Paradoxes in Classical Statistics

Example 1. To estimate μ in $N(\mu, \sigma^2)$, toss a fair coin. Have a smaple size $n = 2$ if it is a head and take $n = 1000$ if it is a tail. An unbiased estimate of μ is $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ with variance $\frac{1}{2}\left\{ \sigma^2/2 + \sigma^2/1000 \right\} \approx \sigma^2/4$.

Suppose it was a tail. Would you believe $\sigma^2/4$ is a measure of accurary of the estimate ?

Example 2 Let X_1, X_2 be i.i.d. $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. Let $\bar{X} \pm C$ be a 95% confidence interval, $C > 0$ being suitably chosen. Suppose $X_1 = 2$ and $X_2 = 1$. Then we know for sure $\theta = (X_1 + X_2)/2$ and hence $\theta \in (\bar{X} - C, \bar{X} + C)$

Should we still claim we have only 95% Confidence that the confidence interval cover θ ?

Bayesian Inference

To make inference about θ is to learn about the unknown θ from data X, i.e, based on the data, explore which values of θ are probable, what might be estimates of different components of θ and the extent of uncertainty associated with such estimates.

In addition to having a model $f(x|\theta)$ and a likelihood function, the Bayesian needs a distribution to ~~having~~ ~~a~~ for θ.

The distribution is called a prior distribution or simply a prior because it quatitifies her uncertainty about θ prior to seeing data.

Now the conditional probability density of θ given $X = x$ by Bayes formula

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int_{\Theta} \pi(\theta') f(x|\theta') d\theta'} \quad \ldots\ldots (1)$$

where $\pi(\theta)$ is the prior density function and $f(x/\theta)$ is the density of X, which is the conditional density of X given θ.

The symbol θ now represents both a random variable and its value. When θ is discrete, the integral in the denominator of (1) is replaced by a sum.

The Conditional density $\pi(\theta/x)$ of θ given $X = x$ is called the posterior density, a quantification of our uncertainty about θ in the light of data. The transition from $\pi(\theta)$ to $\pi(\theta/x)$ is what we learnt from the data.

A Bayesian can simply report her posterior distribution, or she could report the posterior mean

$$E(\theta|x) = \int_{-\infty}^{\infty} \theta \, \pi(\theta|x) \, d\theta$$

and the posterior variance

$$Var(\theta|x) = \int_{-\infty}^{\infty} (\theta - E(\theta|x))^2 \, \pi(\theta|x) \, d\theta .$$

if she want to estimate θ,

or in the case of testing she would report the posterior odds of the relevant hypotheses.

Example 1 Consider the problem of inference about μ for normally distributed data $N(\mu, \sigma^2)$. The data consist of i.i.d. observation $X_1, X_2, \cdots, X_n$ from this distribution. We assume that σ^2 is known.

A mathematically convenient and reasonably flexible prior distribution for μ is a normal distribution with suitable prior mean η and variance τ^2. The prior variance, τ^2 is a measure of strength of our belief in the prior mean, in the sense that the larger the value of τ^2, the less sure we are about our prior guess about η. Jeffreys has suggested we can calibrate τ^2 by Comparing with σ^2.

For example setting $\tau^2 = \sigma^2/m$ would amount to saying information about η is about as strong as the information in m observations in data.

To illustrate, we take, $n = 10$, $m = 1$, the posterior distribution is normal with posterior mean $E(\mu \mid X) = \left(\frac{1}{\tau^2}\eta + \frac{n}{\sigma^2}\bar{x}\right) \Big/ \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)$

$$= (\eta + 10\,\bar{x})/11 .$$

and posterior variance

$$\left(\sigma^2/n \; \tau^2\right) \Big/ \left(\sigma^2/n + \tau^2\right) = \frac{\sigma^2}{11}$$

i.e., in the light of data, μ shifts from prior guess η towards a weighted average of the prior guess about μ and $\bar{x}$, while the variability reduces from σ^2 to $\sigma^2/11$.

If the prior information is small, implying large τ^2 or these are lots of data i.e., n is large, the posterior mean is close to the MLE $\bar{x}$.

Example 2 Consider an urn with Np red and $N(1-p)$ black balls, p is unknown but N is a known large number. Balls are drawn at random one by one with replacement, selection is stopped after n draws. For $i = 1, 2, \cdots, n$ let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th ball drawn is red} \\ 0 & \text{otherwise.} \end{cases}$$

Then X_i's are i.i.d. $B(1, p)$,

Let p have a prior distribution $\pi(p)$.

We will consider a family of priors for p that simplifies the calculation of posterior.

$$\text{Let } \pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad \cdots\cdots (2)$$

$$0 \le p \le 1 \; ; \; \alpha > 0, \; \beta > 0$$

The prior mean and variance are

$$\frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \text{ respectively.}$$

The posterior density is

$$\pi(p \mid X = x) = c(x)\, p^{\alpha+r-1}(1-p)^{\beta+(n-r)-1}$$

where $r = \sum_{i=1}^{n} x_i$ = number of red balls in n draw and $(c(x))^{-1}$ is the denominator in Bayes formula.

Now the posterior is also a Beta density with ~~$\alpha+\beta$~~ $\alpha+r$ in place of α and $\beta+(n-r)$ in place of β in (2) and

$$C(x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+r)\,\Gamma(\beta+n-r)}$$

The posterior mean and variance are

$$E(p|x) = (\alpha+r)/(\alpha+\beta+n)$$

$$Var(p|x) = \frac{(\alpha+r)(\beta+n-r)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

Again if n is large then the posterior mean is approximately equal to the MLE, $\hat{p} = r/n$ and the posterior ~~mean~~ variance is ~~[illegible]~~ quite small,

We can interpret this as an illustration of a fact mentioned before when we have lots of data, the data tend to wash away the influence of the prior.

The posterior mean can be rewritten as a weighted average of the prior mean and MLE

$$\frac{(\alpha+\beta)}{(\alpha+\beta+n)}\;\frac{\alpha}{(\alpha+\beta)} + \frac{n}{(\alpha+\beta+n)}\;\frac{r}{n}$$

the relative importance of the prior and the data being measured by $(\alpha+\beta)$ and n.

Suppose we want to predict the probability of getting a red ball in a new $(n+1)$-st draw given the above data. It would be natural to use $E(p|x)$.

Some commonly used priors and the corresponding value of $E(p/x_1, x_2, \cdot \cdot, x_n)$

If $\alpha = \beta = 1$, the uniform prior, with posterior mean $\left(\sum_{i=1}^{n} x_i + 1\right)/(n+2)$.

If $\alpha = \beta = \frac{1}{2}$, we have the Jeffreys prior with posterior mean $\left(\sum_{1}^{n} x_i + \frac{1}{2}\right)/(n+1)$

Example 3 suppose we think of the problem as a representation of production of defective and non-defective items in a factory producing switches, we would take red to mean defective and black to mean a good switch.

In this context, there would be some prior information available from the engineers. They may be able to pinpoint the likely value of p, which may be set equal to the prior mean $\alpha/(\alpha+\beta)$. If one has knowledge of prior variability also, one would have two equations from which to determine α and β.

In this ~~particular context, the Jeffreys prior~~ with ~~a lot~~ It can be shown that the uniform, Jeffreys prior produce a posterior mean that is very closed to the MLE even for small n.

Also they make better sense than the MLE in the extreme case when $\hat{p}=0$. In most contexts the estimate $\hat{p}=0$ is absurd, the objective Bayes estimates move it a little towards $p=\frac{1}{2}$. Such a movement is called a shrinkage.

Example 4 Let $X_1, \ldots X_n$ be i.i.d $N(\mu, \sigma^2)$ and σ^2 is known. We want to test $H_0: \mu \leq \mu_0$ vs $H_1: \mu > \mu_0$.

Let $\pi(\mu)$ be the prior. First calculate the posterior density $\pi(\mu|x)$. Then

$$P\{H_0|x\} = \int_{-\infty}^{\mu_0} \pi(\mu|x)\, d\mu$$

and

$$\int_{\mu_0}^{\infty} \pi(\mu|x)\, d\mu = 1 - P\{H_0|x\} = P(H_1|x)$$

One may simply report these numbers or choose one of the two hypotheses if one of the two probabilities is substantially bigger.

Improper Priors

For point and interval estimates and to some extent in testing improper priors are used.

An improper prior density $\pi(\theta)$ is non-negative for all θ but $\int_{\Theta} \pi(\theta)\, d\theta = \infty$.

Such an improper prior can be used in the Bayes formula for calculating the posterior, provided the denominator is finite for all x (or all but a set of x with zero probability) and for all θ,

i.e.,
$$\int_{\Theta} \pi(\theta) f(x/\theta)\, d\theta < \infty .$$

Then the posterior density $\pi(\theta / X = x)$ is a proper density function and can be used at least in inference problems or the posterior ~~density~~ ~~function~~ decision problem where we define and minimize
$$\psi(x,a) = E\left(L(\theta,a)/x\right) = \int_{\Theta} L(\theta,a)\, \pi(\theta/x)\, d\theta$$

which is the posterior risk.

However, for improper priors usually $R(\pi, \delta)$ is not used.

where
$$R(\pi,\delta) = \int_{\Theta} R(\theta,\delta)\, \pi(\theta)\, d\theta$$

and
$$R(\theta,\delta) = E_{\theta}\left(L(\theta, \delta(x))\right)$$

The most common improper priors are

$$\pi_1(\mu) = C, \qquad -\infty < \mu < \infty,$$

$$\pi_2(\sigma) = \frac{1}{\sigma}, \qquad 0 < \sigma < \infty$$

for location and scale parameters μ and σ resp.

