a small number of fractiles needs typically be found to determine the specific choice of the prior.

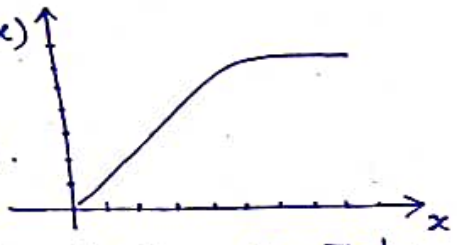
Although this makes the analysis quite simple,

it is troublesome. Since what should be done, ~~what~~ if other fractiles dont agrees with implied choice. The answer is obvious. If there is unavoidable significant disagreement between the densities of certain functional form and subjectively choosen fractiles of the prior that functional form can be discarded as an appropriate model for the prior.

2. Considerably, different functional form can be choosen for the prior density. The quention is whether the choice of functional form is important. The answer is obviously yes, certain functional form are generally superior over others.

4. CDF determination:

Another important technique is the subjective construction of the CDF. The same can be done by determining several α fractiles and plotting the points on the line (x, α). and then sketching a smooth curve by joining them. This can be accurate technique that provide CDF from which the density can be determined.



Note: (A) Histogram and Relative likelihood approaches are usually preferred.

(B) Matching the given functional form is used in following cases

(1). When the density of a standard function form can be found which gives a good match to the prior density obtained by other approaches

(2) When only vague prior is available, one may use the standard functional form in that situation using the vague prior information and then proceed by performing robustness studies.

(1) All the parameters are a priori independent. Then we are left with k univariate priors.

↳ Approximation

i.e. $\pi(\theta_1, \theta_2, \ldots, \theta_k) \xrightarrow{\text{independent}} \pi(\theta_1) \cdot \pi(\theta_2) \cdots \pi(\theta_k)$

Another assumption is,

for bivariate two parameters,

$$\pi(\theta_1, \theta_2) = \pi(\theta_1) \cdot \pi(\theta_2|\theta_1).$$

(both are univariate (one marginal, one conditional).

Different types of priors:

Non-informative priors — Vague
Weakly
Default.

Conjugate prior. (Only objective is mathematical convenience).

↳ [Raiffa and Schlaifer] → 1961.

[Box and Tiao] → 1983 (First Edition)
1991 (second Edition).

Bayesian Theory
Bayesian Applications (MCMC).

Laplace: → Approached Non-informative prior.

(Assign equal probability to members and select an individuals by simple random sampling)

When Range is finite, $\pi(\theta) = \frac{1}{b-a}$, $a \le \theta \le b$.

When Range is infinite, $\pi(\theta) \propto$ a constant.

↓

This is an improper prior.

[Those prior which is not integrable and $\int \text{prior} \neq 1$. Then prior is improper].

(*) But non-informative prior can be chosen sometimes. More important is posterior is proper.

Different kind of priors

1. Non-informative prior (also called vague, weak, default priors).
2. Conjugate prior (often informative)

Prior
- proper prior
- Improper prior (weight)
- Informative (generally prior)
- Non-informative (generally improper)

Conjugate prior:

These are the priors such that prior and posterior belong to the same family.

- Jeffrey's prior.
- Entropy prior
- Reference prior. (By Berger & Bernardo).

Theory of Laplace:

If parameter space is discrete or continuous with finite values then assign probability of $1/k$ to each of them, where k is the number of parameters.

Ⓗ : discrete

Ⓗ : continuous $[a,b]$ $\mathbb{R}^+$ or $\mathbb{R}^-$

- for Ⓗ in $\in [a,b]$, then we consider prior $= 1/(b-a)$, $a \le \theta \le b$.
- for $\mathbb{R}^+$, $\mathbb{R}^-$ and $\mathbb{R}$.

Consider $\pi(\theta) \propto$ a constant for $\theta \in$ Ⓗ.

Note: for infinite values this prior happens to be improper.

Note: $\eta = \exp(\theta)$, this causes the location parameter to change into scale parameter.

For θ, it could be $\pi(\theta) \propto$ a constant

$\pi(\eta) \propto$ a constant

$$d\theta = \frac{1}{\eta} d\eta$$

Lack of invariance transformation in the suggestion of laplace. $\pi(\eta) \propto \frac{1}{\eta}$.

Suggestion: One may choose the most relasonable parameterisation and that lack of prior information, should correspond to a constant density in this parameterisation. That the argument is hard to defend in general. The lack of variance of constant prior requires the search for non-informative priors, which are approximately invariant under transformation.

Non-informative priors for location and scale transformations:-

Jeffreys (1961)
Hartigan (1964)

* Location parameters:

$x \sim$ pdf $f(x|\theta)$

Exp(w).

To If θ is location parameter then $f(x)$ will be in form of $f(x-\theta)$. as for example exp(w).(shifted exponential), Normal$(\theta, 1)$.

To derive non-informative prior for this setup let us assume that, instead of observations x we take y, such that

$$y = x + c,$$ where c is a real constant.

Define $\eta = \theta + c$, then $y \sim f(y-\eta)$.

If x and θ belongs to a real space then y and η must belong to $\mathbb{R}$ therefore $X(\theta), Y(\eta)$ are identical in structure.

It seems reasonable to insist that they have the same non-informative prior.

Another way of thinking is to note that observing y's is same as observing x's, with different shifted origin. Here one has the origin c and not '0'. since the choice of an origin for a unit of ~~mes~~ measurement is quite arbitary, the non-informative prior should be independent of this choice.

Let π and π^* be non-informative prior in (x, θ) and (y, η) the above argument implies that π and π^* should be same.

That is $P^{\pi}(\theta \in A) = P^{\pi^*}(\eta \in A)$ — ①.

For any A belonging to $\mathbb{R}$ $\because \eta = \theta + c$, we can also write

$$P^{\pi^*}(\eta \in A) = P^{\pi}(\theta + c \in A)$$
$$= P^{\pi}(\theta \in A - c) \quad — ②$$

Therefore ① and ② implies,

$$P^{\pi}(\theta \in A) = P^{\pi}(\theta \in A - c) \quad -③ \quad \forall c.$$

Only π having this relationship is said to be a location invariant prior.

Assuming that the prior has density we can write 3 as

$$\int_A \pi(\theta)\,d\theta = \int_{A-c} \pi(\theta)\,d\theta$$
$$= \int_A \pi(\theta - c)\,d\theta$$

If this holds for all sets A, it can be shown that-

$$\pi(\theta) = \pi(\theta - c), \forall c.$$

If $\theta = c$, then $\pi(c) = \pi(0), \forall c$.

$\pi(\theta)$ $\propto$ a constant → location invariant prior.

Scale parameter:

A scale density is in the form

$$X \sim f(x|\sigma) = \sigma^{-1} f(x/\sigma), \quad \sigma > 0.$$

Say for example we have $X \sim$ Normal $(0, \sigma^2)$

$X \sim$ Exponential (θ).

Imagine that instead of imagining X we observe a random variable Y such that $Y = Xc$ $(c > 0)$

Define $\eta = c\sigma$.

Therefore density of y will have the form $Y \sim f(y|\eta)$.

$$Y \sim f(y|\eta) = c^{-1}\sigma^{-1} f(y/c\sigma)$$
$$= \eta^{-1} f(y/\eta)$$

If x is positive, σ is positive real spaces $(\mathbb{R}^+)$ then the sample and prior spaces for the two problems (x, σ) and (Y, η) are same. The two problems are thus identical in structure which again indicates that they should have the same non-informative prior.

Here the transformation can be thought of as simply a change in the scale of measurement say from inch to feet.

Let π and π^* denote the prior in (x, σ) and (Y, η) set up. The above argument means that the equality

$$P^{\pi}(\sigma \in A) = P^{\pi^*}(\eta \in A). \quad -④.$$

should hold for all A in $0 < A < \infty$.

i.e. $\eta = c\sigma$.

Then $P^{\pi^*}(\eta \in A) = P^{\pi}(c\sigma \in A) = P^{\pi}(\sigma \in c^{-1}A) \quad -⑤.$

Putting ④ and ⑤ together we can write

$$P^{\pi}(\sigma \in A) = P^{\pi}(\sigma \in c^{-1}A) - ⑥ \quad , c>0$$

Any π for which this holds is called scale invariant..

$$\Rightarrow \int_A \pi(\sigma)d\sigma = \int_{c^{-1}A} \pi(\sigma)d\sigma$$

$$= \int \pi(c^{-1}\sigma)c^{-1}d\sigma.$$

$$\pi(\sigma) = c^{-1}\pi(c^{-1}\sigma), \forall \sigma \& c.$$

The space of integration is same but the integrands are different.

Therefore $\pi(c) = c^{-1}\pi(1) = c^{-1} \forall c.$

$$\pi(\sigma) \propto \frac{1}{\sigma} \quad , \text{ not constant.}$$

↳ Scale invariant prior.

Note: The derivations of the non-informative prior in the two scenarios should not be considered as compelling. There is indeed a logical flaw in the analysis caused by the fact that the final priors are improper.

The difficulty arises in the argument that if the two problems have identical structures they should have the same informative priors. The problem here is that when improper, the non-informative prior are not unique.

Multiplying an improper prior by a constant K, results in an equivalent prior in the sense that all decisions and inferences in Bayesian Analysis will be identical for the priors π and $K\pi$. Thus there is no reason to insist that π and π^* are same rather they need only be constant multiples of each other.

In first scenario for instance the mild restriction will give instead of ① the relationship.

$$P^{\pi}(A) = h(c)\, P^{\pi^*}(A)., \text{ where } h(c) \text{ is positive}$$

function of c. The other equation ② should also ~~not~~ remain valid. i.e. $P^{\pi^*}(A) = P^{\pi}(A-c)$

Therefore $P^{\pi}(A) = h(c)\, P^{\pi}(A-c).$

Integrating both sides,

$$\int_A P\, \pi(\theta)d\theta = h(c) \int_{A-c} \pi(\theta)d\theta$$

$$= h(c) \int_A \pi(\theta - c)d\theta$$

$$\pi(\theta) = h(c)\, \pi(\theta - c).$$

letting $\theta = c$, $\pi(c) = h(c)\pi(0)$.

The conclusion is that π needs to satisfy,

$$\pi(\theta) = \frac{\pi(c)}{\pi(0)}\pi(\theta - c)$$

$$\Rightarrow \pi(\theta - c) = \frac{\pi(0)}{\pi(c)}\pi(\theta). \quad -(1)$$

There are many improper priors besides uniform which satisfy this relationship.

A prior satisfying this relationship is called relatively location invariant.

Jeffrey's Prior:

Suggestion: (i) $\pi(\theta) \propto$ a constant $(-\infty, \infty)$.

(ii) $\pi(\theta) \propto \frac{1}{\theta}$, $(0, \infty)$ (scale).

(iii) $\pi(\theta) \propto (I(\theta))^{1/2}$

↳ Fisher's Information.

$$I(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\log L\right\} = -n\,E\left\{\frac{\partial^2}{\partial\theta^2}\log f\right\}$$

$$= E\left[\frac{\partial}{\partial\theta}\log L\right]^2$$

In case of θ has multiple parameter

$$\pi(\theta) \propto |I(\theta)|^{1/2}$$

In general Jeffrey's prior provide improper prior but quite often it provides proper prior.

* Jeffrey's prior may give improper prior but mostly proper posterior.

Disadvantage:

Difficult to calculate like in the case of gamma distribution (di-gamma or tri-gamma).

Derivative of Gamma(k) with respect to k.

1. $X \sim \text{Normal}(\theta, \sigma^2)$

$$I(\theta) = -E\begin{bmatrix} \frac{\partial}{\partial\theta^2}\left[-\frac{(x-\theta)^2}{2\sigma^2}\right] & \frac{\partial^2}{\partial\theta\partial\sigma}\left[-\frac{(x-\theta)^2}{2\sigma^2}\right] \\ \frac{\partial^2}{\partial\theta\partial\sigma}\left[-\frac{(x-\theta)^2}{2\sigma^2}\right] & \frac{\partial^2}{\partial\sigma^2}\left[-\frac{(x-\theta)^2}{2\sigma^2}\right] \end{bmatrix}$$

$$= -E\begin{bmatrix} \frac{\partial}{\partial\theta}\left[2(x-\theta)/2\sigma^2\right] & \frac{\partial}{\partial\theta}\left[-\frac{(x-\theta)^2}{2}\cdot\frac{-2}{\sigma^3}\right] \\ \frac{\partial}{\partial\theta}\left[-\frac{(x-\theta)^2}{2}\cdot\frac{-2}{\sigma^3}\right] & \cdot\frac{\partial}{\partial\sigma}\left[-\frac{(x-\theta)^2}{2}\cdot\frac{-2}{\sigma^3}\right] \end{bmatrix}$$

$$= -E\begin{bmatrix} -1/\sigma^2 & \frac{-2}{\sigma^3}(x-\theta) \\ \frac{2(\theta-x)}{\sigma^3} & (x-\theta)^2.(-3/\sigma^4) \end{bmatrix}$$

$$= \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & \frac{3}{\sigma^4}.\sigma^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 3/\sigma^2 \end{bmatrix}$$

$$\pi(\theta) \propto |I(\theta)|^{1/2}$$

$$\propto (3/\sigma^4)^{1/2}$$

$$\propto 1/\sigma^2.$$

Hence the prior is considered to be proportional to $\frac{1}{\sigma^2}$.

It is interesting to show that the prior from an invariance under transformation argument.

A non-informative prior for the situation can also be derived by assuming independence of θ and σ and multiplying the non-informative prior obtained earlier. The result is

$$\pi(\theta,\sigma) \propto \frac{1}{\sigma} \quad -(A)$$

This is actually the non-informative prior recommended by Jeffrey's. Jeffrey's remarked that these sometimes gives an non-informative prior.

Another important feature of Jeffrey's Non-informative prior is that it is not affected by the restriction on parameter space. In such a situation we simply assume that, the non-informative prior is that which is inherited from the unrestricted parameter space.

Suppose $\varphi(\theta)$ is one to one function

$$I(\theta) = I\{\varphi(\theta)\} \cdot \left(\frac{\partial \varphi(\theta)}{\partial \theta}\right)^2$$

$$\Rightarrow I(\theta)^{1/2} = I\{\varphi(\theta)\}^{1/2} \left|\frac{\partial \varphi(\theta)}{\partial \theta}\right|$$

$$\Rightarrow I(\theta)^{1/2} d\theta = I\{\varphi(\theta)\}^{1/2} d\varphi(\theta)$$

The conclusion is that Jeffrey's prior preserves scale invariant parameter.

Let $X \sim N(\mu, 1)$

$$\pi(\mu) \propto 1.$$

$$\varphi(\mu) = \exp(\mu)$$

$$\left[I(\varphi(\mu))\right]^{1/2} = \{I(\mu)\}^{1/2} \left|\frac{\partial \varphi(\mu)}{\partial \mu}\right|^{-1}$$

$$\Rightarrow I\{\varphi(\mu)\}^{1/2} = 1 \cdot (\exp(\mu))^{-1} = \exp|-\mu|.$$

So, it preserves scale invariant parameter.

Critism in the use of non-informative prior

1. Since the method of deriving prior depends on the experimental structure one may avoid the likelihood principle in using non-informative prior.

2. Marginalisation paradox is another important criticism.
[Univariate → Joint distribution is possible.
But Joint → Marginal is not possible].

3. other criticism arrives from the fact that we are often in a position to take various non-informative priors and try to construct examples where they perform poorly

(*) Why comparision of Priors is vague?
→ priors are based on subjective probabilities or rather we can say beliefs. Comparision of different beliefes is can not be done.

4. most embarassing feature is that, often there are a number of non-informative priors.

Conjugate prior: (Raiffa & schlaifer) ↳1961

The name comes from the authors who presented a formal development of conjugate prior distribution. Intituitively a conjugate prior distribution say $g(\theta)$ for a given sampling distribution say $f(x,\theta)$ is such that the posterior distribution $g(\theta|x)$ and the prior $g(\theta)$ are the members of the same family.

So, in developing the theory, Raiffa schlaifer seek a family of priors to satisfy the following criterion —

1. Analytical Tractibility:
   a. Easy to determine posterior from the sample and prior distribution.
   b. Easy to obtain expectations of utility function of interest.
   c. Priors and posteriors should be members of the same family of distributions.

2. Flexible and Rich:
   A wide variety of information and belief should be describable by the family of priors.

3. Easy to interprete:
   The parameters (hyperp.) of the priors are such that an experimentalr can readily ~~relived~~ relate his/her beliefs to characterisize ~~of~~ the prior distributions.

Methodology to obtain conjugate priors

The methodologies relies of the concept of sufficient and complete statistics.

Let the likelihood function is

$$L(\underline{x},\theta) = \prod f(x_i|\theta)$$
$$= g(t,\theta)\, h(x)$$

where T is sufficient statistic.

Let us now assume $g(\theta|\alpha)$ represent a family of pdf, where α denotes the parameters that index the specific member of the family. If for every pair of pdf $g(\theta|\alpha_1)$ and $g(\theta|\alpha_2)$ there is another pdf $g(\theta|\alpha_3)$ such that

$$g(\theta|\alpha_3) = g(\theta|\alpha_1) \times g(\theta|\alpha_2).$$

we say that the family is closed under multiplication.

The above discussions allows us to construct a conjugate prior by determining a family of pdf the satisfy the following two criterions—

1. The likelihood function of θ for any given sample must be proportional to the member of family.
2. The family must be closed under multiplication.

Generally if sufficient statistic exists, the conjugate prior also exists.