Both the improper priors may be interpreted as a sort of limit of the proper priors.

$$\pi_{1,L}(\mu) = \begin{cases} \frac{1}{2L} & \text{if } -L < \mu < L \\ 0 & \text{otherwise} \end{cases}$$

and

$$\pi_{2;L}(\sigma) = \begin{cases} A/\sigma & \text{if } 0 < \frac{1}{L} < \sigma < L \\ 0 & \text{otherwise}; \end{cases}$$

where $A = \frac{1}{(2\log L)}$, in the sense that the posteriors for π_1 and π_2 may be obtained by making $L \to \infty$ in $\pi_{i,L}(\theta \mid x)$

Point Estimates

For a real value θ, standard Bayes estimates are the posterior mean or the posterior median. The posterior man is the Bayes estimate corresponding with squared error loss and the posterior median is Bayes estimate for absolute deviation loss.

Along with the posterior mean one reports the posterior variance of θ. If one choose to work with the posterior median, one should report atleast first and third posterior quartiles.

If ~~p~~ the posterior is unimodal then the posterior mode is another choice. It is similar to the MLE of classical statistics. Indeed if the prior is uniform, both are identical. Along with the posterior mode one can report a suitable highest posterior density credible interval as a measure of posterior variability.

If the paramater is a vector, common choices for reporting are the posterior mean vector and the posterior dispersion matrix.

Testing

We want to test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. If Θ_0 and Θ_1 are of the same dimension as for one-sided null and alternative hypotheses, it is convenient and easy to choose a prior distribution that assigns positive prior probability to Θ_0 and Θ_1. One then calculates the posterior probabilities $P\{\Theta_i | x\}$ as well as the posterior odds ratio (or simply posterior odds), namely,

$$P\{\Theta_0 | x\} / P\{\Theta_1 | x\}$$

One would then find a threshold to decide what constitutes evidence against H_0. The Bayes rule for 0-1 loss is to choose the hypothesis with higher posterior probability.

Let π_0 and $1-\pi_0$ be the prior probabilites of Θ_0 and Θ_1. Let $g_i(\theta)$ be the prior p.d.f. of θ under Θ_i, so that $\int_{\Theta_i} g_i(\theta)\, d\theta = 1$

The prior is then.

$$\pi(\theta) = \pi_0 g_0(\theta) \; I\{\theta \in \Theta_0\} + (1-\pi_0) g_1(\theta) \, I\{\theta \in \Theta_1\}.$$

We do not require any longer that Θ_0 and Θ_1 are the same dimension. So sharp null hypotheses are also covered. We can now proceed as before and report posterior probabilities or posterior odds.

③.

To compute these posterior quantities, note that the marginal density of X under the prior π can be expressed as

$$m_\pi(x) = \int_{\Theta} f(x|\theta)\,\pi(\theta)\,d\theta$$

$$= \pi_0 \int_{\Theta_0} f(x|\theta)\, g_0(\theta)\, d\theta + (1-\pi_0)\int_{\Theta_1} f(x|\theta) g_1(\theta) d\theta$$

and hence the posterior density of θ given the data $X = x$ as

$$\pi(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{m_\pi(x)}$$

$$\therefore \ \pi(\theta|x) = \begin{cases} \pi_0 f(x|\theta)\, g_0(\theta) / m_\pi(x) & \text{if } \theta \in \Theta_0 \\ (1-\pi_0)\, f(x|\theta)\, g_1(\theta) / m_\pi(x) & \text{if } \theta \in \Theta_1 . \end{cases}$$

It follows that

$$P^\pi(H_0|x) = P^\pi(\Theta_0|x) = \frac{\pi_0}{m_\pi(x)} \int_{\Theta_0} f(x|\theta)\, g_0(\theta)\, d\theta$$

and $$P^\pi(H_1|x) = P^\pi(\Theta_1|x) = \frac{(1-\pi_0)}{m_\pi(x)} \int_{\Theta_1} f(x|\theta)\, g_1(\theta)\, d\theta$$

Now the Bayes. factor, of Ho relative to H_1 is defined as

$$BF_{01} = \frac{\int_{\Theta_0} f(x|\theta)\, g_0(\theta)\, d\theta}{\int_{\Theta_1} f(x|\theta)\, g_1(\theta)\, d\theta}$$

which ~~is~~ does not depends on π_0.

Clearly $BF_{10} = \dfrac{1}{BF_{01}}$

The posterior odds ratio of H_0 relative to H_1 is

$$\left(\frac{\pi_0}{1-\pi_0}\right) BF_{01},$$

which is same as BF_{01} if $\pi_0 = \frac{1}{2}$.

The smaller the value of BF_{01}, the stronger the evidence against H_0.

Example Consider a blood test conducted for determining the sugar level of a person with diabetes two hours after he had his breakfast. It is of interest to see if his medication has controlled his blood sugar level. Assume that the test result X is $N(\theta, 100)$, where θ is the true level. In the appropriate population (diabetic but under this treatment), θ is distributed according to a $N(100, 900)$. Then marginally X is $N(100, 1000)$, and the posterior distribution of θ given $X = x$ is normal with mean $\frac{900}{1000} x + \frac{100}{1000} 1000$

$= 0.9x + 10$ and variance $= \frac{100 \times 900}{1000} = 90$

Suppose we want to test $H_0 : \theta \leq 130$ vs $H_1 : \theta > 130$. If the blood test shows a sugar level of 130, what can be concluded?

Note that, given this test result, the true-mean blood sugar level, θ, may be assumed to be $N(127, 90)$. Consequently we obtain,

$$P(\theta \leq 130 / x = 130) = \Phi\left(\frac{130-127}{\sqrt{90}}\right) = 0.624.$$

Therefore Posterior odds ratio $= \frac{0.624}{1-0.624} = 1.66$.

Bacuse $\pi_0 = P^{\pi}(\theta \leq 130) = \Phi(1) = 0.8413$,

the prior odds ratio is $\frac{\Phi(1)}{1-\Phi(1)} = 5.3$ and thus

the Bayes factor turns out to be $\frac{1.66}{5.3} = 0.313$.

It can be noted here that in one-sided testing situations when a continuous prior π can be specified readily for the entire parameter space, there is no need to express it in form of

$$\pi(\theta) = \pi_0\, g_0(\theta)\, I\{\theta \in \Theta_0\} + (1-\pi_0)\, g_1(\theta)\, I\{\theta \in \Theta_1\}.$$

However, the problem of testing a point null hypothesis turns out to be quite different.

Testing of a Point Null Hypothesis

The problem is to test

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

Consider the following example.

In a statistical quality control situation, θ is the size of a unit and acceptable units are with $\theta \in (\theta_0 - \delta, \theta_0 + \delta)$. Then one would like to test $H_{0\delta} : |\theta - \theta_0| \leq \delta$.

In this problem the length of the interval, 2δ, can be explicitly specified. On the other hand, this is not in the case ~~of~~ the following example.

Suppose we want to test the hypothesis,

H_0 : Vitamin C has no effect on the commom cold.

Clearly this is not meant to be thought of as an exact point null; surely vitamin C has 'some' effect, though perhaps a very minnuscule effect. Thus, in reality, this is still the case of an interval null hypothesis, with a very small unspecified interval. However, it would be better represented as a point null hypothesis.

Now consider the hypothesis such as

H_0 : Astology cannot predict the future.

Can perhaps be represented as an exact point null.

Conceptually testing a point null is not a different problem, but there are complications. First of all, it is not possible to use a continuous prior density because any such prior will necessarily assign prior probability zero to the null hypothesis. Consequently, the posterior probability of the null hypothesis will also be zero. Therefore, a prior probability of $\pi_0 > 0$ needs to be assigned to ~~as~~ the point θ_0 and the remaining probability of $\pi_1 = 1 - \pi_0$ will be spread over $\{\theta \neq \theta_0\}$ using a density g_1.

Simply take θ_0 to be a point mass at θ_0.

If the point null hypothesis approximates an interval ~~vea~~ null hypothesis, $H_0 : \theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$, then π_0 is the probability assigned to the interval $(\theta_0 - \epsilon, \theta_0 + \epsilon)$ by a continuous prior.

The complication now is that the prior π is of the form $\pi(\theta) = \pi_0 I\{\theta = \theta_0\} + (1-\pi_0) g_1(\theta) I\{\theta \neq \theta_0\}$ and hence has both discrete and continuous part.