- Let us consider an example.

$$f(x|\theta) = \underbrace{\binom{n}{x}}_{=h(x)} \underbrace{\theta^x (1-\theta)^{n-x}}_{\propto g(t,\theta).}, \quad x = 0, 1, 2, \ldots$$
$$\theta > 0, \quad 0 < \theta < 1.$$

Therefore $L(x,\theta) \propto \theta^x (1-\theta)^{n-x}$

$$\propto \text{Beta}(x+1, n-x+1)$$

- Product of two Beta Distributions,

$$B(\alpha_1, \beta_1) \text{ and } B(\alpha_2, \beta_2)$$

Then $B(\alpha_1, \beta_1) \times B(\alpha_2, \beta_2) = B(\alpha_1+\alpha_2+3, \beta_1+\beta_2+3)$.

(*) Efficiency and unbiasedness has no scope here.

(*) sufficiency and consistency is still there Bayesian Inference.

1. Let $x_1, x_2, \ldots, x_n$ from $B(1,\theta)$.

Here we have to find Jeffrey's prior of θ.

Solution: $x_1, x_2, \ldots, x_n \sim$ Bernoulli$(1,\theta)$.

$I(\theta)$=Fisher's information $= -nE\left[\frac{\partial^2}{\partial\theta^2}\log f\right]$

$$= -nE\left[\frac{\partial^2}{\partial\theta^2}\log\{\theta^x(1-\theta)^{x-1-x}\}\right]$$

$$= -nE\left[\frac{\partial^2}{\partial\theta^2}\log\{x\log\theta+(1-x)\log(1-\theta)\}\right]$$

$$= -nE\left[\frac{\partial}{\partial\theta}\left\{\frac{x}{\theta}+\frac{(1-x)}{(1-\theta)}(-1)\right\}\right]$$

$$= -nE\left[\frac{-x}{\theta^2}+\frac{-(1-x)}{(1-\theta)^2}\right]$$

$$= -n\left[-\frac{\theta}{\theta^2}+\frac{1-\theta}{(1-\theta)^2}\right]$$

$$= -n\left[-\frac{1}{\theta}+\frac{1}{1-\theta}\right]$$

$$= +n\left[\frac{+1-\theta+\theta}{\theta(1-\theta)}\right]$$

$$= \frac{n}{\theta(1-\theta)}$$

Then Jeffrey's prior is

$$\pi(\theta) \propto |I(\theta)|^{1/2}$$

$$\propto \sqrt{\frac{1}{\theta(1-\theta)}}$$

$$\propto \theta^{-1/2}(1-\theta)^{-1/2}$$

$\sim$ Beta$(\frac{1}{2},\frac{1}{2})$, This is proper prior.

What is the posterior distribution then, considering a random sample of size n.

$$p(\theta|x) = \frac{L(x,\theta)\,\pi(\theta)}{\int L(x,\theta)\,\pi(\theta)\,d\theta}$$

$$\propto L(x,\theta)\,\pi(\theta)$$

$$= \theta^{\Sigma x}(1-\theta)^{n-\Sigma x_i}\,\theta^{-1/2}(1-\theta)^{-1/2}$$

$$= \theta^{\Sigma x_i - 1/2}(1-\theta)^{n-\Sigma x_i - 1/2}$$

$\sim$ Beta$\left(\Sigma x_i+\frac{1}{2},\ n-\Sigma x_i+\frac{1}{2}\right)$

$$\hat{\theta}_B = \frac{\Sigma x_i + 1/2}{n+1}$$

2. $X \sim \text{Poisson}(\theta)$

$$f(x,\theta) = \frac{\theta^x \cdot e^{-\theta}}{x!}$$

$$I(\theta) = -n\, E\left[\frac{\partial^2}{\partial\theta^2} \log f\right]$$

$$= -n\, E\left[\frac{\partial^2}{\partial\theta^2} \log\left\{\frac{\theta^x e^{-\theta}}{x!}\right\}\right]$$

$$= -n\, E\left[\frac{\partial^2}{\partial\theta^2}\{x\log\theta - \theta + c\}\right]$$

$$= -n\, E\left[\frac{\partial}{\partial\theta}\left\{\frac{x}{\theta} - 1\right\}\right]$$

$$= -n\, E\left[-x/\theta^2\right] = -n \cdot \frac{-\theta}{\theta^2} = n/\theta$$

Then Jeffrey's prior will be

$$\pi(\theta) \propto |I(\theta)|^{1/2}$$

$$\propto |1/\theta|^{1/2}$$

$$\propto \sqrt{1/\theta} \quad \text{(improper)}$$

$$\propto \theta^{-1/2}(1-\theta)^{1-1} \quad \left.\right\} \times \rightarrow \text{will not form pdf}$$

$$\propto \text{Beta}\left(\tfrac{1}{2}, 1\right)$$

$$P(\theta|x) \propto L(x,\theta)\,\pi(\theta)$$

$$\propto \frac{\theta^{\Sigma x_i} e^{-n\theta}}{x_1!\, x_2! \ldots x_n!}\, \theta^{-1/2}$$

$$\propto \frac{\theta^{\Sigma x_i - 1/2} e^{-n\theta}}{x_1!\, x_2! \ldots x_n!}$$

$$\propto \theta^{\Sigma x_i - 1/2} e^{-n\theta}$$

$$P(\theta|x) \propto \text{Gamma}(n, \Sigma x_i + 1/2)$$

Squared error loss function is justified here.

3. $X \sim \text{Normal}(\mu, 1)$.

Jeffrey's prior $\pi(\mu) \propto$ [illegible].

posterior $\sim$ Normal

4. $X \sim N(0, \lambda)$, posterior $\sim$ Inverted Gamma

$X \sim N(0, \psi)$, $\psi = \lambda^{-1}$, is the precision parameter.

↳ posterior $\sim$ Gamma

5. conjugate prior for Poi(λ).

6. conjugate prior for Normal $(\mu, 1)$
Normal $(0, \sigma)$
Normal $(0, \varphi)$, $\varphi = 1/\sigma$

Inferences based on posterior distribution are insensitive based for the different choices of prior parameters large sample size.

3. solution:

$X \sim$ Normal $(\mu, 1)$

$$f(x, \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$

$$I(\theta) = -n E\left[\frac{\partial^2}{\partial \mu^2} \log f\right]$$

$$= -n E\left[\frac{\partial^2}{\partial \mu^2} \log\left\{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}\right\}\right]$$

$$= -n E\left[\frac{\partial^2}{\partial \mu^2}\left\{\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x-\mu)^2\right\}\right]$$

$$= -n E\left[\frac{\partial}{\partial \mu} - \frac{2}{2} x(x-\mu)(-1)\right]$$

$$= -n E\left[\frac{\partial}{\partial \mu}\left\{(x-\mu)\right\}\right]$$

$$= -n E[-1] = n$$

Jeffrey's prior will be $\pi(\theta) \propto |I(\theta)|^{1/n}$

$$\propto 1.$$

We consider $\pi(\theta) \propto c$, c be any real constant.

posterior distribution is

$$p(\theta|x) \propto 1 \cdot e^{\frac{1}{2}\Sigma(x_i-\mu)^2}$$

$$\propto e^{-\frac{1}{2}\Sigma(x_i-\mu)^2}$$

$$P(\theta|x) \propto e^{-\frac{1}{2}(\Sigma x_i^2 - 2\mu n\bar{x} + \mu^2 n)}$$

$$\propto e^{-\frac{n}{2}(\mu^2 - 2\mu\bar{x} + \bar{x}^2 - \bar{x}^2)} \cdot e^{-\frac{1}{2}\Sigma x_i^2}$$

$$\propto e^{-\frac{n}{2}(\bar{x}-\mu)^2}$$

$$\propto e^{-\frac{1}{2}\cdot\left(\frac{1}{1/n}\right)\cdot(\bar{x}-\mu)^2}$$

$$p(\theta|x) \propto \sim \text{Normal}(\bar{x}, 1/n).$$

4. $X \sim \text{Normal}(0, \lambda)$

$$f(x;\lambda) = \frac{1}{\sqrt{\lambda}\cdot\sqrt{2\pi}} e^{-\frac{1}{2}(x^2/\lambda)}.$$

$$I(\theta) = -n E\left[\frac{\partial^2}{\partial \lambda^2} \log f\right]$$

$$= -n E\left[\frac{\partial^2}{\partial \lambda^2} \log\left\{\frac{1}{\sqrt{\lambda}\sqrt{2\pi}} e^{-\frac{1}{2}(x^2/\lambda)}\right\}\right]$$

$$= -n E\left[\frac{\partial^2}{\partial \lambda^2}\left\{\log \sqrt{\lambda} - \frac{1}{2}\log(x^2/\lambda)\right\}\right]$$

$$= -n E\left[\frac{\partial}{\partial \lambda}\left\{\frac{-1}{\sqrt{\lambda}}\cdot\frac{1}{2\sqrt{\lambda}} - \frac{1}{2}\cdot x^2\cdot\left(\frac{-1}{\lambda^2}\right)\right\}\right]$$

$$= -n E\left[\frac{\partial}{\partial \lambda}\left\{\frac{-1}{2\lambda} + \frac{x^2}{2\lambda^2}\right\}\right]$$

$$= -n E\left[\frac{-1}{2}\left(\frac{-1}{\lambda^2}\right) + \frac{x^2}{2}\cdot\frac{-2}{\lambda^3}\right]$$

$$= -n E\left[\frac{+1}{2\lambda^2} - \frac{2x^2}{\lambda^3}\right]$$

$$= n E\left[\frac{-1}{2\lambda^2} + \frac{x^2}{\lambda^3}\right]$$

$$= n\left\{\frac{-1}{2\lambda^2} + \frac{1}{\lambda^2}\right\} = \frac{n}{\lambda^2}\left(\frac{1}{2}\right) = \frac{n}{2\lambda^2}$$

Jeffrey's prior $\pi(\theta) \propto |I(\theta)|^{1/2}$

$$\propto \left|\frac{1}{2\lambda^2}\right|^{1/2}$$

$$\propto \lambda^{-1}$$

$$p(\theta|x) \propto \frac{1}{(\sqrt{\lambda}\sqrt{2\pi})^n} e^{-\frac{1}{2}\Sigma\left(\frac{x^2}{\lambda}\right)}\cdot\frac{1}{\lambda}$$

$$\propto \frac{1}{\lambda(\sqrt{\lambda})^n} e^{-\frac{1}{2}\Sigma(x^2/\lambda)}.$$

$$\propto \frac{1}{\lambda^{n/2+1}} e^{-\frac{\Sigma}{2}(x^2/\lambda)}.$$

$$\propto \lambda^{-n/2} e^{-\frac{1}{2}\Sigma x^2/\lambda}\cdot\lambda^{-\left(\frac{n}{2}+1\right)}$$

$$\propto \left(\frac{1}{\lambda}\right)^{\frac{n}{2}+1} e^{-\frac{1}{2}\left(\frac{1}{\lambda}\right)\Sigma x^2}$$

$$p(\theta|x) \sim \text{Inverted Gamma}\left(\frac{\Sigma x_i^2}{2}, \frac{n}{2}\right)$$

Predictive Distribution and Inferences

$X \sim f(x, \theta)$.

$p(\theta|\underline{x})$, $x_1, x_2, \ldots, x_n \rightarrow$ observed data (informative)

y: future observation (yet to be observed).
↳ (predictive data).

Now, for one observation, $p(\theta|x) = \dfrac{f(x|\theta)\,\pi(\theta)}{\int f(x|\theta)\,\pi(\theta)\,d\theta}$

Intended distribution

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{\int f(x,y,\theta)\,d\theta}{\int f(x,\theta)\,d\theta}$$ → Trivariate distribution.

$$= \frac{f(y,x|\theta)\,\pi(\theta)\,d\theta}{\int f(x|\theta)\,\pi(\theta)\,d\theta}$$

[Assumption:
- Assuming present observations and future observations are independent.
- populations are considered to be identical]

continuing, we have.

$$f(y|x) = \frac{\int f(y|\theta)\cdot f(x|\theta)\,\pi(\theta)\,d\theta}{\int f(x|\theta)\cdot\pi(\theta)\,d\theta}$$

$$= \int f(y|\theta)\,p(\theta|x)\,d\theta.$$

Posterior Expectation of $Y|x$ is $E_{(\theta|x)}[f(y|\theta)]$

Pluggin Principle:
- Used in classical Inference for predictive purpose.

[θ is unknown. and the problem is we don't even have sample to infer about θ.]

(*) It is not logical to use samples of today to actual predictive θ and use it in $f(y|\theta)$ for future distribution. This is the plugin principle which is used in classical inference.

Bayesians are averaging $f(y|\theta)$ over all possible values of θ. so we are taking $E_{(\theta|x)}[f(y|\theta)]$ in Bayesian inference.

[Monte carlo Estimation:

$$\hat{\varphi}(\theta) = \frac{1}{n}\Sigma\varphi(\theta_i) = E[\varphi(\theta)] = \int \varphi(\theta)\,\pi(\theta)\,d\theta.$$

θ is the random variable and we take random samples based on that.]

If we have any samples (- - - -) from $p(\theta|x)$ then we can have the Monte carlo estimator $\hat{f}(y|\theta) = \frac{1}{n}\sum_{i=1}^{n} f(y|\theta_i)$.

(*) Predictive point estimator of y = predictive mean under squared error loss function.

for hypothesis testing $H_0: \theta \le \theta_0$ Vs $H_1: \theta > \theta_0$.

- It is based on notion of probability.
- We can compare $P(\theta \le \theta_0|x)$ and $P(\theta > \theta_0|x)$ and the one giving higher value will be considered.

If we are willing to test $H_0: y < y_0$ Vs $H_1: y > y_0$.

Simply obtain two predictive probabilities and compare.

Little difference is there which is about loss function.

The loss function here is defined as

~~$(\theta - y)^2$~~ $L(y - d(x))^2$

(1).

Let $x_1, x_2, \ldots, x_n \sim$ Bernoulli $(1, \theta)$.

$\theta \sim$ Beta (α, λ).

Find predictive distribution of $y|x$.

prior $\propto \theta^{\alpha-1}(1-\theta)^{\lambda-1}$

posterior distribution

$$p(\theta|x) \propto \theta^{\Sigma x_i + \alpha - 1}(1-\theta)^{n - \Sigma x_i + \lambda - 1}$$

$$p(\theta|x) \sim \text{Beta}(\Sigma x_i + \alpha,\ n - \Sigma x_i + \lambda).$$

Therefore $f(y|x) = \int_0^1 f(y|\theta).\,p(\theta|x)\,d\theta.$

$$= \int_0^1 \frac{\Gamma(n+\alpha+\lambda)}{\Gamma(\Sigma x_i + \alpha)\,\Gamma(n - \Sigma x_i + \lambda)}\,\theta^{y + \Sigma x_i + \alpha - 1}\,(1-\theta)^{n+1 - \Sigma x_i + \lambda - y - 1}\,d\theta$$

$$= \frac{\Gamma(n+\alpha+\lambda)}{\Gamma(\Sigma x_i+\alpha)\,\Gamma(n-\Sigma x_i+\lambda)} \cdot \frac{\Gamma(y+\Sigma x_i+\alpha)\,\Gamma(1-y+n-\Sigma x_i+\lambda)}{\Gamma(n+\alpha+\lambda+1)}$$

$$\hat{p} = \frac{1}{(n+\alpha+\lambda)} \cdot \frac{\Gamma(y+\Sigma x_i+\alpha)\,\Gamma(1+n-y+\lambda-\Sigma x_i)}{\Gamma(\Sigma x_i+\alpha)\,\Gamma(n-\Sigma x_i+\lambda)}$$

$$P(y=0|x) = \frac{n-\Sigma x_i+\lambda}{n+\alpha+\lambda}$$

$$P(y=1|x) = \frac{\Sigma x_i+\alpha}{n+\alpha+\lambda}$$

2. Poisson (λ) distribution with non-informative prior
   ↳ Negative Binomial.

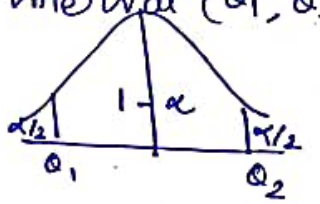
3. Normal Distribution $(\theta, 1)$. Find the predictive distribution considering a proper prior
   ↳ Final answer → Normal.

Bayes Interval Estimation:

$$\int_{\theta_1}^{\theta_2} P(\theta|x)\,d\theta = 1-\alpha = 0.95$$ → only one equation need another one to solve for θ_1 and θ_2.

The interval based on posterior distribution of θ to find $\int_{\theta_1}^{\theta_2} P(\theta|x)\,d\theta = 1-\alpha = 0.95$ is called credible interval with coverage probability $1-\alpha$. This interval contains $1-\alpha$ fraction of one's degree of belief such that $1-\alpha = P[\theta_1 < \theta < \theta_2]$. An equal tail $1-\alpha$ credible interval (θ_1, θ_2) is given by

$$\frac{\alpha}{2} = \int_{-\infty}^{\theta_1} P(\theta|x)\,d\theta = \int_{\theta_2}^{\infty} P(\theta|x)\,d\theta.$$



* To obtain the shortest $1-\alpha$ credible interval one has to minimize $I = \theta_2 - \theta_1$ subject to the condition

$$\int_{\theta_1}^{\theta_2} P(\theta|x)\,d\theta = 1-\alpha = 0.95 \quad -(1)$$

which requires $P(\theta_1|x) = P(\theta_2|x)$. $-(2)$

The interval which simultaneously satisfied (1) and (2) is called the shortest $(1-\alpha)$ credible interval.

An interval I which satisfies the following condition simultaneously

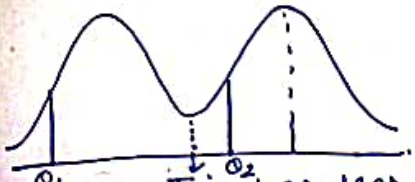
(i) The interval is shortest possible values of θ.

(ii) $p(\theta|x), \theta \in I > p(\theta|x), \theta \in I$.

The interval has more probable values of θ and excludes less probable values of θ.

This interval is called highest posterior density region. or HPD interval.

If the posterior distribution is unimodal, but not necessarily symmetric, the shortest interval and HPD interval are going to be same. This was first observed by Evans in 1976.



This has less probability than some point outside the interval. In this case obtaining HPD is not easy.

(*) This is Bayes interval with coverage probability $1-\alpha$ and not a confidence interval. [Bayer Estimation, SK Sinha]

Problem:

1. $f(x|\theta) = \frac{1}{\theta} \exp[-x/\theta]$

A random sample of size n, $x_1, x_2, \ldots, x_n$

$$L = \frac{1}{\theta^n} \exp(-s/\theta), \text{ where } s = \sum_{i=1}^{n} x_i$$

We consider Jeffrey's prior $g(\theta) \propto \frac{1}{\theta}$., scale parameter.

$$p(\theta|x) = \frac{s^n}{\theta^{n+1} \Gamma n} \exp(-s/\theta).$$

Now, $\frac{2s}{\theta} \sim \chi^2_{2n}$

Therefore $(1-\alpha) = P\left[\chi^2_1 < \frac{2s}{\theta} < \chi^2_2\right]$

$$= P\left[\frac{2s}{\chi^2_2} < \theta < \frac{2s}{\chi^2_1}\right],$$

where $\chi^2_1 = \chi^2_{1-\frac{\alpha}{2};2n}$

$\chi^2_2 = \chi^2_{\frac{\alpha}{2};2n}$

and $\chi^2(p;\nu) = 100p\%$ upper point of χ^2 with ν degrees of freedom.

Then equal tail credible interval is

$$c_1 = 2s/\chi^2_{\frac{\alpha}{2};2n}$$

$$c_2 = 2s/\chi^2_{1-\frac{\alpha}{2};2n}$$

To find the HPD interval,

$$(1-\alpha) = \int_{H_L}^{H_U} p(\theta|x)\,d\theta \quad \text{— ①}$$

$$p(H_L|x) = p(H_U|x), \text{ as chi-square is unimodal.}$$

Solving ①, we have

$$1-\alpha = \int_{H_L}^{H_U} \frac{s^n}{\theta^{n+1}\,\overline{n}} \exp\left[\frac{-s}{\theta}\right] d\theta.$$

$$\Rightarrow \exp\left[-s\left(\frac{1}{H_L} - \frac{1}{H_U}\right)\right] = \left(\frac{H_L}{H_U}\right).$$

Practical:

1. Random numbers from Uniform, using LCG method / convolution method

a sample of size 20.

2. exponential
3. Weibull
4. Binomial
5. Poisson
6. Gamma, with shape parameter > 1.
7. Normal, using Acceptance Rejection method
8. Lognormal
9. chi-square
10. Beta Distribution.
11. Double Exponential

19/04/23.

(*) Bayesian Interval and classical interval matches for exponential distribution. but interpretation is different.

Homework: ① Credible interval for $N(\mu, 1)$

② Credible interval for $N(0, \lambda)$.

③ Consider $B(1, \theta)$.

Take n samples prior $B(\alpha, \beta)$.

Posterior will also be $Beta(\alpha, \beta)$.

[Change value of β and see the effect of it on the posterior. Similarly do that with α]

⇒ Check of sensitivity of posterior distribution.

④ Consider $N(\mu, 1)$ where $\mu \sim N(\gamma, \lambda)$.

Decision Theory

$\theta \in \Theta$: parameter space. — values of θ.

$a \in A$: Action space.

$L(\theta, a)$: loss function

Nature never loses and the statistician is always the one who looses.

(Θ, A, L)

$X \sim f(x|\theta)$ → observable random variable.

Statistician is allowed to see x, before taking any decision.

After that statistician defines $d : \mathcal{X} \to A$.

He takes a decision, then loss function becomes $L(\theta, d(x))$.

Both becomes random.

Minimization of random quantity can not be done.

That is why we minimize

$$R(\theta, d) = E[L(\theta, d(x))]$$

↓ Risk.

In Bayesian, we assume $\theta \sim \pi(\theta)$.

Then Bayes Risk is defined as

$$r(\theta, d) = \int R(\theta, d)\, \pi(\theta)\, d\theta$$ — average taken with respect to different values of θ.

(*). Bayes Rule is one which has minimum Bayes Risk.

This is equivalent to minimization of posterior loss for different values of x.

Bayes Testing

Action → Finite Action

↳ Infinite Action

θ be the parameter on $\mathbb{R}$.

$H_0 : \theta \in \Theta_0$.

$H_1 : \theta \in \Theta_1$.

Let us consider we have group of action $\{a_1, a_2, \ldots, a_k\}$ with $L(\theta, a_i)$ when θ is the true value of the parameter while considering the action a_i.

Now, let $H_0 : \theta \in \Theta_0$

$H_1 : \theta \in \Theta_1$

and we have two actions a_0 and a_1.
(a_0 ↓ accept H_0; a_1 ↳ accept H_1.)

Then $L(\theta, a_0) = \begin{cases} 0, & \text{if } \theta \in \Theta_0 \\ 1, & \text{if } \theta \in \Theta_1 \end{cases}$

→ (When we are taking action a_0 and θ actually belongs to Θ_0, then we are committing 0 loss. In other case, we are committing maximum loss that is one).

Similarly $L(\theta, a_1) = \begin{cases} 0, & \text{if } \theta \in \Theta_1 \\ 1, & \text{if } \theta \in \Theta_0 \end{cases}$

Then posterior expected loss is

$$E_{\theta|x}[L(\theta, a_i)]$$

$$= \int_{\theta \in \Theta} L(\theta, a_1)\, p(\theta|x)\, d\theta$$

$$= \int_{\theta \in \Theta_0} 1 \cdot p(\theta|x)\, d\theta + \int_{\theta \in \Theta_1} 0 \cdot p(\theta|x)\, d\theta$$

$$= p(\Theta_0|x) \rightarrow \text{posterior probability of null hypothesis.}$$

Similarly, $E_{\theta|x}[L(\theta, a_1)] = p(\Theta_1|x)$.

The action with minimum posterior expected loss will be considered.

* When we want to accept Θ_0, minimize the posterior probability of other hypothesis Θ_1. That is maximize the posterior probability of that hypothesis.

Suppose $L(\theta, a_i) = \begin{cases} 0, & \text{if } \theta \in \Theta_i \\ k_i, & \text{if } \theta \in \Theta_j,\ j \neq i = 0, 1. \end{cases}$

$$E_{\theta|x}[L(\theta, a_i)]$$

$$= \int_{\theta \in \Theta_i} L(\theta, a_i)\, p(\theta|x)\, d\theta$$

$$= \int_{\theta \in \Theta_i} 0 \cdot p(\theta|x)\, d\theta + \int_{\substack{\theta \in \Theta_j \\ j \neq i}} k_i\, p(\theta|x)\, d\theta$$

$$= k_i\, p(\Theta_j|x).$$

$$E_{\theta|x}[L(\theta, a_0)] = \int_{\theta \in \Theta_1} L(\theta, a_0) p(\theta|x) d\theta$$

$$= K_0 P(\Theta_1|x).$$

$$E_{\theta|x}[L(\theta, a_1)] = \int_{\theta \in \Theta_0} L(\theta, a_1) p(\theta|x) d\theta$$

$$= K_1 P(\Theta_0|x).$$

The null hypothesis is rejected if

$$E[L(\theta, a_1)] < E[L(\theta, a_0)]$$

$$\Rightarrow K_1 P(\Theta_0|x) < K_0 P(\Theta_1|x).$$

$$\Rightarrow \frac{K_1}{K_0} < \frac{P(\Theta_1|x)}{P(\Theta_0|x)}$$

$$\Rightarrow \frac{K_0}{K_1} > \frac{P(\Theta_0|x)}{P(\Theta_1|x)} = \frac{1 - P(\Theta_1|x)}{P(\Theta_1|x)}.$$

$$\Rightarrow P(\Theta_1|x) > \frac{K_1}{K_0 + K_1}$$

According to classical statistical inference

$$c = \left\{x : P(\Theta_1|x) > \frac{K_1}{K_0 + K_1}\right\}.$$

Example: $X \sim \text{Normal}(\theta, \sigma^2)$ (σ^2 known)

prior of θ is $\theta \sim N(\mu, \tau^2)$, where both μ, τ^2 known

Find the posterior distribution.

Consider a single observation.

$$f_x(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2}.$$

posterior distribution is given as

$$P(\theta|x) \sim N\left(\mu(x), \rho^{-1}\right), \text{ where}$$

$$\mu(x) = \left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2}\right).$$

$$\rho = \frac{1}{\tau^2} + \frac{1}{\sigma^2}$$

The hypothesis is considered as

$$\left.\begin{aligned} H_0 &: \theta \geq \theta_0 \\ H_1 &: \theta < \theta_0 \end{aligned}\right\}.$$

Then Bayes test rejects H_0 if

$$P(\Theta_1|x) > \frac{K_1}{K_0 + K_1}$$

which implies

$$\frac{K_1}{K_0+K_1} < P(\Theta \in \Theta_1 | x) = \int_{-\infty}^{\theta_0} p(\theta|x)\,d\theta$$

$$= \int_{-\infty}^{\theta_0} \frac{\sqrt{\rho}}{\sqrt{2\pi}} \exp\left[-\frac{\rho}{2}(\theta-\mu(x))^2\right] d\theta$$

Considering the transformation

$$\eta = \sqrt{\rho}(\theta - \mu(x)).$$

we have,

$$P(\Theta_1 | x) = \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}} \exp\left[-\eta^2/2\right] d\eta$$

Therefore $\sqrt{\rho}(\theta_0 - \mu(x)) > z\left(\frac{K_1}{K_0+K_1}\right)$

This can be written as (by substituting values of $\mu(x)$ and ρ)

$$x < \theta_0 + \frac{\sigma^2}{\tau^2}(\theta_0 - \mu) - \rho^{1/2}\sigma^2 z\left(\frac{K_1}{K_0+K_1}\right).$$

~~That is~~ $x < \theta_0 + \sigma z(\alpha)$:

Here K_0 and K_1 are the parameters associated with loss function

Here in Bayesian Paradigm, the critical Region is based on all the parameters and hyperparameters we considered from the starting of the testing.

In classical Paradigm, an advoc quantity α is considered to perform any testing of hypothesis. This advocism is not present in Bayesian Paradigm. Hence Bayesian Inference is more rational.

⊛ In Bayesian Inference, there is nothing like level of significance. α does not exist in Bayesian Paradigm.

The next goal is to how to normalise the effect of prior distribution.

Bayes Testing

- Classical Testing :

  let $\theta \in \Theta$, then $H_0 : \theta \in \Theta_0$

  vs $H_1 : \theta \in \Theta_1$

We are assuming that

$$\Theta_0 \cup \Theta_1 = \Theta$$
$$\Theta_0 \cap \Theta_1 = \phi$$

They are the disjoint partitions.

Now, $x \sim f(x|\theta)$

$x_1, x_2, \ldots, x_n$: x is the random sample

1. We decide to reject H_0 when it is true → committed to type-I error.

2. We decide to accept H_0 when H_1 is true → committed type-I error.

Rejection Region:

$R = \{x$: observing x will lead to rejection of $H_0\}$.

$P(R|\theta)$, if $\theta \in \Theta_0$

$1 - R(R|\theta)$ if $\theta \in \Theta_1$

Difference with classical test:

In classical approach we consider the probability of R to which the observations does or does not belong consequently we are concern not merely with a single set of observation we actually made, but also with others we might have made but did not.

Example: Here $x \sim$ Normal$(\theta, 1)$

Then $H_0: \theta = 0$ vs $H_1: \theta > 0$.

Consider one observation $x = 3$.

Then the rejection region is $P[x \geq 3 | H_0] = 0.001350$ which implies rejects H_0 at $\alpha = 0.005$ level of significance.

While we are rejecting we are considering all observations ≥ 3., that we are not observed.

This is the problem of classical inference

Jeffrey's Remark:

" what the use of p-value implies, therefore is that a hypothesis that may be true may be rejected because it has not predicted the observable results that have not occured!"

Bayes Approach:

$p_0 = P[\theta \in (H)_0 | x]$ → posterior probability of H_0.

$p_1 = P[\theta \in (H)_1 | x]$ → posterior probability of H_1.

$\pi_0 = P[\theta \in (H)_0]$
$\pi_1 = P[\theta \in (H)_1]$
} prior probability of H_0 and H_1.

Now, $p_0 + p_1 = 1$ and $\pi_0 + \pi_1 = 1$

Then $\frac{\pi_0}{\pi_1}$ – prior odds on H_0 against H_1.

p_0/p_1 – posterior odds on H_0 against H_1.

If $\pi_0/\pi_1 \simeq 1$, then both apriori be
> 1, H_0 is more likely to accepted as a prior
< 1, H_1 is more likely to be accepted as a prior.

Then $\frac{p_0}{p_1} \simeq 1$, the both are same
> 1, H_0 is more likely to be accepted
< 1, H_1 is more likely to be accepted

Then Bayes factor is written as

$$B = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0 \pi_1}{p_1 \pi_0}$$

where B is the Bayes Factor in favour of H_0 against H_1 or posterior odds to prior odds.

The quantity B is used for testing of hypothesis.

OR $$p_0 = \frac{1}{\left[1 + \left\{\frac{1-\pi_0}{\pi_0}\right\} B^{-1}\right]}$$

proof: We know $B = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0 \pi_1}{p_1 \pi_0}$

$$\Rightarrow B \cdot \frac{p_1}{p_0} = \frac{\pi_1}{\pi_0}$$

$$\Rightarrow \frac{1-p_0}{p_0} = B^{-1}\left(\frac{1-\pi_0}{\pi_0}\right).$$

$$\Rightarrow \frac{1}{p_0} - 1 = B^{-1}\left(\frac{1-\pi_0}{\pi_0}\right).$$

$$\Rightarrow \frac{1}{p_0} = 1 + \left(\frac{1-\pi_0}{\pi_0}\right) B^{-1}.$$

$$\Rightarrow p_0 = \frac{1}{1 + \left(\frac{1-\pi_0}{\pi_0}\right) B^{-1}}$$

prior odds $= \pi_0/\pi_1$

Posterior odds $= p_0/p_1$

Bayes Factor $= \dfrac{p_0/p_1}{\pi_0/\pi_1}$

Testing simple vs simple hypothesis

$\mathcal{H}_0 = \{\theta_0\}$, $\mathcal{H}_1 = \{\theta_1\}$.

$p_0 \propto \pi_0 f(x|\theta_0)$

$p_1 \propto \pi_1 f(x|\theta_1)$

Then $p_0 = \dfrac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}$

$p_1 = \dfrac{\pi_1 f(x|\theta_1)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}$.

$\dfrac{p_0}{p_1} = \dfrac{\pi_0}{\pi_1} \dfrac{f(x|\theta_0)}{f(x|\theta_1)}$

$\Rightarrow B = \dfrac{f(x|\theta_0)}{f(x|\theta_1)}$ → Likelihood Ratio, this exactly parallels to what we have in classical testing.

When hypothesis are composite.

Let us define $g_0(\theta) = \dfrac{g(\theta)}{\pi_0}$, for $\theta \in \mathcal{H}_0$

$g_1(\theta) = \dfrac{g(\theta)}{\pi_1}$, for $\theta \in \mathcal{H}_1$

} both are normalized on their respective regions

where π_0 is prior probability

Then, $\int_{\mathcal{H}_0} g_0(\theta)\,d\theta = \dfrac{1}{\pi_0} \int_{\mathcal{H}_0} g(\theta)\,d\theta = 1.$ ($\int_{\mathcal{H}_0} g(\theta)\,d\theta = \pi_0$)

Now, $p_0 = P[\theta \in \mathcal{H}_0 | x]$

$= \int_{\mathcal{H}_0} p(\theta|x)\,d\theta$

$\propto \int_{\theta \in \mathcal{H}_0} g(\theta) f(x|\theta)\,d\theta$

$\propto \pi_0 \int_{\theta \in \mathcal{H}_0} f(x|\theta)\, g_0(\theta)\,d\theta$

Similarly, we can write

$p_1 \propto \pi_1 \int_{\theta \in \mathcal{H}_1} f(x|\theta)\, g_1(\theta)\,d\theta$

Therefore, $\dfrac{p_0}{p_1} = \dfrac{\pi_0 \int_{\theta \in \mathcal{H}_0} f(x|\theta)\, g_0(\theta)\,d\theta}{\pi_1 \int_{\theta \in \mathcal{H}_1} f(x|\theta)\, g_1(\theta)\,d\theta}.$

Then, Bayes Factor B is

$$\frac{p_0/p_1}{\pi_0/\pi_1} = B = \frac{\int_{\theta\in\Theta_0} f(x|\theta)g_0(\theta)d\theta}{\int_{\theta\in\Theta_1} f(x|\theta)g_1(\theta)d\theta}$$

↷ averaging with respect to $g_0(\theta)$ and $g_1(\theta)$

↓ weighted likelihood Ratio

If $g_0(\theta)$ and $g_1(\theta)$ are kind of vague, then weighted likelihood ratio can be considered as simple likelihood Ratio.

17/4

1. The Electro weak Theory predicted the existence of a new particle, the ω-particle of a mass m of 82.4 ± 1.1 GeV. Experimental results showed that such a particle existed to and had a mass 82.1 ± 1.7 GeV. Consider the mass to have a normal prior and a ~~normal~~ likelihood and assumed the values after $\pm$ signs represent no. of sta

If we are prepare to take both theory and experiment into account, this can be regard this mess to be less than 83.0 GeV. Calculate prior odds, posterior odds (2.47) and Bayes Factor (1.43). 2.43.