Now the marginal density of X is

$$m(x) = \pi_0 f(x|\theta_0) + (1-\pi_0) m_1(x)$$

where

$$m_1(x) = \int_{\theta \neq \theta_0} f(x|\theta)\, g_1(\theta)\, d\theta$$

Therefore,

$$\pi(\theta_0|x) = \frac{f(x|\theta_0)\pi_0}{m(x)}$$

$$= \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + (1-\pi_0) m_1(x)}$$

$$= \left\{ 1 + \frac{1-\pi_0}{\pi_0} \frac{m_1(x)}{f(x|\theta_0)} \right\}^{-1}.$$

It follows then the posterior odds ratio is given by

$$\frac{\pi(\theta_0|x)}{1-\pi(\theta_0|x)} = \frac{\pi_0}{(1-\pi_0)} \frac{f(x|\theta_0)}{m_1(x)},$$

and hence the Bayes factor of H_0 relative to H_1, which is the ratio of the above posterior odds ratio to the prior odds ratio of $\pi_0/(1-\pi_0)$ is

$$B = B(x) = BF_{01}(x) = \frac{f(x \mid \theta_0)}{m_1(x)}$$

Thus, $$\pi(\theta_0 \mid x) = \left\{ 1 + \frac{1-\pi_0}{\pi_0} BF_{01}^{-1}(x) \right\}^{-1}$$

- Example Suppose $X \sim B(n, \theta)$ and we want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Under the alternative hypothesis, suppose $\theta \sim \text{Beta}(\alpha, \beta)$. Then ~~$m_1(x)$~~)

$$m_1(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\,\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}$$

so that, $$BF_{01}(x) = \frac{\binom{n}{x} \theta_0^x (1-\theta_0)^{n-x}}{\cancel{\binom{n}{x}}\,\Gamma\ m_1(x)}$$

$$= \frac{\theta_0^x \cancel{(1-\theta_0)}^{n-x}}{\cancel{\frac{\Gamma(\alpha+\beta)}{\Gamma(}}}$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \theta_0^x (1-\theta_0)^{n-x}$$

Hence we obtain,

$$\pi(\theta_0 \mid x) = \left\{ 1 + \frac{1-\pi_0}{\pi_0} BF_{01}^{-1}(x) \right\}^{-1}$$

$$= \left\{ 1 + \frac{1-\pi_0}{\pi_0} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)} \cdot \frac{1}{\theta_0^x(1-\theta_0)^{n-x}} \right\}^{-1}$$

Credible Intervals

For $0 < \alpha < 1$, a $100(1-\alpha)\%$ credible set for θ is a subset $C \subseteq \Theta$ such that

$$P\{C / X = x\} = 1 - \alpha$$

Usually C is taken to be an interval.

Let θ be a continuous random variable; $\theta^{(1)}, \theta^{(2)}$ be $100\,\alpha_1\,\%$ and $100(1-\alpha_2)\,\%$ quantiles with $\alpha_1 + \alpha_2 = \alpha$. Let $C = [\theta^{(1)}, \theta^{(2)}]$. Then $P(C / X = x) = 1 - \alpha$.

Usually $\alpha_1 = \alpha_2 = \alpha/2$ are choosen.

If θ is discrete, usually it would be difficult to find an interval with exact posterior probability $1-\alpha$. There the condition is relaxed to $P(C \mid X = x) \geqslant 1 - \alpha$, with the inequality being as closed to an equality as possible.

HPD (Highest Posterior Density) interval

Suppose the posterior density for θ is unimodal. Then the HPD interval for θ is the interval

$$C = \{\theta : \pi(\theta \mid X = x) \geqslant k\}$$

where k is chosen such that

$$P(C \mid X = x) = 1 - \alpha$$

Prediction of Future Observation

Suppose the data are $x_1, \cdots, x_n$ where $X_1, \cdots, X_n$ are i.i.d. with density $f(x/\theta)$. We want to predict the unobserved X_{n+1} or set up a predictive credible interval for X_{n+1}.

Prediction by a single number $t(x_1, \cdots, x_n)$ based on $x_1, \cdots, x_n$ with squared error loss amounts to considering prediction loss

$$E\{(X_{n+1} - t)^2 / x\} = E\left[\{(X_{n+1} - E(X_{n+1}/x)) - (t - E(X_{n+1}/x))\}^2 / x\right]$$

$$= E\{(X_{n+1} - E(X_{n+1}/x))^2 / x\} + (t - E(X_{n+1}/x))^2$$

which is minimum at $t = E(X_{n+1}/x)$.

To calculate the predictor we need to calculate the predictive distribution

$$\pi(x_{n+1}/x) = \int_{\Theta} \pi(x_{n+1}/x, \theta)\, \pi(\theta/x)\, d\theta$$

$$= \int_{\Theta} f(x_{n+1}/\theta)\, \pi(\theta/x)\, d\theta$$

Let $\mu(\theta) = \int_{-\infty}^{\infty} x f(x/\theta)\, dx$.

Then $E(X_{n+1}/x) = E(\mu(\theta)/x) = \int_{\Theta} \mu(\theta)\, \pi(\theta/x)\, dx$

Now if $f(x/\theta)$ is a $N(\mu, \sigma^2)$ density with σ^2 known

$\mu(\theta) = \mu$ and hence the predictor is

$\int_{-\infty}^{\infty} \mu\, \pi(\mu/x)\, d\mu$ = posterior mean of μ.

Exchangeability

A set of observation $(X_1, X_2, \ldots, X_n)$ are ~~ea~~ called exchangeable, if their joint distribution function is left enaltered if the arguments are permuted.

~~Thus if~~ i.e., $P\{X_1 \le x_1, \ldots, X_n \le x_n\} = P\{X_1 \le x_{i_1} \ldots X_n \le x_{i_n}\}$ for all $n!$ permutations $x_{i_1}, \ldots, x_{i_n}$ of $x_1, \ldots x_n$.

Clearly, if $X_1, X_2, \ldots X_n$ are i.i.d. then they are exchangeable. A simple way of generating exchangeable random variables is to choose an indexing random parameter η and have the random variables conditionally i.i.d given η.

We say X_i, $i = 1, 2, \ldots, n, \ldots$ is a sequence of exchangeable random variables if $\forall n > 1$, $X_1, X_2, \ldots, X_n$ are exchangeable.

Theorem. Suppose X_i's constitute an exchangeable sequence and each X_i takes only values 0 or 1. Then, for some π,

$$P\{X_1 = x_1, \ldots, X_n = x_n\} = \int_0^1 \eta^{\sum_{i=1}^n x_i} (1-\eta)^{n - \sum_{i=1}^n x_i} d\pi(\eta)$$

$\forall n$, $\forall x_1, \ldots, x_n$ equal to 0 or 1, i.e., given η, $X_1, X_2, \ldots, X_n$ are conditionally i.i.d. Bernoulli with parameter η and η has distribution π.

Exercise 1. Three prisoners, A, B, and C, are each held in solitary confinement. A knows that two of them will be hanged and one will be set free but he does not know who will go free. Therefore, he reasons that he has ⅓ chance of survival. He asks the guard who will go free, but has no success there, then we comes up with the following question for the guard:

If two of us must die, then I know that either B or C must die and possibly both. Therefore, if you tell me the name of one who is to die, I learn nothing about my fate; further, because we are kept apart, I cannot reveal it to them. So tell me the name of one of them who is to die.

The guard like this logic and tells A that C will be hanged. A now argues that either he or B will go free, and so now has $\frac{1}{2}$ chance of survival. Is this reasoning correct?

2. There are three chambers, one of which has a prize. The master of ceremonies will give the prize to you if you guess the right chamber correctly. You first make a random guess. Then he shows you one chamber which is empty (The chamber you guessed first has not been opened). You have an option to stick to your original guess or switch to the remaining other chamber. what should you do.

(12). Example 3

Suppose $X|\mu \sim N(\mu, \sigma^2)$, σ^2 known and $\mu \sim N(\eta, \tau^2)$, η and τ^2 known

(a) Show that the joint density $g(x,\mu)$ of X and μ can be written as $g(x,\mu) = \pi(\mu) f(x|\mu) = \frac{1}{2\pi\sigma\tau} \exp\left\{-\frac{1}{2}\left[\frac{(\mu-\eta)^2}{\tau^2} + \frac{(x-\mu)^2}{\sigma^2}\right]\right\}$

$$= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left(-\frac{(x-\eta)^2}{2(\tau^2+\sigma^2)}\right) \times \sqrt{\frac{\tau^2+\sigma^2}{2\pi\tau^2\sigma^2}} \exp\left\{-\frac{\tau^2+\sigma^2}{2\tau^2\sigma^2}\left(\mu - \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\left(\frac{\eta}{\tau^2}+\frac{x}{\sigma^2}\right)\right)^2\right\}$$

(b) Show that the marginal density $m(x)$ of X is

$$m(x) = \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left(-\frac{(x-\eta)^2}{2(\tau^2+\sigma^2)}\right),$$

and the posterior density $\pi(\mu|x)$ of $\mu|X=x$ is

$$\pi(\mu|x) = \sqrt{\frac{\tau^2+\sigma^2}{2\pi\tau^2\sigma^2}} \exp\left\{-\frac{\tau^2+\sigma^2}{2\tau^2\sigma^2}\left(\mu - \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\left(\frac{\eta}{\tau^2}+\frac{x}{\sigma^2}\right)\right)^2\right\}.$$

(c) What are the posterior mean and posterior s.d. of μ given $X=x$?

(d) Instead of a single observation X as above, consider a random sample $X_1, \ldots, X_n$. What is the mininaal sufficient statistics and what is the likelihood function for μ now?

Work out (b) and (c) in this case.

4. Suppose X_1 and X_2 are i.i.d. having the discrete distribution:

$$X = \begin{cases} \theta - 1/2 & \text{w.p. } 1/2 \\ \theta + 1/2 & \text{w.p. } 1/2 \end{cases}$$

Where θ is an unknown real number.

(a) Show that the set C given by

$$C = \begin{cases} \{(x_1+x_2)/2\} & \text{if } x_1 \neq x_2 \\ \{x_1 - 1\} & \text{if } x_1 = x_2 \end{cases}$$

is a 75 % confidence set for θ.