One-sided test:

consider $\theta_0 < \theta_1$ whenever $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$.
or $\theta_0 > \theta_1$ whenever $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$

From Bayesian viewpoint, there is nothing particular, except that the use of p-values might be advocated.

$X \sim N(\theta, \sigma^2)$, where σ^2 is known, single observation is considered.

$\pi(\theta) \propto 1$. → Jeffrey's prior.

Then $H_0: \theta \le \theta_0$ vs $H_1: \theta > \theta_0$.

posterior: $p(\theta|x) \sim \text{Normal}(x, \sigma^2)$.

$$\text{Then } p_0 = P[\theta \le \theta_0 | x]$$
$$= \int_{-\infty}^{\theta_0} p(\theta|x)\,d\theta$$
$$= \int_{-\infty}^{\theta_0} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\theta - x)^2\right] d\theta$$

Let us consider a transformation $z = (\theta - x)/\sigma$

$$\text{Then } p_0 = \int_{-\infty}^{(\theta_0 - x)/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}z^2\right] dz = \Phi\left(\frac{\theta_0 - x}{\sigma}\right).$$

1. $X \sim \text{Poisson}(\lambda)$

$$L \propto e^{-n\lambda}\, \lambda^{\Sigma x_i}$$

The prior is $\lambda \sim \text{Gamma}(\alpha, \beta)$.

$$g(\lambda) \sim \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$p(\lambda|x) = \frac{e^{-n\lambda}\, \lambda^{\Sigma x_i}\, \lambda^{\alpha-1} e^{-\beta\lambda}}{\int_0^\infty e^{-n\lambda} \lambda^{\Sigma x_i} \lambda^{\alpha-1} e^{-\beta\lambda}\, d\lambda}$$

$$= \frac{e^{-(n+\beta)\lambda}\, \lambda^{\Sigma x_i + \alpha - 1}}{\int_0^\infty e^{-(n+\beta)\lambda}\, \lambda^{\Sigma x_i + \alpha - 1}\, d\lambda}.$$

$$= \frac{e^{-(n+\beta)\lambda}\, \lambda^{\Sigma x_i + \alpha - 1}}{\frac{1}{(n+\beta)^{\Sigma x_i + \alpha}} \lfloor \Sigma x_i + \alpha \cdot} = \frac{(n+\beta)^{\Sigma x_i + \alpha}\, e^{-(n+\beta)\lambda}\, \lambda^{\Sigma x_i + \alpha - 1}}{\lfloor \Sigma x_i + \alpha}$$

Then predictive distribution is

$$f(y|x) = \int_0^\infty f(y|\lambda)\, p(\lambda|x)\, d\lambda .$$

$$= \int_0^\infty \frac{\lambda^{\alpha-1} e^{-\beta\lambda} \cdot \beta^\alpha}{\Gamma\alpha} \cdot \frac{(n+\beta)^{\Sigma x_i+\alpha} e^{-(n+\beta)\lambda} \cdot \lambda^{\Sigma x_i+\alpha}}{\Gamma(\Sigma x_i+\alpha)} d\lambda .$$

Then predictive distribution is

$$f(y|x) = \int_0^\infty f(y|\lambda) \cdot p(\lambda|x)\, d\lambda$$

$$= \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!} \frac{(n+\beta)^{\Sigma x_i+\alpha} e^{-(n+\beta)\lambda} \lambda^{\Sigma x_i+\alpha-1}}{\Gamma(\Sigma x_i+\alpha)} d\lambda .$$

$$= \frac{1}{\Gamma(\Sigma x_i+\alpha)} (n+\beta)^{\Sigma x_i+\alpha} \cdot \frac{1}{y!} \int_0^\infty e^{-(n+\beta+1)\lambda} \lambda^{\Sigma x_i+\alpha+y-1} d\lambda .$$

$$= \frac{(n+\beta)^{\Sigma x_i+\alpha}}{\Gamma(\Sigma x_i+\alpha)} \cdot \frac{1}{y!} \frac{\Gamma(\Sigma x_i+\alpha+y)}{(n+\beta+1)^{\Sigma x_i+\alpha+y}}$$

$$= \frac{1}{y!} \frac{\Gamma(\Sigma x_i+\alpha+y)}{\Gamma(\Sigma x_i+\alpha)} \left(\frac{n+\beta}{n+\beta+1}\right)^{\Sigma x_i+\alpha} \cdot \left(\frac{1}{n+\beta+1}\right)^y$$

$$= \frac{(\Sigma x_i+\alpha+y-1)!}{y!\,(\Sigma x_i+\alpha-1)!} \left(\frac{n+\beta}{n+\beta+1}\right)^{\Sigma x_i+\alpha} \left(\frac{1}{n+\beta+1}\right)^y .$$

$$= \binom{\Sigma x_i+\alpha+y-1}{\Sigma x_i+\alpha} \left(\frac{n+\beta}{n+\beta+1}\right)^{\Sigma x_i+\alpha} \left(1-\frac{n+\beta}{n+\beta+1}\right)^y .$$

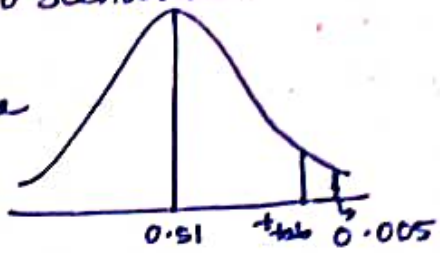
Continuation of one-sided test:

P-value: $P = P[T \geq t_{cal} \mid H_0]$

$= \int_{t_{cal}}^{\infty} T\, dt \simeq 0.005$ (0.51) $>$ two scenarios.

↓

Observed level of significance



Here $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$

$p\text{-value} = P[x \geq x \mid \theta = \theta_0]$

$= \frac{1}{\sigma\sqrt{2\pi}} \int_x^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x-\theta_0)^2\right] dx$

$= 1 - \Phi\left(\frac{x-\theta_0}{\sigma}\right)$

$= \Phi\left(\frac{\theta_0 - x}{\sigma}\right) = p_0$

Here posterior probability under is the p-value.

Then posterior probability p_1 is $1 - P\text{-value}$.

Then $\frac{p_0}{p_1} = \frac{p\text{-value}}{1 - p\text{-value}}$.

Then π_0 and π_1 are all infinity.

Then Bayes Factor is $B = \frac{p_0/p_1}{\pi_0/\pi_1}$.

Hence this strategy doesn't work for improper prior.

considering π_0 and π_1 to be very large, that is $\frac{\pi_0}{\pi_1} \simeq 1$

Then, Bayes Factor is $B \propto p_0/p_1 = \frac{p\text{-value}}{1 - p\text{value}}$.

Then $p_0 = P\text{-value} = (B^{-1}+1)^{-1}$.

$p_1 = (1+B^{-1})^{-1}$.

The result is significant if $p_0 \leq \alpha$,

which is equivalent to $p_1 \geq 1-\alpha$.

Testing point null hypothesis:

$H_0: \theta = \theta_0$

vs $H_1: \theta \neq \theta_0$

In classical, in this situation we use UMPU test.

Here, we are testing single point against a big interval.

The testing of point null hypothesis is often performed in inappropriate circumstance. It will virtually never be the cases, where one seriously entertains the hypothesis $H_0: \theta = \theta_0$. Exactly, a point which classical statistician also fully admits. A more reasonable null hypothesis can be $H_0: \theta \in \Theta_0 = (\theta_0 - \epsilon, \theta_0 + \epsilon)$, $\epsilon > 0$. where ϵ is choosen such that all θ belonging to Θ can be considered indistinguishable from Θ_0. The question arises, if a relalistic hypothesis is $H_0: \theta \in \Theta_0 = (\theta_0 - \epsilon, \theta_0 + \epsilon)$, $\epsilon > 0$, when it is reasonable to be approximated by $H_0: \theta = \theta_0$. From a Bayesian viewpoint- it will be reasonable if and only if the posterior probabilities are closed. This will happen if the likelihood function is considerably constant in $(\theta_0 - \epsilon, \theta_0 + \epsilon)$. But this is a very strong condition which is often difficult to obtain in real life.

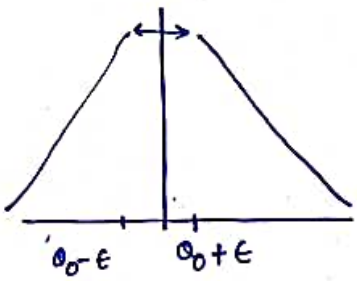
A more reasonable prior can be somewhat similar to

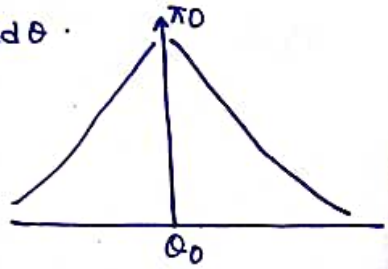
$$g(\theta) = \begin{cases} \pi_0 & , \text{if } \theta = \theta_0 \\ \pi_1 p_1(\theta) & , \text{if } \theta \neq \theta_0 \end{cases}, \quad \pi_0 + \pi_1 = 1.$$

(Even after so much ambiguity, we are taking $\theta = \theta_0$ only as it in that point we have most belief. Rest of probability is measured by $p_1(\theta)$ except the point $\theta = \theta_0$. Removing one odd point will not change density.

$\int p_1(\theta) d\theta$.

Here $\int_{\Theta} g(\theta) d\theta = \pi_0 + \pi_1 \int_{\theta \neq \theta_0} p_1(\theta) d\theta$.





Let $\underline{x} = x_1, x_2, \ldots, x_n$

$$p(\underline{x}) = \int f(x|\theta)\, g(\theta)\, d\theta \rightarrow \text{prior predictive density}$$

$$= \pi_0 f(x|\theta_0) + \pi_1 \int_{\theta \neq \theta_0} f(x|\theta)\, p_1(\theta)\, d\theta$$

$$= \pi_0 f(x|\theta_0) + \pi_1 p_1(x), \text{ say } p_1(x) = \int_{\theta \neq \theta_0} f(x|\theta)\, p_1(\theta)\, d\theta$$

Here, our objective is to find Bayes Factor.

Then, $p_0 = P[\theta \in \Theta_0 | x]$

$$= \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 p_1(x)}$$

$$= \frac{\pi_0 f(x|\theta_0)}{p(x)} \rightarrow \text{Total marginal.}$$

Similarly, $p_1 = P[\theta \in \Theta_1 | x]$

$$= \frac{\pi_1 p_1(x)}{\pi_0 f(x|\theta_0) + \pi_1 p_1(x)}$$

$$= \frac{\pi_1 p_1(x)}{p(x)}$$

Then Bayes Factor is $B = \frac{f(x|\theta_0)}{p_1(x)}$

If we have a sufficient statistic T, then

$$f(x|\theta) = f(t|\theta) \cdot f(x|t)$$

[In Bayes sufficiency, we can ~~zero~~ make posterior distribution more compact by removing $h(x)$.]

Then, $p_1(x) = \int p_1(\theta) f(t|\theta) f(x|t) \, d\theta$.

$$= f(x|t) p_1(t)$$

Therefore $p_0 = \frac{\pi_0 f(t|\theta_0)}{p(t)}$

$$p_1 = \frac{\pi_1 p_1(t)}{p(t)}$$

Then Bayes Factor will be $B = \frac{f(t|\theta_0)}{p_1(t)}$

Example:

Bayes Factor:

$x_1, x_2, \ldots, x_n \sim N(\theta, \varphi)$., φ is known.

$\bar{x} \sim \text{Normal}(\theta, \varphi/n)$.

Let $p_1(\theta) \sim \text{Normal}(\mu, \varphi)$ under H_1. (density defined everywhere except θ_0)

Assumption:

1. Let $\mu = \theta_0$ (?). It is sensible to take $\mu = \theta_0$ as values close to θ_0 are more likely to be true than those far away.

2. $\sqrt{\psi}$ is considerably greater than the width 2ϵ.
[values outside $(\theta_0-\epsilon, \theta_0+\epsilon)$ will also be considered but not to far away].

and we have

$\bar{x}-\theta \sim N(0, \psi/n)$ → independent of θ.

$(\bar{x}-\theta)$ and $p_1(\theta)$ are independently distributed.

Then $\tilde{x} \sim N(\theta_0, \varphi + \psi + \frac{\varphi}{n})$.

Then $p_1(\bar{x}) = \int p_1(\theta) \cdot f(\bar{x}|\theta)\, d\theta$

$= N(\theta_0, \psi + \frac{\varphi}{n})$ — ①

Then $B = \frac{f(\bar{x}|\theta_0)}{p_1(\bar{x})} = \frac{\{2\pi\varphi/n\}^{-1/2} \exp\left[-\frac{1}{2}(\bar{x}-\theta_0)^2/(\varphi/n)\right]}{\{2\pi(\psi+\frac{\varphi}{n})\}^{-1/2} \exp\left[-\frac{1}{2}(\bar{x}-\theta_0)^2/(\psi+\varphi/n)\right]}$

Therefore $B = \left\{1+\frac{n\psi}{\varphi}\right\}^{1/2} \cdot \exp\left[-\frac{1}{2}(\bar{x}-\theta_0)^2 \varphi^{-1} n \left\{1+\frac{\varphi}{n\psi}\right\}^{-1}\right]$

Now, $z = \frac{|\bar{x}-\theta_0|}{\sqrt{\varphi/n}}$

Therefore B can be written as

$$B = \left\{1+\frac{n\psi}{\varphi}\right\}^{1/2} \exp\left[-\frac{1}{2} z^2 \left\{1+\frac{\varphi}{n\psi}\right\}^{-1}\right]$$

Then $p_0 = \frac{1}{\left\{1+\frac{1-\pi_0}{\pi_0} B^{-1}\right\}}$

Lindley's Paradox:

(Also Known as Jeffrey's Paradox).

Let us consider $\pi_0 = 1/2$, $\psi = \varphi$

(variability of x ≃ variability of average of x)

Let $z = 1.96$

$n = 15$.

Then $B = (1+15)^{1/2} \exp\left[-\frac{1}{2}(1.96)^2 \left\{1+\frac{1}{15}\right\}^{-1}\right]$

$\simeq 0.66$

Then $p_0 = (1+0.66^{-1})^{-1} = 0.40$

$p_1 = 0.60$

Here $Z \sim N(0,1)$

Then $|Z| \geq 1.96$, will reject H_0 at 5% level of significance.

However, the Bayesian results provider a prior $p_0 = 0.4$.

2-tailed p-value	z	n = 1	5	10	20	50	100	..	10^3
0.1	1.645	0.418	0.442	0.442	0.558	...	.	...	0.891
0.05	1.960	0.351	0.331	0.367	0.424	...	....	....	0.823
0.01	2.576	0.212	0.134	0.140	0.163	...	....	....	0.535
0.001	3.291	0.086	0.021	0.026	0.021	0.031	...	....	0.121
									p_0

Here p_0 approaches to 1 as n increases irrespective of p-value. This implies that null hypothesis will not never be rejected. This is Lindley's Paradox.

Lindley's Paradox: (Jeffrey's observed it first)

The paradox concerns the fact that in testing H_0 vs H_1 with a fixed π_0 and P_1, a fixed α for each sample size, will results in $p_0 \to 1$ as $n \to \infty$, No matter how small α is. Thus when n is very large, the Bayesian test will frequently yield a posterior probability of H_0 near unity even when α is very small.

Observations:

1. We assume prior probability of null hypothesis is 1/2. $\pi_0 = \pi_1 = 1/2$. This assumption does seem natural and would be considered objective. In either case a slight change in the value of π_0 would not make much difference in the quali
2. We also assume prior density of θ, under alternative hypothesis was normal with mean θ_0 and variance ψ. In fact a precise change choice of $P_1(\theta)$ does not make a great deal of difference under $|\bar{x} - \theta_0|$ is large.
3. Lindley proved $P_1(\theta) \sim U\left(\theta_0 - \frac{\sqrt{\psi}}{2}, \theta_0 + \frac{\sqrt{\psi}}{2}\right)$ ob over an interval centered on θ_0. Jeffrey's argued to consider a cauchy distribution $P_1(\theta) = \frac{1}{\pi} \frac{\sqrt{\psi}}{\psi + (\theta - \theta_0)^2}$. Although his argument could not convince anyone else. Moreover, the choices of $P_1(\theta)$ ultimately resulted in more or less similar result that we have found earlier.

3. There is also a scale parameter ψ in the distribution of $p_1(\theta)$. Although it seems reasonable, that ψ_0 should be proportional to ψ. There does not seem to be any convencing argument for choosing this to have a particular value. It can be easily seem that effect of $\psi = k\psi$ on U and p_0 is just same as taking $\psi_0 = \psi = \psi$ and n is multiplied by a factor k. Obviously it is not Sensible to use a procedure that always provides the null hypothesis as a posterior probability of unity.

A Bound that dosen't depend on prior distribution.

$$p_1(\bar{x}) = \int p_1(\theta) f(\bar{x}|\theta) d\theta$$
$$\leq f(\bar{x}|\hat{\theta}), \quad \hat{\theta} \text{ is MLE}$$

Therefore $p_1(\bar{x}) \leq f(\bar{x}|\hat{\theta})$

$$= f(\bar{x}|\bar{x}), \quad \hat{\theta} = \bar{x} \text{ for normal distribution.}$$
$$= (2\pi\psi/n)^{-1/2}$$

Then Bayes Factor is $B = \dfrac{p(\bar{x}|\theta_0)}{p_1(\bar{x})}$

$$\geq \frac{(2\pi\psi/n)^{-1/2} \exp\left[-\frac{1}{2}(\bar{x}-\theta_0)^2/(\psi/n)\right]}{(2\pi\psi/n)^{-1/2}}$$

$$= \exp\left[-\frac{1}{2}(\bar{x}-\theta_0)^2/(\psi/n)\right]$$

Considering $z = \dfrac{|\bar{x}-\theta_0|}{\sqrt{\psi/n}}$

Then $B \geq \exp\left[-\frac{1}{2}z^2\right]$

The Bound is independent of prior distribution.