(b) Calculate $P\{C \text{ covers } \theta \mid x_1 - x_2\}$.

5. Let $X_1, X_2, \ldots X_n$ be i.i.d $N(\mu, \sigma^2)$, σ^2 known. Suppose μ has the $N(\eta, \tau^2)$ prior distribution with known η and τ^2

a) Construct the $100(1-\alpha)\%$ HPD credible interval for μ.

b) Construct a $100(1-\alpha)\%$ predictive interval for X_{n+1}.

(c) Consider the uniform prior for this problem by letting $\tau^2 \to \infty$. Work out (a) and (b) in this case.

6. Let $X_1, \ldots X_m$ and $Y_1, \ldots Y_n$ be independent random samples, respectively from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ where σ^2 is known. Construct a $100(1-\alpha)\%$ credible interval for $(\mu_1 - \mu_2)$ assuming a uniform prior on (μ_1, μ_2).

Coherence

There is an alternative way of justifying a Bayesian approach to decision making on the basis of the notion of coherence.

Suppose A stands for a set in the space of θ and x and $A_x = \{\theta : (\theta, x) \in A\}$. Given x, the decision maker (DM)'s uncertainity about A is given by $q(x, A_x)$.

An MC (master of ceremonires) chooses a betting system (A, b), where A is as above and b is a bounded real valued function of x. The DM accepts the ~~gra~~ gamble with pay-off

$$\psi(\theta, x) = b(x)\left[I_A(\theta, x) - q(x, A_x) \right].$$

She gets $b(x)\, q(x, A_x)$ or pays $b(x)\left[1 - q(x, A_x)\right]$ depending on whether θ lies in A_x or not.

The expected pay-off is

$$E(\theta) = \int \psi(\theta, x)\, p(dx \mid \theta)$$

If she accepts k such gambles defined as above by $(A_{(1)}, b_{(1)}), \ldots, (A_{(k)}, b_{(k)})$, then her expected pay off is the sum of the k expected pay-offs. She will face sure loss if

$$\inf_{\theta} \left(\sum_{i=1}^{k} E_i(\theta) \right) > 0$$

Any rational choice of q should avoid sure loss as defined above. Such a choice is said to be coherent if no finite combination of acceptable bets can lead to sure loss.

Robustness and Sensitivity

Intuitively, robaustness means lack of sensitivity of the decision or inference to assumptions in the analysis that may involve a certain degree of uncertainity. In an inference problem, the ~~as~~ assumptions usually involve choice of the model and prior, whereas in a decision problem there is the additional assumption involving the choice of the loss or utility function. An analysis to measure the sensitivity is called sensitivity analysis. Clearly, robustness with respect to all ~~these~~ of the components is desirable. That is to say that reasonable variations from the choice used in the analysis for the model, prior, and loss function do not lead to unreasonable variations in the conclusions arrived at.

Example This example illustrates why sensitivity to the choice of prior can be an important consideration. Suppose we observe X, which follows Poisson (θ) distribution. Further, it is felt a priori that θ has a continuous distribution with median 2 and upper quartile 4, i.e. $P^{\pi}(\theta \leq 2) = 0{\cdot}5 = P^{\pi}(\theta \geq 2)$ and $P^{\pi}(\theta \geq 4) = 0{\cdot}25$.

If these are the only prior inputs available, the following three are candidates for such a prior:

(i) π_1 : $\theta \sim$ exponential (a) with $a = \frac{\log 2}{2}$

(ii) π_2 : $\log(\theta) \sim N\left(\log(2), \left((\log 2)/z_{0.25}\right)^2\right)$; and

(iii) π_3 : $\log(\theta) \sim$ Cauchy $(\log(2), \log(2))$.

Then (i) Under π_1, $\theta | x \sim$ Gamma $(a+1, x+1)$,

so that the posterior mean is $\frac{(a+1)}{(x+1)}$.

(ii) under π_2, if we let $\gamma = \log(\theta)$ and

$\tau = \frac{\log(2)}{z_{0.25}} = \frac{\log(2)}{0.675}$, we obtain

$$E^{\pi_2}(\theta | x) = E^{\pi_2}(\exp(\gamma) | x)$$

$$= \frac{\int_{-\infty}^{\infty} \exp(-e^{\gamma}) \exp(\gamma(x+1)) \exp(-(\gamma - \log 2)^2 / 2\tau^2) d}{\int_{-\infty}^{\infty} \exp(-e^{\gamma}) \exp(\gamma x) \exp(-(\gamma - \log 2)^2 / 2\tau^2) \, d\gamma}$$

and (iii) under π_3, again if let $\gamma = \log \theta$ we get

$$E^{\pi_3}(\theta | x) = E^{\pi_3}(\exp(\gamma) | x)$$

$$= \frac{\int_{-\infty}^{\infty} \exp(-e^{\gamma}) \exp(\gamma(x+1)) \left[1 + \left(\frac{\gamma - \log 2}{\log 2}\right)^2\right]^{-1} d\gamma}{\int_{-\infty}^{\infty} \exp(-e^{\gamma}) \exp(\gamma x) \left[1 + \left(\frac{\gamma - \log 2}{\log 2}\right)^2\right]^{-1} d\gamma}$$

Posterior Means under π_1, π_2 and π_3 as follows:

			x					
π	0	1	2	3	4	5	10	20
π_1	·749	1·485	2·228	2·971	3·713	4·456	8·169	15·595
π_2	·950	1·480	2·106	2·806	3·559	4·353	8·660	17·945
π_3	·761	1·562	2·094	2·633	3·250	3·950	8·867	19·178

To see if the choice of prior matters, simply examine the posterior means under the three different priors.

For small or moderate x ($x \le 10$), there is robustness; the choice of prior does not seem to matter too much. For large values of x, the choice does matter. The inference that a conjugate prior obtains then is quite different from what a heavier tailed prior would obtain. It is now clear that there are situations where it does matter what prior one chooses from a class of priors, each of which is considered reasonable given the available prior information.

Classes of Priors

The goals ~~is~~ to choose a class of priors, Γ, are,

(i) to ensure that as many 'reasonable' priors as possible are included,

(ii) to try to eliminate 'unreasonable' priors,

(iii) to ensure that Γ does not require prior information which is difficult to elicit, and

(iv) to be able to compute measures of robustness without much difficulty.

Example. Suppose θ is a real-valued parameter, prior beliefs about which indicate that it should have a continuous prior distribution, symmetric about 0 and having the third quartile, Q_3, between 1 and 2, Consider, then

$$\Gamma_1 = \{ N(0, \tau^2) : 2 \cdot 19 < \tau^2 < 8 \cdot 76 \} \text{ and}$$

$$\Gamma_2 = \{ \text{symmetric priors with } 1 < Q_3 < 2 \}$$

Even though Γ_1 can be appropriate in some ~~sense~~ cases, it will mostly be considered "rather small". because it contains only sharp-tailed distribution. On the other hand, Γ_2 will typically be "too large", containing priors, shapes of some of which will be considered unreasonable. Starting with Γ_2 and imposing reasonable

Constraints such as unimodality on the priors can lead to sensible classes such as

$$\Gamma_3 = \{\text{unimodal symmetric priors with } 1 < Q_3 < 2\}$$

$\Gamma_1 \subseteq \Gamma_3 \subseteq \Gamma_2$.

Conjugate class

The class considering of conjugate priors is one of the easiest classes of priors to work with.

If $X \sim N(\theta, \sigma^2)$ with known σ^2, the conjugate priors for θ are the normal priors $N(\mu, \tau^2)$.

So one could consider

$$\Gamma_c = \{ N(\mu, \tau^2) : \mu_1 \le \mu \le \mu_2, \tau_1^2 \le \tau^2 \le \tau_2^2 \}$$

for some specified values of μ_1, μ_2, τ_1^2, and τ_2^2.

The advantage with the conjugate class is that posteriors can be calculated in closed form.

In the above case, if $\theta \sim N(\mu, \tau^2)$, then $(\theta \mid X = x) \sim N(\mu^*(x), \delta^2)$, where $\delta^2 = \dfrac{\tau^2\sigma^2}{\tau^2+\sigma^2}$

and $$\mu^*(x) = \frac{\tau^2}{\tau^2+\sigma^2}\, x + \frac{\sigma^2}{\tau^2+\sigma^2}\, \mu .$$

Minimizing and maximizing posterior quantities then becomes any easy task.

Neighbourhood Class

If π_0 is a single elicited prior, then uncertainity in this elicitation can be modeled using the class

$$\Gamma_N = \{ \pi \text{ which are in the neighborhood of } \pi_0 \}$$

A natural and well studied class is the ϵ-contamination class,

$$\Gamma_\epsilon = \{ \pi : \pi = (1-\epsilon)\,\pi_0 + \epsilon q \,,\; q \in Q \}$$

ϵ reflecting the uncertainity in π_0 and Q specifying the contaminations. Some choices for Q are all unimodal distributions with mode θ_0, and all unimodal symmetric distributions with mode θ_0.

The ϵ-contamination class with appropriate choice of Q can provide good robustness.

Density Ratio class

Assuming the existence of densities for all the priors in the class, the density ratio class is defined as

$$\Gamma_{DR} = \{ \pi : L(\theta) \le \alpha \pi(\theta) \le U(\theta) \text{ for some } \alpha > 0 \}$$

$$= \left\{ \pi : \frac{L(\theta)}{U(\theta')} \le \frac{\pi(\theta)}{\pi(\theta')} \le \frac{U(\theta)}{L(\theta')} \text{ for all } \theta, \theta' \right\}$$

for specified non-negative functions L and U.