For $\pi_0 = 1/2$

p-value	z	Bound on B.	Bound on p_0
0.1	1.645	0.258	0.205
0.05	1.960	0.146	0.128
0.01	2.576	0.036	0.035
0.001	3.291	0.004	0.004.

Thus if $z = 1.96$, Bayes factor B is atleast 0.146. and p_0 is atleast 0.128.

1. Let $x_1, x_2, \ldots, x_n \sim N(\theta, \varphi)$, where φ is unknown.

Then Bayes Factor $B = \dfrac{f(x|\theta_0)}{p_1(x)}$

$$f(x|\theta,\varphi) = (2\pi)^{-1/2}\{\varphi^{-1/2}\exp\left[-\tfrac{1}{2}\frac{\theta^2}{\varphi}\right]\}\cdot\exp\left(\frac{x\theta}{\varphi} - \frac{1}{2}\cdot\frac{x^2}{\varphi}\right)$$

Then

$$L = \Pi f(x_i|\theta,\varphi)$$
$$= \varphi^{-n/2}\exp\left[-\tfrac{1}{2}\{s + n(\bar{x}-\theta)^2\}/\varphi\right]$$

where $s = \Sigma(x_i-\bar{x})^2$

Then $f(\underline{x}|\theta_0,\varphi) \propto \varphi^{-n/2}\exp\left[-\tfrac{1}{2}\{s + n(\bar{x}-\theta_0)^2\}/\varphi\right]$

we consider prior for φ as $g(\varphi) \propto \frac{1}{\varphi}$, as scale parameter

Then $f(\underline{x}|\theta_0) = \int f(x|\theta_0,\varphi)\, g(\varphi)\, d\varphi$

$$= \int f(\underline{x},\varphi|\theta_0)\, d\varphi.$$

$f(x|\theta_0) \propto (1 + t^2/2)^{-(\nu+1)/2}$, where $\nu = n-1$

$$t = \frac{\sqrt{n}(\theta_0-\bar{x})}{s}$$

$$s^2 = s/(n-1).$$

Then we have to find $p_1(\underline{x})$.

$g(\theta) \sim N(\theta_0, \phi)$.

$$p_1(x|\varphi) = \int f(x|\theta,\varphi)\cdot g(\theta)\, d\theta$$
$$\propto \varphi^{-n/2}\left(1+\frac{n\phi}{\varphi}\right)^{-1/2}\exp\left[-\frac{1}{2}\left(s + \frac{n(\bar{x}-\theta_0)^2}{1+\frac{n\phi}{\varphi}}\right)/\varphi\right]$$

Bayes factor $B = \dfrac{(1+t^2/\nu)^{-\frac{\nu+1}{2}}}{(1+nk)^{-1/2}\left\{1 + t^2(1+nk)^{-1}/\nu\right\}^{-\frac{\nu+1}{2}}}$

as $n \to \infty$

$$B = \frac{\exp\left[-\frac{1}{2}t^2\right]}{(1+nk)^{-1/2}\exp\left[-\frac{1}{2}t^2(1+nk)\right]^{-1}}$$

[A result which is used here is $\lim_{x\to\infty}\left(1+\frac{a}{x}\right)^x = e^a$].

Then $B \simeq (1+nk)^{1/2}\exp\left[-\frac{1}{2}t^2\frac{nk}{nk+1}\right]$.

As $n \to \infty$, $B \to \infty$.

Arnold Journal.

Bayesian Computation

Bayesian Statistics
- → Bayesian Theory.
- → Bayesian Applications
- → Bayesian Computation

Books
- Peter Congdon
- CP Robert

$X \sim f(x|\theta)$, $\theta = \theta_1, \theta_2, \ldots, \theta_k$, $k \geq 2$

samples: $x_1, x_2, \ldots, x_n$.

$$p(\theta|\underline{x}) = \frac{l(\theta)\, g(\theta)}{\int \cdots \int l(\theta)\, g(\theta)\, d\theta_1\, d\theta_2 \ldots d\theta_k}$$

proportionality not the case when sampling plan doesn't allow.

$$p(\theta_i|\underline{x}) = \int\int \cdots \int p(\theta|\underline{x})\, d\theta_i \quad , i = 1, 2, \ldots, k.$$

$$= \frac{\int \cdots \int l(\theta)\, g(\theta)\, d\theta_i}{\int \cdots \int l(\theta)\, g(\theta)\, d\theta_1 \cdots d\theta_k}$$

$\to (k-1)$ integrals.

$\to (k)$ integrals

$p(\theta_i \theta_j|\underline{x})$, $i \neq j = 1, 2, \ldots, k$.

$$= \int\int \cdots \int p(\theta|\underline{x}) \cdot d\theta_i\, d\theta_j \to (k-2) \text{ integrals.}$$

For trivariate distributions, $i \neq j \neq k = 1, 2, \ldots p$

$$p(\theta_i \theta_j \theta_k|\underline{x}) = \int\int\int \cdots \int p(\theta|\underline{x})\, d\theta_i\, d\theta_j\, d\theta_k \to (p-3) \text{ integrals.}$$

Then $E\{\varphi(\theta)|\underline{x}\} = \int\int\int \cdots \int \varphi(\theta)\, p(\theta|\underline{x})\, d\theta_1\, d\theta_2 \ldots d\theta_k \to k$ integrals.

If $\varphi(\theta) = \theta$, we will have posterior mean.

$\varphi(\theta) = \theta^2$, will give $E(\theta^2|x)$

Then posterior variance is $E(\theta^2|x) - \{E(\theta|x)\}^2$

That is $\text{var}(\theta|x) = E(\theta^2|x) - \{E(\theta|x)\}^2$

predictive distribution

$$p(y|\underline{x}) = \int \cdots \int f(y|\theta)\, p(\theta|x)\, d\theta$$

When θ have discrete distribution, Replace $\int$ by Σ.

And show that series is convergent for its existence.

But it is not logical always to take θ taking discrete values.

Until and unless we are very sure which values θ is taking.

• In continuous case, the integrals become very complex.

In 1702-1761, Bayesian Paradigm was not developed because of two reasons —

1. Prior Distribution → 1950- } 1990
2. Computation. → 1970-

1. Conjugate Analysis
2. Analytical / Asymptotic Approximation.
3. Numerical Integration.
4. Monte Carlo Simulation.
5. Markov chain methodology Monte Carlo simulation.

Emperical Bayes and Hierarchical Bayes

$\theta \sim g(\theta|a)$ → hyperparameter.

We give prior optimum amount of weightage.

Now, what should be the value of a.

when prior is insensitive to value of hyperameter.

This type of prior is called Robust prior.

posterior risk $(\underline{x}, a)$

plot posterior risk $(\underline{x}, a)$ V/S a. and take that a, the minimizes posterior risk.

Sandwich procedure: classical and Bayesian Mixture = Emperical Bayes.

$\theta \sim g(\theta|a)$ → a is estimated by classical Inference.

Then $g(\theta|\hat{a}_{classical})$ is considered as prior.

Hierarchical Bayes: $\theta \sim g(\theta|a)$ → hyperparameter.

prior is getting weak.

$a \sim g_1(a|b)$ → hyper-hyperparameter.

$b \sim g_2(b|c)$

⋮

$c \sim g_3(c)$ → Non-informative prior

$(\theta, a, b, c) \sim g(\theta|a) \cdot g_1(a|b) \cdot g_2(b|c) \cdot g_3(c)$.

$\int_c \int_b \int_a g(\theta, a, b, c) \sim h(\theta)$ → marginal distribution of θ.

This is hierarchical prior.

This is highly used in MCMC simulation.

Economeric ⇄ Bayesian Statistics.

Bayes computation

Conjugate Analysis

- flexible
- rich
- math tractability

1. $X \sim \text{Normal}(\mu, \sigma^2)$ ↳ known

$$\mu \sim g(\mu) = N(\mu_0, \tau_0^2)$$

$$x_1, x_2, \ldots, x_n \rightarrow p(\mu|x) = N\left(\frac{\tau_0^2 \Sigma x_i + \mu_0 \sigma^2}{n\tau_0^2 + \sigma^2}, \frac{\tau_0^2 \sigma^2}{n\tau_0^2 + \sigma^2}\right).$$

This is conjugate prior.

2. $X \sim N(\mu, \sigma^2)$ ↳ known

$\tau = \frac{1}{\sigma^2}$ = precision parameter.

i.e. $X \sim N(\mu, \tau)$.

prior of τ is $\tau \sim g(\tau) = \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right)$.

posterior $p(\tau|x) \sim \text{Gamma}\left(\frac{n+\delta_0}{2}, \frac{\gamma_0 + \Sigma(x_i - \mu)^2}{2}\right)$.

3. $X \sim \text{Normal}(\mu, \sigma^2)$, μ, σ^2 both unknown

$\tau = 1/\sigma^2$ precision parameter.

$g(\mu, \tau) = g_1(\mu|\tau) \cdot g_2(\tau)$.

considering $g_1(\mu|\tau) \sim N(\mu_0, \sigma_0^2/\tau)$.

$g_2(\tau) \sim \text{Gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right)$.

↳ Book: Raiffa & Schlaifer.

Family	prior
1. Binomial	Beta
2. Poisson	Gamma
3. Exponential	Gamma ($\theta e^{-x\theta}$) or Inverted Gamma ($\frac{1}{\theta} e^{-x/\theta}$).
4. Normal	Normal / Gamma
5. Multinomial	Dirichlet → (highly used in medical sciences / Biostatistics).

Asymptotic or Analytical Approximation.

1. Normal Approximation to the posterior:

Suppose we have the posterior density $p(\theta)$

↳ There be an unique mode → m. i.e. unimodal.

The posterior density must have a concave structure.

This method is not for Bimodal or multimodal distribution.

Taylor's Expansion says

$$\log p(\theta) = \log p(m) + \left[\frac{\partial}{\partial\theta}\log\theta\right]_m(\theta-m) - \frac{1}{2!}(\theta-m)'\left[\frac{-\partial^2}{\partial\theta\partial\theta'}\log p(\theta)\right](\theta-m) + R(\theta$$

where $R(\theta)$ denotes the remainder terms of order 3 and above, in the components of $(\theta-m)$ and can be *

$$= \log p(m) - \frac{1}{2!}(\theta-m)'\left[\frac{-\partial^2}{\partial\theta\partial\theta'}\log p(\theta)\right]_m(\theta-m) + R(\theta)$$

[Here $\left[\frac{\partial}{\partial\theta}\log\theta\right]_m = 0$, as m is mode, as obtained by solving $\frac{\partial}{\partial\theta}\log\theta = 0$, This will give $\hat{\theta} = m$.

(*) neglected in the approximation.

Generally we have the posterior mode of proportional $p^*(\theta) = cp(\theta)$, where

$$c = \int p^*(\theta)d\theta.$$

So, the expansion of $\log p^*(\theta)$ can be analogously done in the same way. Therefore $p^*(\theta)$ can be written as

$$p^*(\theta) = p^*(m)\exp\left[-\frac{1}{2}(\theta-m)'V^{-1}(\theta-m)\right]$$

where $c = p^*(m)(2\pi)^{k/2}|V|^{1/2}$

and $V = \left[\frac{-\partial^2}{\partial\theta\partial\theta'}\log p(m)\right]^{-1} = \left[-\frac{\partial^2}{\partial\theta\partial\theta'}\log p^*(m)\right]^{-1}$

which is the Hessian matrix of $\log p^*(\theta)$ at m.

Therefore $\theta \sim N(m, V)$ → Multivariate Normal.

Therefore the approximate Bayesian Inference proceeds as in normal case. The approximation is similar to the asymptotic result of MLE if the prior is constant.

Here m is the MLE and V is the asymptotic variance.

Any posterior can be always be approximated by a Normal distribution provided it is concave.

This is Bayesian version of Central limit theorem.

Difficulties:

1. Calculation of Derivations of logarithms of posterior distribution.
2. The most outstanding problem is Determination of posterior mode.
3. Difficulties to check if it is concave it not.

[Multidimensional Approx Maximization is required to approximate the mode, which is very much complicated.].

2. Standard Laplace Approximation:

[Also known as Lindley's Approximation].

One way to improve earlier approximation is to include higher order terms in Taylors series expansion of log posterior. The inclusion of third order terms leads to

$$p^*(\theta) \propto p^*(m) \exp\left[-\frac{1}{2}(\theta-m)'V^{-1}(\theta-m) + \frac{1}{3!}R(\theta)\right]$$

where $$R(\theta) = \sum_i \sum_j \sum_l p^{ijl}(\theta_i - m_i)(\theta_j - m_j)(\theta_l - m_l)$$

where $$p^{ijl} = \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_l} \log p^*(\theta)\Big|_m.$$

The previous approximation was of order $O(n^{-1})$.

But this approximation is of order $O(n^{-2})$. This is better than previous as it provides substantial improvement over the normal approximation discussed earlier. The only problem is evaluation of third of order derivatives and it can be real challenge if the dimension of θ is large enough.

Laplace Approximation (Exponential form).

posterior expectation of some function of θ, i.e. $\varphi(\theta)$ is

$$E[\varphi(\theta)|x] = \int \varphi(\theta)\, p(\theta|x)\, d\theta$$

$$= \frac{\int \varphi(\theta)\, l(\theta)\, g(\theta)\, d\theta}{\int l(\theta)\, g(\theta)\, d\theta}$$

We shall use the same idea of previous approximation that is to approximate the integrals by quadratic

form, but here this will be done separately for both denominator and numerator

Let $\varphi(\theta) > 0$

$\Rightarrow L(\theta) = \log l(\theta) + \log g(\theta)$

$L^*(\theta) = \log \varphi(\theta) + \log l(\theta) + \log g(\theta)$.

m^* $L^*(\theta)$ is the value that maximizes $L(\theta)$ and v^* is the negative of inverse Hessian of L^* at m^*.

Then $E[\varphi(\theta)|x] = \dfrac{\int \exp[L^*(\theta)]d\theta}{\int \exp[L(\theta)]d\theta}$.

Here L^* upto second order. is

$L^*(\theta) = L^*(m^*) - \frac{1}{2}(\theta - m^*)\, v^{*-1}(\theta - m)$

Then $E[\varphi(\theta)|x] = \left(\dfrac{|v^*|}{|v|}\right)^{1/2} \exp\left[L^*(m^*) - L(m)\right]$

If $\varphi(\theta) < 0$, add some constant to it, to make it positive and proceed similarly, although $\varphi(\theta)$ is rarely negative.

(*)(*) [Taylor's Expansion ≡ Asymptotic Approximation]

What is Analytical Approximation:

Integrands are difficult to visualize.

Numerical Integration

Consider the integral $I = \int_a^b g(\theta)d\theta$.

Quadratic rule approximate the integral $\hat{I} = \sum_{i=1}^{n} w_i g(\theta_i)$

A simple rule is to take n equally spaced points with equal weights for one dimensional problem taking n of order 10^2 happens to be sufficient for a good approximation. for further improvement the form of the integrand is taken under consideration.

Gaussian Rules:

Gaussian Rules were developed and regulated when the integrand is approximated by $\int h(\theta)\, p(\theta)d\theta$, where $h(\cdot)$ is a polynomial function of θ and $p(\theta)$ is the density function.

That is $\int g(\theta)d\theta = \int \dfrac{g(\theta)}{p(\theta)} \cdot p(\theta)d\theta$, where $h(\theta) = \dfrac{g(\theta)}{p(\theta)}$.

$= \int h(\theta)\, p(\theta)\, d\theta$

- When ~~where~~ $p(\theta) = U[-1, 1]$, the resulting formula is called Gauss-Legendre or Gauss Jacobi formula.
- When $p(\theta)$ is a gamma density, then $0 < p(\theta) < \infty$, then it is Gauss-Laguerre formation formula.
- When $p(\theta)$ is normal density, then the range of $p(\theta)$ is $-\infty$ to ∞, the resulting formula is Gauss-Hermite Quadrature Formula.

The approximation is exact is $h(.)$ is a polynomial function of degree 2 or less.

The approximation is appropriate if $h(.)$ is ~~a polynomi~~ well approximated by a polynomial of degree ~~2 and~~ $2n-1$.

In multivariate case, the integral may be evaluated by cartesian products with assumption of independence of components. One can use iterative strategies such as iterative product rules or iterative spherical strategies. An outstanding problem is that the number of function evaluation increases exponentially with the dimension of θ. The major drawback with this technique however is that they require unrealistic level of expertise on the part of users in selecting appropriate rules, starting values and specifying often appropriate reparameterization of the problem.

Smith (1991) .g wrote a paper about Analytical approach.

Bayesian Computation Via Simulation:

- Monte Carlo simulation
- MCMC.

Process of simulation

set of Assumption → simplified / complicated

↓

Derivation → managable / impossible

↓

Inferences.

⇒ It will be easier, if we directly switch over from the assumptions to final inferences without going through the derivations. This is what actually simulation does.

⇒ Simulation is the entire scenario of creating a hypothetical framework.

⇒ We can always generate random numbers from any pmf/pdf.

⇒ Given a random sample we can create the probability function.

pdf/pmf ⟷ Random sample

⇒ The main problem in Bayesian Inference is the posterior pdf. Here we don't have the actual pdf and we only have function which proportional to the actual pdf. The problem is that we don't to know the proportionality constant.

[Bootstrap:
• Bayesian Weighted Bootstrap is a non-parametric procedure.]