If we take $L \equiv 1$ and $U \equiv c$ then

$$\Gamma_{DR} = \left\{ \pi : c^{-1} \le \frac{\pi(\theta)}{\pi(\theta')} \le c \,, \text{ for all } \theta, \theta' \right\}$$

Global Measures of Sensitivity

Example : Suppose $X_1, X_2, \cdots X_n$ are i.id $N(\theta, \sigma^2)$, with σ^2 known and Γ be all $N(0, \tau^2)$, $\tau^2 > 0$, priors for θ. Then the variation in the posterior mean is $\left(\inf_{\tau^2 > 0} E(\theta | \bar{x}) , \sup_{\tau^2 > 0} E(\theta | \bar{x}) \right)$.

Because for fixed τ^2, $E(\theta | \bar{x}) = \frac{\tau^2}{\tau^2 + \sigma^2} \bar{x}$, this range can easily be seen to be $(0, \bar{x})$ or $(\bar{x}, 0)$ according as $\bar{x} \geq 0$ or $\bar{x} < 0$. If $\bar{x}$ is small in magnitude, this range will be small. Thus the robustness of the procedure of using posterior mean as the Bayes estimate of θ will depend crucially on the magnitude of the observed value of $\bar{x}$.

As can be seen from the above example, a natural global measure of sensitivity of the Bayesian quantity to the choice of prior is the range of this quantity as the prior varies in the class of priors of interest. Further there are three categories of Bayesian quantities of intereot.

(i) Linear functionals of the prior ;

$\rho(\pi) = \int_{\Theta} h(\theta) \pi(d\theta)$ where h is a given function.

If h is taken to be the likelihood function l, we get an important linear functional, the marginal density ~~df~~ of data, i.e.; $m(\pi) = \int_{\Theta} l(\theta)\, \pi(d\theta)$.

(ii) Ratio of linear functionals of the priors :

$$\rho(\pi) = \frac{1}{m(\pi)} \int_{\Theta} h(\theta)\, l(\theta)\, \pi(d\theta)$$

for some given function h.

~~(iii)~~ If we take $h(\theta) = \theta$, $\rho(\pi)$ is the posterior mean.

For $h(\theta) = I_C(\theta)$, the indicator function of the set C, we get the posterior probability of C.

(iii) Ratio of nonlinear functionals :

$$\rho(\pi) = \frac{1}{m(\pi)} \int_{\Theta} h(\theta, \phi(\pi))\, l(\theta)\, \pi(d\theta)$$

for some given h. For $h(\theta, \phi(\pi)) = (\theta - \mu(\pi))^2$ where $\mu(\pi)$ is the posterior mean, we get

$\rho(\pi)$ is the posterior variance.

Note that extreme values of linear functionals of the prior as it varies in a class Γ are easy to compute if the extreme points of Γ can be identified.

Example

Suppose $X \sim N(\theta, \sigma^2)$, with σ^2 known and the class Γ of interest is

$$\Gamma_{SU} = \{ \text{all symmetric unimodal distributions with mode } \theta_0 \}$$

Then ϕ denoting the standard normal density,

$$m(\pi) = \int_{-\infty}^{\infty} \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right) \pi(\theta)\, d\theta .$$

Note that any unimodal symmetric (about θ_0) density π is a mixture of uniform densities symmetric about θ_0. Thus the extreme points of Γ_{SU} are $U(\theta_0 - r, \theta_0 + r)$ distribution.

Therefore, $\inf\limits_{\pi \in \Gamma_{SU}} m(\pi)$

$$= \inf_{r>0} \frac{1}{2r} \int_{\theta_0 - r}^{\theta_0 + r} \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right) d\theta$$

$$= \inf_{r>0} \frac{1}{2r} \left\{ \Phi\left(\frac{\theta_0 + r - x}{\sigma}\right) - \Phi\left(\frac{\theta_0 - r - x}{\sigma}\right) \right\}.$$

$$\sup_{\pi \in \Gamma_{SU}} m(\pi) = \sup_{r>0} \frac{1}{2r} \int_{\theta_0 - r}^{\theta_0 + r} \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right) d\theta$$

$$= \sup_{r>0} \frac{1}{2r} \left\{ \Phi\left(\frac{\theta_0 + r - x}{\sigma}\right) - \Phi\left(\frac{\theta_0 - r - x}{\sigma}\right) \right\}$$

Lemma

Suppose C_T is a set of probability measures on the real line given by $C_T = \{\nu_t : t \in T\}$, $T \subset R^d$, and let $\mathcal{C}$ be the convex hull of C_T. Further suppose h_1 and h_2 are real-valued functions defined on R such that $\int |h_1(x)| dF(x) < \infty$ for all $F \in \mathcal{C}$, and $K + h_2(x) > 0$ for all x, for some constant K. Then, for any k,

$$\sup_{F \in \mathcal{C}} \frac{k + \int h_1(x)\, dF(x)}{K + \int h_2(x)\, dF(x)} = \sup_{t \in T} \frac{k + \int h_1(x)\, \nu_t(dx)}{K + \int h_2(x)\, \nu_t(dx)}$$

$$\inf_{F \in \mathcal{C}} \frac{k + \int h_1(x)\, dF(x)}{K + \int h_2(x) dF(x)} = \inf_{t \in T} \frac{k + \int h_1(x)\, \nu_t(dx)}{K + \int h_2(x)\, \nu_t(dx)} .$$

Proof Because $\int h_1(x)\, dF(x) = \int h_1(x) \int_T \nu_t(dx)\, \mu(dt)$,

for some probability measure μ on T, using Fubini's theorem,

$$k + \int h_1(x)\, dF(x) = \int (k + h_1(x)) \int_T \nu_t(dx)\, \mu(dt)$$

$$= \int_T \left(\int (k + h_1(x))\, \nu_t(dx) \right) \mu(dt)$$

$$= \int_T \left[\left(\frac{\int (k + h_1(x)\, \nu_t(dx)}{\int (K + h_2(x)\, \nu_t(dx)} \right) \int (K + h_2(x))\, \nu_t(dx) \right] \mu(dt)$$

$$\leq \left(\sup_{t \in T} \frac{\int (K + h_1(x)) \, \nu_t(dx)}{\int (K + h_2(x)) \, \nu_t(dx)} \right) \left(K + \int h_2(x) \, dF(x) \right) :$$

Therefore,

$$\sup_{F \in \mathcal{C}} \frac{K + \int h_1(x) \, dF(x)}{K + \int h_2(x) \, dF(x)} \leq \sup_{t \in T} \frac{\int (K + h_1(x)) \, \nu_t(dx)}{\int (K + h_2(x)) \, \nu_t(dx)}$$

However, because $\mathcal{C} \supset C_T$,

$$\sup_{F \in \mathcal{C}} \frac{K + \int h_1(x) \, dF(x)}{K + \int h_2(x) \, dF(x)} \geq \sup_{t \in T} \frac{\int (K + h_i(x)) \, \nu_t(dx)}{\int (K + h_2(x)) \, \nu_t(dx)}$$

Hence the proof for the supremum, and the proof for the infimum is similar.

Therem Consider the class Γ_{SU} of all symmetric unimodal prior distributions with mode θ_0 . Then it follows that

$$\sup_{\pi \in \Gamma_{SU}} E^{\pi}(g(\theta) \mid x) = \sup_{r > 0} \frac{\frac{1}{2r} \int_{\theta_0 - r}^{\theta_0 + r} g(\theta) f(x \mid \theta) \, d\theta}{\frac{1}{2r} \int_{\theta_0 - r}^{\theta_0 + r} f(x \mid \theta) \, d\theta} ,$$

$$\inf_{\pi \in \Gamma_{SU}} E^{\pi}(g(\theta) \mid x) = \inf_{r > 0} \frac{\frac{1}{2r} \int_{\theta_0 - 0}^{\theta_0 + r} g(\theta) f(x \mid \theta) \, d\theta}{\frac{1}{2r} \int_{\theta_0 - r}^{\theta_0 + r} f(x \mid \theta) \, d\theta}$$

Proof Note that $E^{\pi}(g(\theta) \mid x) = \dfrac{\int g(\theta) f(x \mid \theta) \, d\pi(\theta)}{\int f(x \mid \theta) \, d\pi(\theta)}$,

where $f(x \mid \theta)$ is the density of the data x.

Now using the above lemma and recalling that any unimodal symmetric distribution is a mixture of symmetric uniform distribution, we get the result.

Example Suppose $X \mid \theta \sim N(\theta, \sigma^2)$ and robustness of the posterior mean w.r.t. Γ_{SU} is of interest. Then, range of posterior mean over this class can be easily computed using the previous theorem. Thus

$$\sup_{\pi \in \Gamma_{SU}} E^{\pi}(\theta \mid x) = \sup_{r>0} \frac{\frac{1}{2r} \int_{\theta_0 - r}^{\theta_0 + r} \frac{\theta}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right) d\theta}{\frac{1}{2r} \int_{\theta_0 - r}^{\theta_0 + r} \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right) d\theta}$$

$$= x + \sup_{r>0} \frac{\phi\left(\frac{\theta_0 - r - x}{\sigma}\right) - \phi\left(\frac{\theta_0 + r - x}{\sigma}\right)}{\Phi\left(\frac{\theta_0 + r - x}{\sigma}\right) - \Phi\left(\frac{\theta_0 - r - x}{\sigma}\right)}$$

$$\inf_{\pi \in \Gamma_{SU}} E^{\pi}(\theta \mid x) = x + \inf_{r>0} \frac{\phi\left(\frac{\theta_0 - r - x}{\sigma}\right) - \phi\left(\frac{\theta_0 + r - x}{\sigma}\right)}{\Phi\left(\frac{\theta_0 + r - x}{\sigma}\right) - \Phi\left(\frac{\theta_0 - r - r}{\sigma}\right)}$$

Example Suppose $X \mid \theta \sim N(\theta, \sigma^2)$ and it is of interest to test $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$. Again, suppose that Γ_{SU} is the class of priors to be considered and robustness of this class is to be examined. Because

$$P^{\pi}(H_0 \mid x) = P^{\pi}(\theta \leq \theta_0 \mid x)$$

$$= \frac{\int_{-\infty}^{\infty} I_{[-\infty, \theta_0]}(\theta) f(x \mid \theta)\, d\pi(\theta)}{\int_{-\infty}^{\infty} f(x \mid \theta)\, d\pi(\theta)}$$

appling the previous theorem, we get

$$\sup_{\pi \in \Gamma_{SU}} p^{\pi}(H_0|x) = \sup_{r>0} \frac{\frac{1}{2r}\int_{\theta_0 - r}^{\theta_0} \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right) d\theta}{\frac{1}{2r}\int_{\theta_0 - r}^{\theta_0 + r} \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right) d\theta}$$

$$= \sup_{r>0} \frac{\Phi\left(\frac{\theta_0 - x}{\sigma}\right) - \Phi\left(\frac{\theta_0 - r - x}{\sigma}\right)}{\Phi\left(\frac{\theta_0 + r - x}{\sigma}\right) - \Phi\left(\frac{\theta_0 - r - x}{\sigma}\right)}$$

and similarly,

$$\inf_{\pi \in \Gamma_{SU}} p^{\pi}(H_0|x) = \inf_{r>0} \frac{\Phi\left(\frac{\theta_0 - x}{\sigma}\right) - \Phi\left(\frac{\theta_0 - r - x}{\sigma}\right)}{\Phi\left(\frac{\theta_0 + r - x}{\sigma}\right) - \Phi\left(\frac{\theta_0 - r - x}{\sigma}\right)}$$

It can be seen that the above bounds are, respectively, 0·5 and α, where $\alpha = \Phi\left(\frac{x-\theta_0}{\sigma}\right)$, the P-value.

We shall now consider the density-ratio class that was metioned earlier is given by

$$\Gamma_{DR} = \{\pi : L(\theta) \leq \alpha\pi(\theta) \leq U(\theta) \text{ for some } \alpha > 0\}$$

for specified non-negative functions L and U.

For $\pi \in \Gamma_{DR}$ and any real valued π- integrable function h on the parameter space Θ, let

$\pi(h) = \int_{\Theta} h(\theta)\,\pi(d\theta)$. and let $h \equiv h^{+} - h^{-}$,

Theorem

For U-integrable functions h_1 and h_2, with h_2 positive a.s. w.r.t. all $\pi \in \Gamma_{DR}$,

$$\inf_{\pi \in \Gamma_{DR}} \frac{\pi(h_1)}{\pi(h_2)}$$ is the unique solution λ of

$$U(h_1 - \lambda h_2)^- + L(h_1 - \lambda h_2)^+ = 0$$

$$\sup_{\pi \in \Gamma_{DR}} \frac{\pi(h_1)}{\pi(h_2)}$$ is the unique solution λ of

$$U(h_1 - \lambda h_2)^+ + L(h_1 - \lambda h_2)^- = 0$$

Proof Let $\lambda_0 = \inf\limits_{\pi \in \Gamma_{DR}} \frac{\pi(h_1)}{\pi(h_2)}$, $c_1 = \inf\limits_{\pi \in \Gamma_{DR}} \pi(h_2)$ and $c_2 = \sup\limits_{\pi \in \Gamma_{DR}} \pi(h_2)$. Then $0 < c_1 < c_2 < \infty$, and $|\lambda_0| < \infty$. Because $U(h_1 - \lambda h_2)^- + L(h_1 - \lambda h_2)^+ = \inf\limits_{\pi \in \Gamma_{DR}} \pi(h_1 - \lambda h_2)$ for any λ,

note that $\lambda_0 \geq \lambda$ iff $U(h_1 - \lambda h_2)^- + L(h_1 - \lambda h_2)^+ \geq 0$

However, $\lambda_0 > \lambda$ iff $U(h_1 - \lambda h_2)^- + L(h_1 - \lambda h_2)^+ > 0$.

A Si Hence $\lambda_0 = \lambda$ iff $U(h_1 - \lambda h_2)^- + L(h_1 - \lambda h_2)^+ = 0$

A similar argument for the supremum.

Example

Suppose $X \sim N(\theta, \sigma^2)$, with σ^2 known. Consider the class Γ_{DR} with L being the Lebesgue measure and $U = kL$, $k > 1$. Because the posterior mean is

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

we have that $\inf\limits_{\pi \in \Gamma_{DR}} E^{\pi}(\theta|x)$ is the unique solution λ of

$$k \int_{-\infty}^{\lambda} (\theta - \lambda) f(x|\theta)\, d\theta + \int_{\lambda}^{\infty} (\theta - \lambda) f(x|\theta)\, d\theta = 0$$

and similarly, $\sup\limits_{\pi \in \Gamma_{DR}} E^{\pi}(\theta|x)$ is the unique solution λ of

$$\int_{-\infty}^{\lambda} (\theta - \lambda) f(x|\theta)\, d\theta + k \int_{\lambda}^{\infty} (\theta - \lambda) f(x|\theta)\, d\theta = 0$$

Noting that $f(x|\theta) = \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right) = \frac{1}{\sigma} \phi\left(\frac{\theta - x}{\sigma}\right)$, and letting λ_1 be the minimum and λ_2 the maximum, the above equations may be rewritten as

$$(k-1)\left[\left(\frac{\lambda_1-x}{\sigma}\right)\Phi\left(\frac{\lambda_1-x}{\sigma}\right)+\phi\left(\frac{\lambda_1-x}{\sigma}\right)\right]=\frac{\lambda_1-x}{\sigma},$$

$$(k-1)\left[\left(\frac{\lambda_2-x}{\sigma}\right)\Phi\left(\frac{\lambda_2-x}{\sigma}\right)+\phi\left(\frac{\lambda_2-x}{\sigma}\right)\right]=k\left(\frac{\lambda_2-x}{\sigma}\right).$$

Now let $k\left(\frac{\lambda_1-x}{\sigma}\right)=\gamma$, Then $\lambda_2 = x+\sigma\frac{\gamma}{k}$

Put $\lambda_0 = x-\sigma\frac{\gamma}{k}$ or $\frac{\lambda_0-x}{\sigma} = -\frac{\gamma}{k}$,

Then we see from the second equation above that

$$(k-1)\left[\left(\frac{\lambda_0-x}{\sigma}\right)\Phi\left(\frac{\lambda_0-x}{\sigma}\right)+\phi\left(\frac{\lambda_0-x}{\sigma}\right)\right]$$

$$=(k-1)\left[-\frac{\gamma}{k}\Phi\left(-\frac{\gamma}{k}\right)+\phi\left(-\frac{\gamma}{k}\right)\right]$$

$$=(k-1)\left[-\frac{\gamma}{k}\left(1-\Phi(\gamma/k)\right)+\phi\left(\frac{\gamma}{k}\right)\right]$$

$$=(k-1)\left[\gamma/k\,\Phi(\gamma/k)+\phi(\gamma/k)\right]-(k-1)\gamma/k$$

$$=0$$

implying that once λ_2 is obtained, say $\lambda_2 = x+\sigma\cdot\gamma/k$, the solution for λ_1 is simply $x-\sigma\frac{\gamma}{k}$.

Different Methods of Construction of Objective Priors:

To construct objective priors under general regularity conditions, one may do one of the following things.

1. Define a uniform distribution that takes into account the geometry of the parameter space.

2. Minimize a suitable measure of information in the prior.

3. Choose a prior with some form of frequentist ideas because a prior with little information should lead ~~lead~~ to inference that is similar to frequentist inference.

Although the usual uniform prior $\pi(\theta) = c$ has come in for a lot of criticism, these criticisms help one understand the motivation behind (1) and (2). The fact is that both (1) and (2) lead to the Jeffreys prior,

namely, $\pi(\theta) = [\det(I_{ij}(\theta))]^{1/2}$

where $(I_{ij}(\theta))$ is the Fisher information matrix.

In the one-dimensional case (3) also leads to the Jeffreys prior.

Uniform Distribution and Its Criticisms

The first objective prior ever to be used is the uniform distribution over a bounded interval.

A common argument, based on "ignorance", seems to have been that if we know nothing about θ, why should we attach more density to one point than another? ~~The argumen~~

A second argument is that the uniform maximizes the shannon entropy.

The principle of ignorance has been criticized by many statisticians. Essentially, the criticism is based on an invariance argumat.

Let $\eta = \psi(\theta)$ be a one-to-one function of θ. If we know nothing about θ, then we know nothing for η also. So the principle of ignorance applied to η will imply our prior for η is uniform on $\psi(\Theta)$ just as it had led to a uniform prior for θ. But this leads to a contradiction. To see this suppose ψ is a differentiable and $p(\eta) = c$ on $\psi(\Theta)$.