$x_1, x_2, \ldots, x_r, \ldots, x_n$ → censored observation.

we will consider likelihood function of $x_1, x_2, \ldots, x_r$ and survival function of remaining $(n-r)$ unobserved values.

For example, Normal and Gamma have no definite form of survival function. Then its become impossible to calculate the likelihood function.

Simulation / sample based approaches:
↳ posterior ∝ LF × prior.

1. Density Estimates ⇒ ⊙ Histogram
   ⊙ Kernel Density Estimates.
   ⋮
2. Emporical Distribution Function.
3. Consistent estimates of posterior expectation.
4. Predictive samples

} Sample based inferences

Standard Importance sampling

Monte Carlo Integral

Consider $I = \int t(\theta) p(\theta)\, d\theta$

$= \int \frac{t(\theta)}{q(\theta)} \cdot p(\theta) \cdot q(\theta)\, d\theta = E_q\left[\frac{t(\theta) p(\theta)}{q(\theta)}\right]$

suppose $\theta_1, \theta_2, \ldots, \theta_n$ is chosen from $q(\theta)$.

* $q(\theta)$ should be chosen in such a way samples can easily be generated from it.

Then Monte Carlo Estimator is

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} \frac{t(\theta_i)\, p(\theta_i)}{q(\theta_i)}$$

This is method of moment estimator of I.

- Here we are assuming all these observations are equally likely.
- It is also an unbiased estimator of I. i.e. $E(\hat{I}) = I$.
- It's variance is calculated as

$$\text{Var}(\hat{I}) = \frac{\sigma^2}{n}, \text{ where } \sigma^2 = \text{Var}\left\{\frac{t(\theta_i)\, p(\theta_i)}{q(\theta_i)}\right\}.$$

as θ_is are independent and identically distributed, so no covariance term

If $n \to \infty$, then $\text{Var}(\hat{I}) \to 0$

$\Rightarrow$ This estimator is becoming one of the best estimator.

There is a problem also, which is the more n, the better result. So, what should be the optimum value of n is an important question.

- Then we have $\frac{\hat{I} - I}{\sigma/\sqrt{n}} \sim N(0,1)$, as $n \to \infty$.

That implies $\hat{I}$ follows central limit theorem.

- $\hat{I}$ is also a consistent estimator of I.
  i.e. $\hat{I} \xrightarrow{P} I$, as $n \to \infty$.

But this properties are not that ~~im~~relevant in Bayesian Paradigm.

* Unbiasedness has no relevance here.
* The way variance is calculated is not that relevant here.
* CLT is calculated based on posterior distribution in Bayesian Paradigm, not based on sample values only.
* Consistency is considered highly Bayesian Inference.

To calculate the estimate of I, i.e. $\hat{I}$, we calculate $\hat{I}$ for different sizes of n.

500	$\hat{I}_0$
1000	$\hat{I}_1$
2000	$\hat{I}_2$
⋮	⋮
5000	$\hat{I}_3$
10,000	$\hat{I}_4$

} consider upto required decimal placer.

⊛ Important question is, are the samples really independent.

The generating $q(\cdot)$ is usually called importance density and sampling from $q(\cdot)$ is called Importance sampling.

sometimes $q \propto p$, and this will help in minimizing the variance of.

» Normal, Gamma, Uniform. → this choices of q usually helps the algorithm to work well.

Although choice q is a challenge.

Sampling - Importance - Resampling

$p(\theta|x) \propto LF \times$ prior.

suppose we have a random sample from prior.

» The problem is can we update the random sample obtained from prior and get random samples from posterior distribution if the updating mechanism is present.

→ The answer is Yes.

» What if the prior is improper. That is the prior ~~of~~ of θ is not a pdf. $p_1 \propto L_1 g_1(\theta)$.

$p_2 \propto L_2 g_2(\theta)$.

Then $p_1 \propto \left\{\frac{L_1 g_1(\theta)}{L_2 g_2(\theta)}\right\} \cdot p_2$.

Now we take random samples p_2 to update it to get the samples from p_1. Where p_2 is the dummy posterior.

If we take $L_1 = L_2 = L$, then $p_1 \propto \left\{\frac{L \cdot g_1(\theta)}{L \cdot g_2(\theta)}\right\} \cdot p_2$

$\Rightarrow p_1 \propto \left\{\frac{g_1(\theta)}{g_2(\theta)}\right\} p_2$

We can change either prior or likelihood or both to generate sample from desired posterior. This is called sampling-resampling

» Consider a random sample from $\theta_1, \theta_2, \ldots, \theta_n \sim h_1(\theta)$ (pdf)

We requires a sample from and another density $h(\theta)$.

They are somewhat ~~to~~ related to each other. can ~~go~~ we generate from $h(\theta)$ using $h_1(\theta)$.

The problem is, given a positive function $h_2(\theta)$ which is normalisable to $h(\theta)$ such that

$$h(\theta) = \frac{h_2(\theta)}{\int h_2(\theta)\, d\theta}.$$

* Can we get to random samples from $h(\theta)$ using $h_1(\theta)$ and $h_2(\theta)$. The answer is yes.

Sampling-Importance Resampling (weighted Bootstrap)

$\theta_1, \theta_2, \ldots, \theta_n \sim h_1(\theta)$

We require sample from $h(\theta)$.

and $h(\theta) = \frac{h_2(\theta)}{\int h_2(\theta) d\theta}$

There are two methods:

1. Rejection Method
2. Weighted Bootstrap Method.

Rejection Method:

Reject if $\frac{h_2(\theta)}{h_1(\theta)} \leq H$, where H is the upperbound such that $H = \sup_{\theta} \frac{h_2(\theta)}{h_1(\theta)}$

In acceptance-Rejection method if $H=1$, then we get one-to-one. That is every generated random number is accepted.

* Best envelope is that where two distribution overlap each other
- ⊙ In order to have most efficient algorithm, we need to have $H=1$.
- ⊙ Rejection method is not that efficient.
- ⊙ Rejection method says retain those θ_i for which that ratio large.

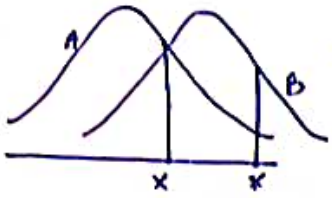
Algorithm:

1. $\theta \sim h_1(\theta)$
2. $U \sim U[0,1]$
3. If $U \leq \frac{h_2(\theta)}{H h_1(\theta)}$ accept θ.
4. Otherwise repeat 1-3.

If ratio $\frac{h_2(\theta)}{H h_1(\theta)}$ is very small $(10^{-2}, 10^{-3})$ probability is very much less. If $\frac{h_2(\theta)}{H h_1(\theta)} \geq 1$ definitely we the generated random number will be accepted.

If $h_2(\theta) \gg h_1(\theta)$, then $\frac{h_2(\theta)}{M h_1(\theta)}$ is going to be large.

If numerator is small enough → high probability of rejection.



If generated random number is in high probable region of target distribution then it is likely to accepted. If in low probable region then it is likely to be rejected.

Now, $M = \sup_{\theta} \frac{h_2(\theta)}{h_1(\theta)}$ is not always easy to obtain. This is why Rejection method is mostly fails.

challenges:
1. Finding envelope distribution
2. Maximization in high dimension.

We have No Bound in sampling Importance Resampling

sampling: Let $\theta_1, \theta_2, \ldots, \theta_n \sim h_1(\theta)$.

$$w_i = \frac{h_2(\theta_i)}{h_1(\theta_i)}.$$

and $q_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$

and $\sum_{i=1}^{n} q_i = 1$.

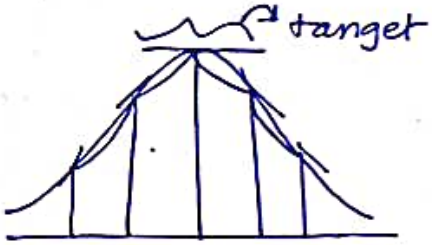
We consider a discrete distribution,

Values: $\theta_1 \; \theta_2 \; \ldots \; \theta_n$
prob : $q_1 \; q_2 \; \ldots \; q_n$

This a is PMF. (Resampling)

Then SIR takes observation θ^* from the above PMF.

Similar concept is here. The highly probable value is drawn.

[

If concave take two three points on both sides and draw tangents for each point This tangents will help us to draw two envelopes, one upper and one lower. Then squeeze them to find the actual distribution. This is called squeezing principle].

⊙ In usual bootstrap, all samples are equally likely. In Weighted Bootstrap, they are not because of the weights associated with samples.

⊙ Here n should be sufficiently large, in order that θ^* becomes an observation from $h(\theta)$.

* The lesser the resemblance between $h_1(\theta)$ and $h(\theta)$, the more the process is going to be difficult. Then we need to have a large large value of n.

Let $p(\theta|x) \propto l(\theta, x) \cdot g(\theta)$.

Then $H = l(\hat{\theta}, x)$ ↓ MLE.

Here $h_2(\theta) = p(\theta|x)$

$h_1(\theta) = g(\theta)$.

Then $\frac{h_2(\theta)}{H h_1(\theta)} = \frac{p(\theta|x)}{H g(\theta)} = \frac{l(\theta, x)}{l(\hat{\theta}, x)}$

The θ will be selected when the likelihood is large enough for each. That is, θ in the prior is sample will get selected in the posterior sample, if it has large likelihood value, is more likely to be retained.

Taking $q_i = \frac{l(\theta_i, x)}{\sum_{j=1}^{n} l(\theta_j, x)}$.

When prior is improper, and samples can not be generated from it and

$$p_2(\theta) \propto \frac{l_2(\theta) g_2(\theta)}{l_1(\theta) g_1(\theta)} p_1(\theta)$$

$$p_2(\theta) \propto \gamma(\theta) p(\theta) \text{ where } \gamma(\theta) = \frac{l_2(\theta) g_2(\theta)}{L_1(\theta) g_1(\theta)}$$

$x_1, x_2, \ldots, x_r, \underbrace{x_{r+1}, \ldots, x_n}_{>r}$ and unobserved.

$L_2 = \prod f(x_i, \theta) \cdot P(x_i > x_r)$ ← survival function

$p_2 \propto L_2 g_2$. , for gamma/normal distribution the survival function can not be obtained in closed form.

So, L_2 can not be calculated directly.

Solution:

Consider only first r observation. Then $L_1 = \prod f(x_i, \theta)$ will be in closed form.

We consider $p_1 \propto L_1 g_2$

Generate random numbers from $p_1(\theta)$.

Then we can use $p_2(\theta) \propto \frac{l_2(\theta) g_2(\theta)}{l_1(\theta) g_1(\theta)} p_1(\theta)$.

This is an advantage of sampling resampling method.

Markov-chain Monte Carlo Simulation

1983-1984 Gelfand and Smith (1991).

This was first developed for high dimensional posterior. Although this can be also for low dimensional as well.

Key-Idea for MCMC simulation:

Suppose we wish to generate a sample from a distribution $\pi(\theta)$, $\theta \in \Theta \subseteq R^k$. from which we are interested for simulation. However, suppose we can construct a markov chain with state space Θ and equilibrium distribution is $\pi(\theta)$. If we then run the chain for a long time, simulated value of the chain can be used as a basis for summarizing features of $\pi(\theta)$ of interest. Under suitable regularity conditions asymptotic results exists which clarify the since in which the sample output can be used as a random sample from $\pi(\theta)$.

Let $\theta^1, \theta^2, \ldots, \theta^t, \ldots$ these are the realizations of the markov chain.

Then $\theta^t \xrightarrow{d} \theta \sim \pi(\theta)$

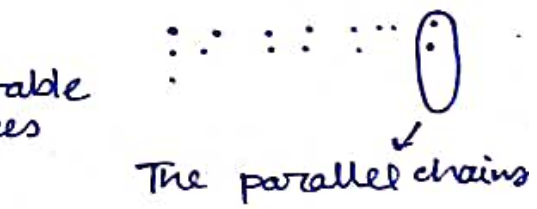
Then $\frac{1}{t} \sum_{i=1}^{t} \phi(\theta^i) \xrightarrow[n \to \infty]{a.s} E_\pi(\phi(\theta))$.

If $\phi(\theta) = \theta$, then $\frac{1}{t} \sum_{i=1}^{t} \theta^i \xrightarrow[t \to \infty]{a.s} E_\pi(\theta)$.

} Ergodic average.

Here we need iid sample. How can we do that?

Two ways → single chain with suitable spaces
→ parallel chains.

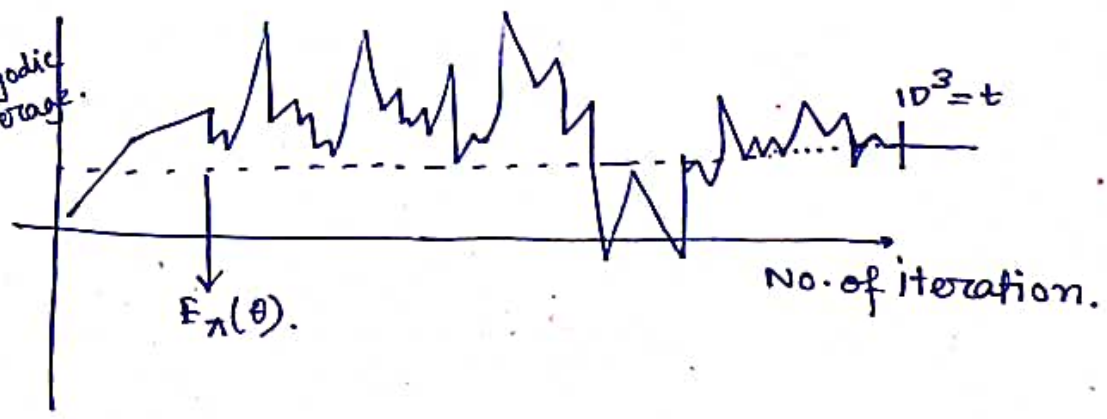


θ^1

$\frac{\theta^1 + \theta^2}{2}$

$\frac{\theta^1 + \theta^2 + \theta^3}{3}$

$\vdots$



$\theta^1, \theta^2, \ldots, \theta^{10^3} \ldots, \theta^{10^3+10} \ldots$

↓ 1st sample ↓ second

1. Gibbs Sampler.
2. Metropolis Hasting Algorithm

Gibbs Sampler

Geman & Geman (1984).

↳ Image Restoration based on Gibbs theorem.

Definition:

Gibbs sampler is a Markovian updating scheme which proceeds to be iterated sampling from various full conditionals specified upto proportionality from the joint posterior, treating for each θ_i in turn of all other quantities as fixed known constants. The joint posterior also needs to specified upto proportionality only.

$$\pi(\theta) \to p(\theta|x) \quad , \theta \in \Theta \subseteq \mathbb{R}^K$$

Gibbs sampler give us an updating mechanism.

$$p(\theta_1 | \theta_2, \ldots, \theta_K, x) \propto LF \cdot g(\theta_1, \ldots, \theta_K)$$

full conditionals

K unidimensional conditions:

$p(\theta_1 | \theta_2, \ldots, \theta_K, x)$.

$p(\theta_2 | \theta_1, \theta_3 \ldots, \theta_K, x)$

$p(\theta_3 | \theta_1, \theta_2, \theta_4, \ldots, x)$

$\vdots$

$p(\theta_K | \theta_{-K}, x)$. ← known

Initial values $\theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_K^0)$

Then $p(\theta_1 | \theta_2^0, \theta_3^0, \ldots, \theta_K^0, x) \sim \theta_1^1$.

↑ variable

One the basis of this we generate one observation that is θ_1^1.

Then θ_1^1 is now the most recent value.

Then $p(\theta_2 | \theta_1^1, \theta_3^0, \theta_4^0, \ldots, \theta_K^0, x) \sim \theta_2^{1}$.

Then now we have θ_1^1, θ_2^1 as upgraded value.

similarly, $\theta_3^1 \sim p(\theta_3 | \theta_1^1, \theta_2^1, \theta_4^0, \ldots, \theta_K^0, x)$.

$$\vdots$$

$$\theta_K^1 \sim p(\theta_K | \theta_{-K}^1, x).$$

Then we have $\theta^1 = (\theta_1^1, \theta_2^1, \ldots, \theta_K^1)$.

Then we will calculate again all the full conditionals by $\theta^1 = (\theta_1^1, \theta_2^1, \ldots, \theta_K^1)$.

Then $\theta_1^2 \sim p(\theta_1 | \theta_2^1, \theta_3^1, \ldots, \theta_K^1, x)$.

$\theta_2^2 \sim p(\theta_2 | \theta_1^2, \theta_3^1, \theta_4^1, \ldots, \theta_K^1, x)$

$$\vdots$$

$\theta_K^2 \sim p(\theta_K | \theta_{-K}^2, x)$

Then we have $\theta^0, \theta^1, \theta^2, \ldots, \theta^t, \ldots$

For Ergodic Average, the sum $A^t = \frac{1}{t}\sum_{i=1}^{t} \theta^i$ should converge.

We need to check if $|A^t - A^{t+1}| \leq \text{ACU} \Rightarrow 10^{-3}$.

This was the strategy was for single chain.

For parallel chain, the entire process is repeated for different set of initial values.

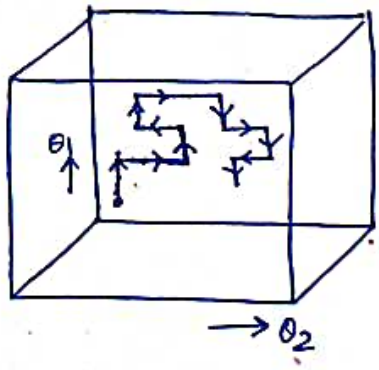
If we consider $K=2$, (minimum 2 is required for Gibbs sampler).