Then the prior $p^*(\theta)$ for θ is

$$p^*(\theta) = p(\eta)\,|\psi'(\theta)| = c\,|\psi'(\theta)|$$

Which is not a constant in general.

This argument also leads to an invariance principle. Suppose we have an algorithm that produces noninformative priors for both θ and η, then these priors $p^*(\theta)$ and $p(\eta)$ should be connected by the equation

$$p^*(\theta) = p(\eta)\,|\psi'(\theta)|$$

i.e., a noninformative prior should be invariant under one-to-one differential transformation.

The second argument in favour of the uniform, based on shannon entropy, is also flawed. Shannon derived a measure of entropy in the finite discrete case from certain natural axioms. His entropy is $H(p) = -\sum_{i=1}^{m} p_i \log p_i$ which is maximized by the discrete uniform, i.e., at $p = (\frac{1}{m}, \dots, \frac{1}{m})$.

Entropy is a measure of the amount of uncertainty about the outcomes of the experiment. A prior that maximizes this will maximize the uncertainity, so it is a noninformative prior. Because such a prior should minimize the information, we take negative of entropy as information.

Shannon's entropy is a natural measure in the discrete case and the discrete uniform appears to be the right noninformative prior. The continuous case is an entriely different matter. Shannon pointed out that for a density p, $H(p) = -\int (\log p(x))\, p(x)\, dx$ is unsatisfactory, clearly it is not desired from axioms, it is not invariant under one-to-one transformation, also, it depends on the measure $\mu(x)\,dx$ w.r.t which the density $p(x)$ is taken. Note also athat the measure $\mu(x)\,dx$ is not non-negative. Just take $\mu(x)=1$ and take $p(x)$ ~~is~~ = uniform on $[0, c]$, Then $H(p) > 0$ iff $c > 1$.

Finally, if the density is take w.r.t $\mu(x)\,dx$, then it is easy to verify that the density is p/μ and

$$H(p) = -\int \left(\log \frac{p(x)}{\mu(x)}\right) \frac{p(x)}{\mu(x)}\, \mu(x)\, dx$$

is maximized at $p = \mu$, i.e., the entropy is maximum at the ~~arbitrary μ~~. arbitrary μ. For all these reasons, we do not think $H(P)$ Is the right entropy to maximize. However, $H(P)$ serves a useful purpose when we have partial information.

Jeffreys Priors as a Uniform Distribution

Suppose $\Theta = \mathbb{R}^d$ and $I(\theta) = (I_{ij}(\theta))$ is the $d \times d$ Fisher information matrix. we assum $I(\theta)$ is positive definite for all θ. Rao had proposed the Riemannian metric ρ related to $I(\theta)$ by

$$\rho(\theta, \theta + d\theta) = \sum_{i,j} I_{i,j}(\theta)\, d\theta_i\, d\theta_j \left(1 + o(1)\right)$$

It is known that this is the unique Rieemannian metric that transforms suitably under one-one differentiable transformations on Θ. Notice that in general Θ does not inherit the usual Euclidean metric that goes with the improper uniform distribution over $\mathbb{R}^d$.

Fix a θ_0 and let $\psi(\theta)$ be a smooth one-to-one transformation such that the information matrix $I^{\psi} = \left[E_{\psi}\left(\frac{\partial \log p}{\partial \psi_i} \cdot \frac{\partial \log p}{\partial \psi_j}\right)\right]$ is the identity matrix I at $\psi_0 = \psi(\theta_0)$. This implies the local geometry in the ψ-space around ψ_0 is Euclidean and hence $d\psi$ is a suitable uniform distribution there.

If we now lift this back to the θ-space by using the Jacobian of transformation and the simple fact

$$\left(\left|\frac{\partial \theta_j}{\partial \psi_i}\right|\right)\left(I_{ij}(\theta)\right)\left(\left|\frac{\partial \theta_j}{\partial \psi_i}\right|\right)' = I^{\psi_0} = I ,$$

we get the Jeffreys prior in the θ-space,

$$d\psi = \left\{ \det \left| \frac{\partial \theta_i}{\partial \psi_j} \right| \right\}^{-1} d\theta = \left\{\det [I_{ij}(\theta)]\right\}^{1/2} d\theta .$$

A similar method present an alternative construction where one takes a compact subset of the parameter space and approximates this by a finite set of points in the so-called Hellinger metric

$$d(P_\theta, P_{\theta'}) = \left[\int (\sqrt{p_\theta} - \sqrt{p_{\theta'}})^2 dx \right]^{1/2}$$

where p_θ and $p_{\theta'}$ are the densities of P_θ and $P_{\theta'}$. One then puts a discrete uniform distribution on the approximating finite set of points and lets the degree of approximation tends to zero. Then the corresponding discrete uniforms converge weakly to the Jeffreys distribution.

Jeffreys Prior as a Minimizer of Information

Let the Shannon entropy associated with a random variable or vector Z be denoted by

$$H(Z) = H(p) = -E_p(\log p(Z))$$

where p is the density of Z. Let $X = (x_1, x_2, \cdot\, x_n)$ have density or probability function $p(x|\theta)$ where θ has prior ~~distribution~~ density $p(\theta)$. We assume $X_1, X_2, \ldots X_n$ are i.i.d and conditions for asymptotic nomality of posterior $p(\theta/x)$ hold. We know that $H(p)$ is not a good measure of entropy and $-H(p)$ not a good measure of information if p is a density. Bernardo suggestedd. Suggested a Kullback-Leibler divergence between prior and posterior, namely,

$$J(p(\theta), X) = E\left\{\log \frac{p(\theta|X)}{p(\theta)}\right\}$$

$$= \int_{\Theta} \left\{ \int_{\mathcal{X}} \left[\int_{\Theta} \log \left\{ \frac{p(\theta'|x)}{p(\theta')} \right\} p(\theta'|x)\, d\theta' \right] p(x|\theta)\, dx \right\} p(\theta)d$$

is a better measure of entropy and $-J$ a better measure of information in the poior.

To get a feeling for this, notice that if the prior is nearly degenerate, at say some θ_0, so will be the posterior. This would imply J is nearly zero. On the other hand if $p(\theta)$ is rather diffuse, $p(\theta|x)$ will ~~be~~ differ a lot from $p(\theta)$, at least for moderate or large n, because $p(\theta|x)$ would be quite peaked. In-fact $p(\theta|x)$ would be approximately normal with mean $\hat{\theta}$ and variance of the order $O(\frac{1}{n})$. The substantial difference between prior and posterior would be reflected by a large value of J.

To ~~sup~~ sum up J is small when p is nearly degenerate and large when p is diffuse, i.e, J captures how diffuse is the prior. It therefore makes sense to minimize J with respect to the prior.

Aftr a lot of calculation one can show that the Jeffreys prior minimize J.

5)

Measure of entropy or information

Shannon introduced missing information in the context of a noisy channel. Any channel has a source that produces (say, per second) messages X with p.m.f $P_X(x)$ and entropy

$$H(X) = -\sum_{x} p_X(x) \log P_X(x).$$

A channel will have an output Y (per second) with entropy $H(Y) = -\sum_{y} P_Y(y) \log P_Y(y)$.

If the channel is noiseless, then $H(Y) = H(X)$.

If the channel is noisy, Y given X is still random. Let $p(x,y)$ denote their joint p.m.f.

The joint entroy is

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y).$$

Clearly, $H(X,Y) = H(X) + H_X(Y) = H(Y) + H_Y(X)$

~~and $H(X,Y)$~~ $H_Y(X)$ is called the equivocation or average ambiguity about input X given only output Y. It is the information about input X that is received given the output Y.

It is the amount of additional information that must be supplied per unit time at the receiving end to correct the received message.

Thus $H_Y(X)$ is the missing information. So amount of information produced in the channel (per unit time) is $H(X) - H_Y(X)$

which may be shown to be non-negative by Shannon's basic results,

$$H(X) + H(Y) \geq H(X,Y) = H(Y) + H_Y(X).$$

In statistical problems, we take X to be θ and Y to be observation vector X. Then $H(\theta) - H_X(\theta)$ is the same measure as before, namely

$$E\left(\log \frac{p(\theta/X)}{p(\theta)}\right)$$

The maximum of $H(X) - H_Y(X)$ with respect to the soure, i.e, w.r.t. $p(x)$ is what Shannon calls the capacity of the channel. Over compact rectangles, the Jeffreys prior is this maximizing distribution for the statistical channel.

Hypothesis testing and Model Selection

Suppose X having density $f(x/\theta)$ is observed, with θ being an unknown element of the parameter space Θ. Suppose that we are interested in comparing two models M_0 and M_1, which is given by

M_0 : X has density $f(x/\theta)$ where $\theta \in \Theta_0$

M_1 : X has density $f(x/\theta)$ where $\theta \in \Theta_1$

For $i = 0, 1$ let $g_i(\theta)$ be the prior density of θ, conditional on M_i being the true model. Then, to compare models M_0 and M_1 on the basis of a random sample $x = (x_1, x_2, \dots x_n)$ one would use the Bayes factor $B_{01}(x) = \dfrac{m_0(x)}{m_1(x)}$

where $$m_i(x) = \int_{\Theta_i} f(x/\theta)\, g_i(\theta)\, d\theta, \qquad i = 0, 1$$

We also use the notation BF_{01} for the Bayes factor, ie which is the ratio of posterior odds ratio of the hypotheses to the corresponding prior odds ratio. Therefore, if the prior probabilities of the hypotheses, $\pi_0 = P^{\pi}(M_0) = P^{\pi}(\Theta_0)$ and $\pi_1 = P^{\pi}(M_1) = P^{\pi}(\Theta_1) = 1 - \pi_0$ are specified, then $P(M_0|x) = \left\{ 1 + \frac{1-\pi_0}{\pi_0} B_{01}^{-1}(x) \right\}^{-1}$.