Then full conditional is $p(\theta_1 | \theta_2, x)$

$p(\theta_2 | \theta_1, x)$

[Best ~~example~~ view of Bivariate Normal is to hang Normal density curve by its mode. Then it will give a three dimensional density curve].

{ Univariate → 2D graph.
{ Bivariate → 3D.



First we fix θ_2 and we move along the axis θ_1.

Then again fix θ_1 and move along θ_2.

If the process is repeated.

This will give an idea of the surface.

And at each point we note the outcomes. and calculate the mean.

$\theta^0 = (\theta_1^0, \theta_2^0)$

(θ_1^1, θ_2^1)

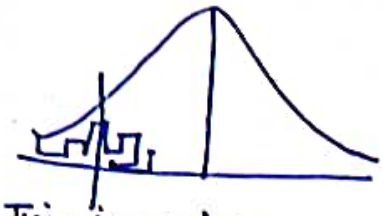
(θ_1^2, θ_2^2)

$\vdots$

(θ_1^t, θ_2^t)

We calculate the means to find the centre of gravity.

But the move actually determines ~~the centre of gravity~~. how much time / iterations it will take to find actually the centre of gravity.



This is only concentrated in tail.

For initial value, we can sometimes consider an estimate (preferably MLE). If not MLE any estimate can be considered as we are eventually compromising.

less dense

Easy convergence.

$corr(\theta_1, \theta_2)$ is very weak, so the mover are easy.

high intra molecular force

high density

convergence is difficult.

Mover are very very small.

correlation (θ_1, θ_2) is very strong.

Solution! Reduce $corr(\theta_1, \theta_2)$.

How to reduce:

Appropriate transformation of y and x can reduce correlation.

Example: Logarithmic transformation.

(*) When some variates are highly correlated the Gibbs Sampler will disrupt the entire process.

$\theta_1, \theta_2, \ldots, \theta_k \quad . \quad (\theta_2, \theta_3), \ldots, \underbrace{(\theta_{10}, \theta_{11})}_{\text{high correlated.}}$

If we don't want to loose both θ_{10}, θ_{11} we consider joint full conditional.

$$p(\theta_{10}, \theta_{11} | \ldots)$$

If all are correlated then we will have full joint full dimensional joint $\{p(\underbrace{\theta_1, \theta_2, \ldots \theta_{10}, \theta_{11}}_{\text{Original posterior (impossible to deal with)}})$

Two solution:

1. transformation.
2. Joint full conditional.

Difficulties in transformation:

1. Not able to find proper transformation.
2. high dimensional Jacobian.

Drawbacks of Gibbs sampler:

1. When the distribution is highly peaked and very less variability or probability difference between two points is large enough.

* winbux software for windows Bayesian Analysis.

$\theta^0, \theta^1, \ldots, \theta^t$.

where $\theta_1 \sim p(\theta_1|x) \rightarrow$ posterior marginal of θ_1.

$\theta_2 \sim p(\theta_2|x)$

componentwise sampler from ~~marginal~~ posterior is obtaing.

$(\theta_1, \theta_2) \sim p(\theta_1, \theta_2|x)$.

Final sample : $\theta_1, \theta_2, \ldots,$ on.

Each θ_i have got k components.

Hence we have $\theta_1, \theta_2, \ldots$ on iid $p(\theta|x)$.

Sometimes, we are interested in $p(\varphi(\theta)|x)$.

One solution to find first the posterior $p(\varphi(\theta)|x)$ by multiplying Jacobian etc...

Another solution is, simply replace values in $\varphi(\theta)$. and then we have $\varphi(\theta^0), \varphi(\theta^1), \ldots, \varphi(\theta^t), \ldots$

These will constitute the sample from $p(\varphi(\theta)|x)$.

Metropolis Hastings Algorithm: Metropolis Algorithm:

- Metropolis and others (1953)
- Metropolis and Hastings (1972).
- This is the ground of Bayesian compution 1990-1995.

Gibbs sampler:

- It requires availability of full conditionals.
- High correlated variates.

In metropolis Hastings Algorithm, no such requirement of full conditionals.
High correlated variates are also easily treated.
There is one transformation orthogonal transformation [orthogonalization].
M.H Algorithm is more efficient than Gibbs sampler.

$\Theta_1, \Theta_2, \ldots \Theta_t, \ldots \Theta$: state space.
$\pi(\theta)$ is equilibrium distribution → posterior.
$q(\theta, \theta')$ be any markovian kernel.
(candidate generating density).
Then suppose current value of θ is $\Theta_t = \theta$.
Then it generate another value $\theta' \sim q(\theta, \theta')$.
But θ' is selected with some probability α.
accept θ' and allow the chain to move.
Then we have $\Theta_{t+1} = \theta'$
(Next stage of markovian chain.)
If it is accepted then θ' otherwise the chain remains in the previous part.

* There can be repeatations.
* We need to check if the observations are mostly accepted so, the chain varies.

This setup defines the markovian chain, and α, the acceptance probability is defined as

$$\alpha(\theta, \theta') = \begin{cases} \min\left\{\frac{\pi(\theta')}{\pi(\theta)}, \frac{q(\theta', \theta)}{q(\theta, \theta')}, 1\right\} & \text{if } \pi(\theta) q(\theta, \theta') > 0 \\ 1, & \text{if } \pi(\theta) q(\theta, \theta') = 0 \end{cases}$$

MH Algorithm ↑

The larger the ratio, the larger the acceptance rate.

If q is symmetric kernel, then generating θ' keeping θ
≡ Generating θ keeping θ' as initial value.
i.e. $q(\theta, \theta') = q(\theta', \theta)$.
Then α term simplifies $\alpha(\theta, \theta') = \min\left\{\frac{\pi(\theta')}{\pi(\theta)}, 1\right\}$.
This ~~simplies~~ is Metropolis Algorithm.

Then replace $\pi(\theta)$ → by $p(\theta|x)$

As we consider posterior in Bayesian inference.

$p(\theta|x)$ = c.L.F. prior.

Metropolis Algorithm as well as Metropolis Hastings algorithm requires posterior upto proportionality only.

Consider $\alpha(\theta, \theta')$ as uniform$(0,1)$.

whatever θ' we have generated, if θ' provides larger posterior probability that it is accepted. If it makes the posterior smaller than $p(\theta'|x)$ then it is more likely to be rejected. This is kind of similar to the acceptance rejection algorithm.

The choice of q:

If $\theta : \theta_1, \theta_2, \ldots, \theta_k$

We consider multivariate normal kernel with mean vector θ and variance-covariance matrix $c\Sigma$, where c is a scaling constant.

Alternatively, we can consider k dimensional uniform distribution (more better if orientation, orthogonal helps in rotation transformation is considered).

In Gibbs sampler we are moving in unidimensional place.

Here in MVN, k values are generated together and then according we are trying to find the surface.

If we know mean centre of gravity, then if consider BVN with same mean they will coincide. If the variability are also coincided. then BVN actually cover the entire surface. Then samples from that BVN will be somewhere from the actual surface.

The scaling constant is usually considered as 0.5-1.0. This is considered because we always want the central values, so posterior tail probability is not that large. Ultimately goal is acceptance probability should be as much high as possible.

orthogonal transformation is required so that both the mean centre of gravity is considered.

Example:

1. $\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$

$x_1, x_2, \ldots, x_n$ is random samples.

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2\right]$$

$$L \propto \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\left\{\Sigma x_i^2 - 2\mu\Sigma x_i + n\mu^2\right\}\right] \times$$

$$\propto \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\left\{\Sigma x_i^2 - 2\mu\bar{x}.n + n\mu^2\right\}\right]$$

$$\propto \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\left\{\Sigma(x_i-\mu)^2 + \bar{x}-\bar{x})^2\right\}\right]$$

$$\propto \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\left\{\Sigma\{(x_i-\bar{x})^2 + (\bar{x}-\mu)^2 + 0\}\right\}\right]$$

$$\propto \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2\right] \cdot \exp\left[-\frac{1}{2\sigma^2}(\bar{x}-\mu)^2\right]$$

prior $\propto \frac{1}{\sigma}$.

$$p(\mu,\sigma|\underline{x}) \propto \frac{1}{\sigma^{n+1}} \cdot \exp\left[-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2\right] \cdot \exp\left[-\frac{1}{2\sigma^2}n(\bar{x}-\mu)^2\right]$$

$$\begin{cases} p(\mu|\sigma,\underline{x}) \propto \exp\left[-\frac{1}{2\sigma^2}\cdot n(\bar{x}-\mu)^2\right] = \exp\left[-K_1(\bar{x}-\mu)^2\right], & K_1 = \\ p(\sigma|\mu,\underline{x}) \propto \frac{1}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2\right] \cdot \exp\left[-\frac{1}{2\sigma^2}n(\bar{x}-\mu)^2\right] \end{cases}$$

$$\propto \frac{1}{\sigma^{n+1}} \exp\left[-K/\sigma^2\right], \text{ where } K = \Sigma(x_i-\bar{x})^2/2$$

full conditionals.

$p(\sigma|\mu,\underline{x})$: Samples from this can be generated by approximating by gamma/ inverted gamma distribution.

$p(\mu|\sigma,\underline{x})$: Samples from this can generated by approximating by normal distribution. Acceptance Rejection can be another option.

$$\frac{\partial}{\partial\mu}\log p(\mu|\sigma,\underline{x}) = \frac{\partial}{\partial\mu}\left[-K(\bar{x}-\mu)^2\right] = 2K_1(\bar{x}-\mu)$$

$$\Rightarrow \frac{\partial^2 \log p}{\partial\mu^2} = -2K_1. \rightarrow \text{second derivative negative}$$

$\rightarrow$ Distribution concave.

(If +ve, then convex).

If some positive, some negative then this implies bimodality of distribution.

⊛ Exponential (θ) can not be done by Gibbs sampler. as only one parameter. We need more than 2 parameter. (atleast 2).

2. For weibull,

$$f(x) = \frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1} \cdot \exp\left[-\left(\frac{x}{\theta}\right)^{\beta}\right]$$

$$L \propto \frac{\beta^n}{\theta^{n\beta}} \cdot \Pi x_i^{\beta-1} \exp\left[-\Sigma\left(\frac{x_i}{\theta}\right)^{\beta}\right]$$

considering prior $\propto \frac{1}{\theta\beta}$.

Then $p(\beta,\theta|x) \propto \frac{\beta^{n-1}}{\theta^{n\beta+1}} \cdot \Pi x_i^{\beta-1} \exp\left[-\Sigma\left(\frac{x_i}{\theta}\right)^{\beta}\right]$.

$$p(\theta|\beta,x) \propto \frac{1}{\theta^{n\beta+1}} \exp\left[-\Sigma x_i^{\beta}/\theta^{\beta}\right]$$

$$\propto \frac{1}{\theta^{n\beta+1}} \exp\left[-k/\theta^{\beta}\right]$$

$$p(\beta|\theta,x) \propto \frac{\beta^{n-1}}{\theta^{n\beta+1}} \cdot \Pi x_i^{\beta-1} \exp\left[-\Sigma(x_i/\theta)^{\beta}\right]$$

Inverted gamma / Gamma

Difficult to go with Acceptance Rejection Method.

$$\frac{\partial}{\partial\beta}\log p(\beta|\theta,x) = \frac{\partial}{\partial\beta}\left[(n-1)\log\beta - (n\beta+1)\log\theta + (\beta-1)\Sigma\log x_i - \Sigma(x_i/\theta)^{\beta}\right]$$

can be manipulated by HH algorithm.

$$= \frac{n-1}{\beta} - n\log\theta + \Sigma\log x_i - \Sigma\left(\frac{x_i}{\theta}\right)^{\beta}\log\left(\frac{x_i}{\theta}\right)$$

$$\frac{\partial^2}{\partial\beta^2}\log p(\beta|\theta,x) = -\frac{(n-1)}{\beta^2} - 0 + 0 - \Sigma\left(\frac{x_i}{\theta}\right)^{\beta}\left\{\log\left(\frac{x_i}{\theta}\right)\right\}^2$$

$$= -\frac{n-1}{\beta^2} - \Sigma\left(\frac{x_i}{\theta}\right)^{\beta}\left\{\log\left(\frac{x_i}{\theta}\right)\right\}^2$$

< 0

Hence the distribution is also concave.

Hybrid Algorithm:

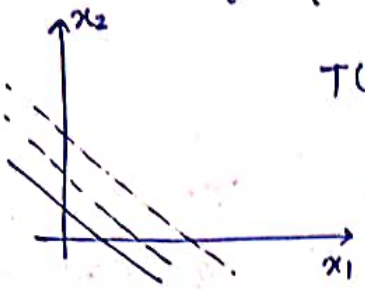
combining different algorithm.

Example: SIR + Metropolis Algorithm + HH Algorithm

Bayesian Sufficiency

If $T(x)$ is a statistics with $\mathcal{f} = \{T(x) : x \in \mathcal{X}\}$ the sufficient partition of $\mathcal{X}$ indicated by T is induced by T is the collection of all sets of the form

$$\mathcal{X}_t = \{x \in \mathcal{X} : T(x) = t\}, \quad t \in \mathcal{f}$$



$T(x) = x_1 + x_2 = t$

Note that if $t_1 \neq t_2$

$\mathcal{X}_{t_1} \cap \mathcal{X}_{t_2} = \phi$

and $\bigcup_{t \in \mathcal{f}} \mathcal{X}_t = \mathcal{X}$.

Then $E_{x|t}[h(x) | T(x) = t]$

$$= \begin{cases} \int_{\mathcal{X}_t} h(x) f_t(x)\, dx. \\ \sum_{x \in \mathcal{X}_t} h(x) f_t(x)\, dx. \end{cases}$$

Sufficiency

Bayesian Definition:

A statistic $T(x)$ is sufficient in Bayesian sense if for all prior $\pi(\theta)$, the posterior distribution of θ is as follows $\pi(\theta|x) \propto \pi(\theta|T(x)=t)$.

Note: In classical sufficiency always implies Bayesian Sufficiency. Thus the posterior density is a function of sufficient statistic.

Theorem: If T is sufficient in classical sense, then it is sufficient in Bayesian also.

The posterior distribution is defined by

$$\pi(\theta|\underline{x}) = \frac{l(\underline{x}|\theta) \cdot \pi(\theta)}{\int_{\theta} l(\underline{x}|\theta) \cdot \pi(\theta) d\theta}.$$

By Neymann-Fisher factorization theorem,

$$l(\underline{x}|\theta) = h(x) \cdot g(T(x), \theta).$$

$$\text{Then } \pi(\theta|\underline{x}) = \frac{h(\underline{x})\, g(T(\underline{x}), \theta) \cdot \pi(\theta)}{h(\underline{x}) \int g(T(\underline{x}), \theta) \pi(\theta) d\theta}$$

$$= \frac{g(T(\underline{x}); \theta)\, \pi(\theta)}{\int g(T(\underline{x}), \theta) \pi(\theta) d\theta}.$$

Converse part:

Assuming Bayesian sufficiency we have

$$\frac{L(\underline{x}|\theta)\pi(\theta)}{f(x)} = \pi(\theta|\underline{x})$$
$$= \pi(\theta|T(\underline{x}))$$
$$= \frac{f(T(x)|\theta).\pi(\theta)}{f(T(x))}.$$

Thus we have,

$$f(\underline{x}|\theta) = f(x) \cdot \frac{f(T(x)|\theta)}{f_T(T(x))}.$$

which can written in the forem of $h(x).g(T(x),\theta)$.

$$= \left\{\frac{f(x)}{f_T(T(x))}\right\} \cdot g\, f(T(\underline{x})|\theta)$$

↓ Independent of θ ; ↳ function of T and θ.

1. Exponential (θ)

$$L(\theta) = \theta^n e^{-\theta \Sigma x_i}$$
$$= \theta^n e^{-\theta T(x)}$$

Where $T(x) = \Sigma x_i \sim$ Gamma (n, θ).

Prior $\rightarrow \theta \sim$ Gamma (a,b).

$$\pi(\theta|\underline{x}) \propto \theta^n e^{-\theta T(x)} \cdot \frac{a^b}{\Gamma b} \theta^{b-1} e^{-\theta a}.$$
$$\propto \theta^{n+b-1} \cdot e^{-\theta(T(x)+a)}.$$
$$\propto \text{Gamma}\,(n+b,\ T(x)+a). \rightarrow \text{conjugate prior.}$$

Another way:

$$\pi(\theta|T(x)) = \frac{f(T(x)|\theta)\cdot\pi(\theta)}{f(T(x))}.$$
$$\propto f(T(x)|\theta)\cdot\pi(\theta).$$
$$\propto \frac{\theta^n}{\Gamma n} T^{n-1} e^{-T\theta} \frac{a^b}{\Gamma b} \theta^{b-1} e^{-\theta a}$$
$$\propto \theta^{n+b-1} e^{-\theta(T+a)}$$
$$\propto \text{Gamma}\,(n+b,\ T(x)+a).$$

Hence $\pi(\theta|\underline{x}) = \pi(\theta|T(x))$.

Exponential Family

$$f(x) = h(x) \exp\{R(\theta)T(x) - A(\theta)\}.$$

The likelihood function is

$$l(\theta) = \prod_{i=1}^{n} h(x_i) \exp\{R(\theta)\sum_{i=1}^{n} T(x_i) - n A(\theta)\}.$$

In order to construct conjugate prior, we have

$$\pi(\theta) \propto \exp\{R(\theta)\gamma - \gamma_0 A(\theta)\}.$$

$$\pi(\theta|\underline{x}) \propto \exp\{R(\theta)(\sum_{i=1}^{n} T(x_i) + \gamma) - (n+\gamma_0)A(\theta)\}$$

Then For exponential (θ), we have

$$\pi(\theta) \propto \theta^{a} e^{-\theta\gamma}$$

$$\pi(\theta) \propto \theta^{n+a} e^{-\theta(\gamma+\Sigma x_i)}.$$

Emperical Bayes

Estimate prior $\hat{\lambda}$ based on data.

① parametric EB

② Non-parametrical EB.

Marginal distribution of X.

$$h(x,\theta) = f(x|\theta)\,\pi(\theta)$$

$$m(x) = \int_{\Theta} f(x|\theta)\,\pi(\theta)\,d\theta.$$

Example: $X \sim N(\theta, \sigma_f^2)$

$$\theta \sim N(\mu_\pi, \sigma_\pi^2)$$

$$\theta|x \sim N\left(\frac{\sigma_f^2\theta}{\sigma_f^2+\sigma_\pi^2} + \frac{\sigma_\pi^2 x}{\sigma_f^2+\sigma_\pi^2}, \frac{\sigma_f^2\sigma_\pi^2}{\sigma_f^2+\sigma_\pi^2}\right).$$

$$m(x) \sim N(\mu_\pi, \sigma_\pi^2+\sigma_f^2)$$

If $m(x|\pi_1) > m(x|\pi_2)$.

Hyperparameters of π_1 are more plaussible than π_2.

Here in classical inference θ is fixed.

Maximum likelihood 2 approach:

$$\Gamma = \{\pi : \pi(\theta) = g(\theta|\lambda),\ \lambda \in \Lambda\}.$$

1. HL-II method
2. Moments Method
3. Distance method.

Definition:

Suppose Γ is a class of prior under consideration and that $\hat{\pi} \in \Gamma$ satisfies

$$m(x|\hat{\pi}) = \sup_{\pi \in \Gamma} m(x|\pi).$$

Then $\hat{\pi}$ will be called type-II maximum likelihood prior.

Then $\sup_{\pi \in \Gamma} m(x|\pi) = \sup_{\lambda \in \Lambda} m(x|g(\theta)|\lambda))$.

let $m_0 \sim N(\mu_\pi, \sigma_f^2 + \sigma_\pi^2)$.

$$m(\underset{\sim}{x}|\pi) = \prod_{i=1}^{p} m_0(x_i|\pi_0).$$

$$= \frac{1}{[2\pi(\sigma_\pi^2+\sigma_f^2)]^{p/2}} \exp\left\{-\sum_{i=1}^{p} \frac{(x_i-\mu_\pi)^2}{2(\sigma_\pi^2+\sigma_f^2)}\right\}.$$

$$= \frac{1}{[2\pi(\sigma_\pi^2+\sigma_f^2)]^{p/2}} \exp\left\{\frac{-ps^2}{2(\sigma_\pi^2+\sigma_f^2)}\right\} \exp\left\{\frac{-p(\bar{x}-\mu_\pi)^2}{2(\sigma_\pi^2+\sigma_f^2)}\right\}.$$

For $\mu_x = \bar{x}$ the likelihood $m(\underset{\sim}{x}|\pi)$ acheives maximum value.

$$\psi(\sigma_\pi^2) = \frac{1}{[2\pi(\sigma_\pi^2+\sigma_f^2)]^{p/2}} e^{-ps^2/2(\sigma_\pi^2+\sigma_f^2)}$$

$$\frac{\partial}{\partial \sigma_\pi^2}\psi(\sigma_\pi^2) = \frac{-p/2}{(\sigma_\pi^2+\sigma_f^2)} + \frac{ps^2}{2(\sigma_\pi^2+\sigma_f^2)} = 0$$

$$\sigma_\pi^2 = s^2 - \sigma_f^2, \quad s^2 \geq \sigma_f^2$$

$$\sigma_\pi^2 = 0, \quad s^2 < \sigma_f^2$$

Then HLE's are

$$\hat{\mu}_\pi = \bar{x}$$

$$\sigma_\pi^2 = \max(0, s^2 - \sigma_f^2)$$

$\Gamma = \{\pi : \pi(\theta) = g(\theta)\}, \quad \theta \in \Lambda\}$

$\lambda \sim g(\theta) = \text{Gamma}(\alpha, \beta)$

$E(\lambda) = ? = \bar{\lambda}$

$\text{Var}(\lambda) = \text{Var}(\lambda)$

where $\bar{\lambda} = \alpha/\beta \quad \frac{\alpha}{\beta^2} = \text{Var}(\lambda)$

Consider, $\mu_m = E_m(X) = \int_{\mathcal{X}} x\, m(x)\, dx.$

$$= \int_{\mathcal{X}} x \left\{ \int_{\Theta} f(x|\theta)\, \pi(\theta)\, d\theta \right\} dx$$

$$= \int_{\Theta} \pi(\theta) \int_{\mathcal{X}} x f(x|\theta)\, dx\, d\theta$$

$$= \int_{\Theta} \pi(\theta)\, \mu_f(\theta)\, d\theta$$

$$= E_{\pi}[\mu_f(\theta)]$$

Likewise,

$$\sigma_m^2 = E_m (x-\mu_m)^2 = \int_{\mathcal{X}} (x-\mu_m)^2 m(x)\, dx$$

$$= \int_{\mathcal{X}} (x-\mu_m)^2 \int_{\Theta} f(x|\theta)\, \pi(\theta)\, d\theta\, dx$$

$$= \int_{\Theta} \pi(\theta) \left\{ \int_{\mathcal{X}} (x-\mu_m)^2 f(x|\theta)\, dx \right\} d\theta.$$

$$= \int_{\Theta} E_f (x-\mu_m)^2 \pi(\theta)\, d\theta$$

$$= E_{\pi}[E_f (x-\mu_m)^2]$$

Consider, $E_\pi E_f [(x-\mu_f(\theta)) + (\mu_f(\theta) - \mu_m)]^2$

$$\Rightarrow \sigma_m^2 = E_\pi \{\sigma_f^2(\theta)\} + E_\pi \{\mu_f(\theta) - \mu_m\}^2$$

(1) If $\mu_f(\theta) = 0$, then $\mu_m = \mu_\pi = E_\pi(\theta)$, prior mean

(2) If $\sigma_f^2(\theta) = \sigma_f^2$ (free from θ)

$$\sigma_m^2 = \sigma_f^2 + \sigma_\pi^2$$

$x \sim N(0,1)$

$\theta \sim N(\mu_\pi, \sigma_\pi^2)$

$x|\pi \sim N(1, 3)$

subjective Approach.

$\mu_\pi = ?$ $\sigma_\pi^2 = ?$

$\mu_\pi = \mu_m = 1.$

$\sigma_\pi^2 = \sigma_m^2 - \sigma_f^2 = 3-1 = 2$

$\theta \sim N(1,2)$

(3) Distance Approach

$\hat{m}(x) = \frac{1}{p}$[The number of x_i equal to x]

$\hat{m}(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta.$

$d(\hat{m}, \hat{m}_\pi) = E_{\hat{m}}\left[\log \frac{\hat{m}(x)}{m_{\hat{\pi}}(x)}\right]$

$= E_{\hat{m}}(\log \hat{m}(x)) - E_{\hat{m}}(\log m_\pi(x))$

free from π

* $\hat{\pi}$ can be obtained by minimizing with respect to hyperparameter.
* Since $E_{\hat{m}}(\log \hat{m}(x))$ is free from π, minimisation of $d(\hat{m}, \hat{m}_\pi)$ is equivalent to maximization of $E_{\hat{m}}(\log m_\pi(x))$.
* If $\pi(\theta_i) = p_i$ with $0 \le p_i \le 1$, $\sum_{i=1}^{k} p_i = 1$

$$= \sum_{j=1}^{n} \frac{1}{n} \log\left(\sum_{i=1}^{k} f(x_i|\theta) p_i\right).$$

The maximization of the last expression over p_i is straight forward LPP.

Non-parametric Bayes

Let $x_i \sim P(\theta_i)$, $i = 1, 2, \ldots, P$.

$\theta_i \sim \pi(\theta)$; $i = 1, 2, \ldots P$

This density can be estimated as

$$\hat{m}(j) = \frac{\text{no. of } x_i\text{'s equal to } j}{P}.$$

We need to obtain Bayes estimate as →

$$\delta_{\pi_0}(x_p) = E(\theta_p | x_p) = \int_{\theta_p} \theta_p \, \pi(\theta_p | x_p) \, d\theta_p$$

($\pi(\theta_p | x_p)$: posterior mean.)

$$= \frac{\int_{\theta_p} \theta_p f(x_p | \theta_p) \pi_0(\theta_p) \, d\theta_p}{m(x_p)}$$

$$= \frac{\int \frac{\theta_p^{x_p+1}}{x_p!} \pi_0(\theta_p) \, d\theta_p}{m(x_p)}.$$

$$= \frac{(x_p+1) \int \frac{\theta_p^{x_p+1} e^{-\theta_p}}{x_{p+1}!} \pi_0(\theta_p) \, d\theta_p}{m(x_p)}$$

$$= \frac{(x_p+1) \int f(x_{p+1} | \theta_p) \pi_0(\theta_p) \, d\theta_p}{m(x_p)}.$$

$$= \frac{(x_p+1) \cdot m(x_p+1)}{m(x_p)}.$$

Bayes Estimate will be →

$$(x_p+1) \cdot \frac{\text{no. of } x_i\text{'s equal to } x_{p+1}}{\text{no. of } x_i\text{'s equal to } x_p}$$

Example: $n = 10$, $x_1, x_2, \ldots, x_n \sim$ Poisson (θ).

The observations are $\{0, 1, 5, 2, 4, 6, 9, 6, 4, 2\}$.

i	0	1	2	3	4	5	6	7	8	9
$m(i)$	1/10	1/10	2/10	0/10	2/10	1/10	2/10	0/10	0/10	1/10

Bayes Estimate of θ, when $x_{obs} = 0$ is

$$\hat{\theta}_0 = (x_n+1) \cdot \frac{m_{x_n+1}}{m_{(x_n)}}$$

$$= (0+1) \cdot \frac{m(1)}{m(0)} = 1 \cdot \frac{1/10}{1/10} = 1.$$

Hierarchical Bayes:

Hierarchical Bayes model is a Bayesian statistical model $(f(x|\theta) \cdot \pi(\theta))$. where prior distribution $\pi(\theta)$ is decomposed in conditional distributions.

$\pi_1(\theta|\theta_1).\pi_2(\theta_1|\theta_2)\ldots\pi_n(\theta_{n-1}|\theta_n)$

and a marginal distribution $\pi_{n+1}(\theta_n)$ such that

$$\pi(\theta) = \int \pi_1(\theta|\theta_1).\pi_2(\theta_1|\theta_2)\ldots\pi_n(\theta_{n-1}|\theta_n)\, d\theta_1 d\theta_2 \ldots d\theta_n.$$
$$\Theta_1 \cdots \Theta_n .$$

where α_i's are called hyper-parameters of level i, $1 \le i \le n$.

Two stage is good, more than two stage is a waste

first stage: → hyperparameter.

$\pi_1(\theta|\lambda)$: $\lambda \in \Lambda$

Second stage:

$\lambda \sim \pi_2(\lambda)$

usually non-informative prior but should be proper prior.

Conditional posterior:

The posterior of θ is $p(\theta|x) = \int \pi_1(\theta|\quad)$

$\theta|\theta_1 \sim \pi_1(\theta|\theta_1)$, where θ_1 hyperparameter.

$\theta_1 \sim \pi_2(\theta_1)$.

So, $\pi(\theta|\theta_1, \underline{x}) = \dfrac{f(x|\theta).\pi_1(\theta|\theta_1).}{m(x|\theta_1).}$

$\Rightarrow m_1(x|\theta_1) = \int_{\Theta} f(x|\theta)\pi_1(\theta|\theta_1)\, d\theta$

$\Rightarrow \pi(\theta_1|x) = \dfrac{m(x|\theta_1)\pi_2(\theta)}{m(x)}$

$\therefore m(x) = \int_{\Theta} m_1(x|\theta_1).\pi_2(\theta)\, d\theta_1$

Posterior moments:

$$E^{\pi}(h(\theta)|\underline{x}) = E^{\pi}(\theta_1|\underline{x}).\left[E^{\pi_1}(h(\theta)|\theta_1,x)\right]$$

where $E^{\pi_1}(h(\theta)|\theta_1,x) = \int_{\Theta} h(\theta)\pi(\theta|\theta_1,\underline{x})\, d\theta$.

Example: (1) $x \sim N(\mu,1)$, $\mu \sim$ Normal $(\mu_\pi, 1)$, $\mu_\pi \sim N(0,1)$

(2) $x \sim N(\mu_0, \sigma^2)$

$\sigma^2 \sim$ Gamma (a,b)

$(a,b) \sim$ Gamma $(1,1)$

(3) $x \sim$ Binomial (n,p)

$p|m \sim$ Beta (m,m)

$m \sim$ Uniform $(0,1)$

or $m \sim$ gamma (1)

Bayes Robustness

1. Model.
2. Prior.
3. Loss function.

$X \sim Poi(\theta)$

Median = 2

$\theta_4 = 4$

Priors: $\pi_1 : \theta \sim$ Exponential (a), $a = \log 2/2$

$\pi_2 : \log(\theta) \sim N\left(\log(2), (\log^2/z_{.05})^2\right)$.

$\pi_3 : \log(\theta) \sim$ Cauchy $(\log 2, \log 2)$

π	0	1	...	50.
π_1	0.749			
π_2	0.950			
π_3	0.761.			

1. To ensure that as many reasonable as possible are included.
2. To try to eliminate unreasonable priors.
3. To ensure ~~the~~ prior family Γ does not require prior information which is difficult to illustrate
4. To be able to compute measures of robustness without much difficulty.

① Conjugate priors $\Gamma_C = \{ N(w, \tau^2) ; \mu_1 \le \mu \le \mu_2 ; \tau_1^2 \le \tau^2 \le \tau_2^2 \}$

② Neighbourhood class

$\Gamma_N = \{ \pi$ which are in the neighbourhood of $\pi_0 \}$.

③ ϵ-contamination class

$\Gamma_\epsilon = \{ \pi : \pi = (1-\epsilon)\pi + \epsilon q : q \in Q \}$.

④ Density Ratio class.

$\Gamma_{DR} = \{ \pi : L(\theta) \le \alpha \pi(\theta) \le U(\theta)$, for some $\alpha > 0 \}$.

$L = 1$, $U = c$, then

$\Gamma_{DR} = \{ \pi : c^{-1} \le \frac{\pi(\theta)}{\pi(\theta')} \le c$, for all $\theta, \theta' \}$.

Posterior Robustness:

(1) Global Measure:

3 functions were suggested.

(i) linear functionals of the prior:

$$\rho(\pi) = \int h(\theta)\, \pi(\theta|x)\, d\theta$$

if h = likelihood.

$$m(x) = \int_{\Theta} l(\theta|x)\, \pi(\theta|x)\, d\theta.$$

(ii) Ratio of linear functionals of prior

$$\rho(\pi) = \frac{1}{m(x)} \int h(\theta)\, l(\theta)\, \pi(\theta|x)\, d\theta$$

if $h(\theta)=\theta$, $\rho(\pi) = \frac{1}{m(x)} \int \theta\, l(\theta)\, \pi(\theta|x)\, d\theta$

(iii) Ratio of non-linear functionals

$$\rho(\pi) = \frac{1}{m(x)} \int h(\theta, \phi(\pi))\, l(\theta)\, \pi(\theta|x)\, d\theta.$$

if $h(\theta, \phi(\pi)) = (\theta - \mu(\pi))^2$, where $\mu(\pi)$ is the posterior mean

Example: Suppose $x \sim N(\theta, \sigma^2)$ → known.

$\Gamma = N(0, \tau^2)$, $\tau^2 > 0$.

Then variation in posterior mean will be

$$(\inf E(\theta|\bar{x}),\ \sup(\theta|\bar{x})).$$

Here $E(\theta|\bar{x}) = \frac{\tau^2}{\tau^2+\sigma^2} \cdot \bar{x}$

$$\lim_{\tau^2 \to 0} E(\theta|x) = 0.$$

$$\lim_{\tau^2 \to \infty} E(\theta|\bar{x}) = \bar{x}.$$

Hence the interval is $(0, \bar{x})$ or $(\bar{x}, 0)$.

Then $R_\pi(x) = \dfrac{(\rho^{\pi}_{(x)} - \rho^{0}_{(x)})^2}{\text{var}^{\pi}(x)}$.

where $\rho^{\pi}_{(x)}$: posterior mean

$\rho^{0}_{(x)}$: posterior mean with respect to nbd class.

Example: $x \sim N(\theta, 1)$

$\pi_0 \sim N(0, 2)$.

$\Gamma = N(0, \tau^2)$, $1 \le \tau^2 \le 10$.

Then $\pi(\theta) = \theta$.

$$\theta|x \sim N\left(\frac{\tau^2 x}{\tau^2+1}, \frac{\tau^2}{\tau^2+1}\right).$$

$$R_\pi(x) = [(x^2-2)^2 \cdot \text{?}] \left\{\frac{\tau^2}{\tau^2+1} - \frac{2}{3}\right\} \cdot x.$$

$$R_\pi(x) = \frac{(\tau^2-2)^2 x^2}{z^2(z^2+1)}.$$

$\sup R_\pi(x) = 6.4\, x^2/99$.

Thus robustness is accepted when x lies between $(0, 4)$.

Local measure of sensitivity.

$$S(\pi, V : x) = \lim_{\epsilon \to 0} \frac{d(\pi^x, V_\epsilon^x)}{d(\pi, V_\epsilon)}.$$

$$V_\epsilon = (1-\epsilon)\pi + \epsilon . V$$

$$d(\pi, V) = \sup_{A \in B} |\pi(A) - V(A)|$$