Thus, if conditional prior densities g_0 and g_1 can be specified, one should simply use the Bayes factor B_{01} for model selection.

If, further π_0 is also specified, the posterior odds ratio of M_0 to M_1 can also be utilized. However, these computations may not always be easy to perform, even when the required prior ingredients are fully specified. A possible solution is the use of BIC as an approximation to a Bayes factor.

Example Consider the problem that is usually called nonparametric regression. Independent responses y_i are observed along with covariates x_i, $i=1,2,\dots n$.

The model of interest is

$$y_i = g(x_i) + \epsilon_i \quad , \quad i=1,2,\dots n$$

where ϵ_i are iid $N(0,\sigma^2)$ errors with unknown error variance σ^2. The function g is called the regression function.

In linear regression, g is a priori assumed to be ~~fully unknown also~~. linear in a set of finite regression coefficients. In general, g can be assumed to be fully unknown also. Now, if model selection involves choosing g from two different fully nonparametric classes of regression functions, this becomes a very difficult problem. Various simplifications including reducing g to be semi-parametric have been studied.

Consider a different model checking problem, that of testing for normality. In its simplest ~~for~~ form, the problem can be stated as checking whether a given random sample $X_1, X_2, \dots, X_n$ arose from a population having the normal distribution. We may write it as

M_0: X is $N(\mu, \sigma^2)$ with arbitrary μ and $\sigma^2 > 0$.

M_1 : X does not have the normal distribution.

However, this looks quite different, because M_1 does not constitute a parametric alternative. Hence it is not clear how to use Bayes factors or posterior odds ratios here for model checking.

P- value and Posterior Probability of Ho as meaoures of Evidence against the Null.

One particular tool from classical statistics that is very widely used in applied sciences for model checking or hypothesis testing is the P-value. It ~~is~~ also happens to be one of the concepts that is highly misunderstood and misused. The basic idea behind the definition of P- value is,

It is the probability ~~of~~ under a null hypothesis of obtaining a value of a test statistic that is at least as extreme as that observed in the sample data.

Suppose that it is desired to test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

and that a classical significance test is available and is based on a test statistic $T(X)$, large values of which are deemed to provide evidence against the null hypothesis. If ~~a~~ data $X = x$ is observed, with corresponding $t = T(x)$, the P- value then is

$$\alpha = P_{\theta_0}(T(X) \geqslant T(x))$$

Example

Suppose we observed $X_1, X_2, \ldots, X_n$ i.i.d from $N(\theta, \sigma^2)$; where σ^2 is known . Then $\bar{X}$ is sufficient for θ and it has the $N(\theta, \sigma^2/n)$ distribution.

Noting that $T = T(\bar{X}) = |\sqrt{n}\,(\bar{X} - \theta_0)/\sigma|$ is a natural test statistics to test

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

one obtains the usual P- value as $\alpha = 2[1 - \Phi(t)]$ where $t = |\sqrt{n}(\bar{x} - \theta_0)/\sigma|$ and Φ is the standard normal cumulative distribution function.

Fisher meant P- value to be used informally as a measure of degree of surprise in the data relative to H_0. This use of P- value as a post-experimental or conditional measure of statistical evidence seems to have some intuitive justification.

Bounds on Bayes Factors and Posterior Probabilities

Consider the following example where P-values and the posterior probabilities are very different.

We observe $\bar{X} \sim N(\theta, \sigma^2/n)$, with know σ^2. Upon using $T = |\sqrt{n}(\bar{X} - \theta_0)/\sigma|$ as the test statistics to test

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

Then the P-values comes out to be $\alpha = 2[1-\Phi(t)]$.

On the set $\{\theta \neq \theta_0\}$, let θ have the density g_1 of $N(\mu, \tau^2)$. Then we have

$$B_{01} = \sqrt{1+\rho^{-2}} \; \exp\left\{-\frac{1}{2}\left[\frac{(t-\rho\eta)^2}{(1+\rho^2)} - \eta^2\right]\right\}$$

where $\rho = \sigma/(\sqrt{n}\,\tau)$ and $\eta = (\theta_0 - \mu)/\tau$.

Now, if we choose $\mu = \theta_0$, $\tau = \sigma$ and $\pi_0 = \frac{1}{2}$ we get $B_{01} = \sqrt{1+\rho^{-2}} \; \exp\left\{-\frac{1}{2}\left[\frac{t^2}{1+\rho^2}\right]\right\}$

For various values of t and n, the different measures of evidence, α = P-value, B = Bayes factor, and $P = P(H_0|x)$ are ~~displayed~~ is as follows.

		n											
		1		5		10		20		50		100	
t	α	B	P	B	P	B	P	B	P	B	P	B	P
1·645	0·10	·72	·42										
1·960	0·05	·54	·35										
2·576	0·01	·27	·21										
3·291	·001	·10	·09										

It may be noted that the posterior probability of Ho varies between 4 to 50 times the corresponding P-value which is an indication of how different these two measures of evidence can be.

Bayesian Computation

Bayesian analysis requires lots of computation of expectations and quantiles of probability distributions that arise as posterior distributions. Modes of the densities of such distributions are also sometimes used. The standard Bayes estimate is the posterior mean, which is also the Bayes rule under the squared error loss. Its accuracy is assessed using the posterior variance, which is again an expected value. Posterior median is sometimes utilized, and to provide Bayesian credible regions, quantilies of posterior distributions are needed. If conjugate priors are not used, as is mostly the case these days, posterior distributions will not be standard distributions and hence the required Bayesian quantities (i.e., posterior quantities of inferential interest) cannot be computed in closed from. The special techniques are needed for Bayesian computations.

Example 1 Suppose X is $N(\theta, \sigma^2)$ with known σ^2 and Cauchy (μ, τ) prior on θ is considered appropriate from ~~θ~~ robustness considerations.

Then $\pi(\theta|x) \propto \exp\left(-(\theta-x)^2/(2\sigma^2)\right)\left(\tau^2+(\theta-\mu)^2\right)^{-1}$

and hence the posterior mean and variance are

$$E^{\pi}(\theta|x) = \frac{\int_{-\infty}^{\infty} \theta \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\left(\tau^2+(\theta-\mu)^2\right)^{-1} d\theta}{\int^{\infty} \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\left(\tau^2+(\theta-\mu)^2\right)^{-1} d\theta}$$

and $$v^{\pi}(\theta|x) = \frac{\int_{-\infty}^{\infty} \theta^2 \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\left(\tau^2+(\theta-\mu)^2\right)^{-1} d\theta}{\int_{-\infty}^{\infty} \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\left(\tau^2+(\theta-\mu)^2\right)^{-1} d\theta} - \left(E^{\pi}(\theta|x)\right)^2$$

Note that the above integrals cannot be computed in closed form, but various numerical integration techniques such as IMSL routines or Gaussian quadrature can be efficiently used to obtain very good approximations of these.

Example 2. Suppose $X_1, X_2, \ldots, X_K$ are independent Poisson counts with $X_i \sim$ Poisson (θ_i). θ_i are a priori considered related, and a joint multivariate normal prior distribution on their logarithm is assumed. Specifically, let $\nu_i = \log(\theta_i)$ be the ith element of $\underset{\sim}{\nu}$ and suppose

$$\underset{\sim}{\nu} \sim N_K\left(\mu \underset{\sim}{1}, \tau^2\{(1-\rho) I_K + \rho \underset{\sim}{1}\underset{\sim}{1}'\}\right)$$

where $\underset{\sim}{1}$ is the K-vector with all elements being 1, and μ, τ^2 and ρ are known constants.

Then, because

$$f(\underset{\sim}{x}|\underset{\sim}{\nu}) = \exp\left(-\sum_{i=1}^{K}\{e^{\nu_i} - \nu_i x_i\}\right) \Big/ \prod_{i=1}^{K} x_i!$$

and $$\pi(\underset{\sim}{\nu}) \propto \exp\left(-\frac{1}{2\tau^2}(\underset{\sim}{\nu} - \mu\underset{\sim}{1})'\left((1-\rho)I_K + \rho\underset{\sim}{1}\underset{\sim}{1}'\right)^{-1}(\underset{\sim}{\nu} - \mu\underset{\sim}{1})\right)$$

we have that $\pi(\underset{\sim}{\nu}|\underset{\sim}{x}) \propto$

$$\exp\left\{-\sum_{i=1}^{K}\{e^{\nu_i} - \nu_i x_i\} - \frac{1}{2\tau^2}(\nu - \mu 1)'\left((1-\rho) I_K + \rho\underset{\sim}{1}\underset{\sim}{1}'\right)^{-1}(\underset{\sim}{\nu} - \mu\underset{\sim}{1})\right\}$$

There fore, if the posterior mean of θ_j is of interest, we need to compute,

$$E^{\pi}(\theta_j|x) = E^{\pi}(\exp(\nu_j)/x) = \frac{\int_{R^k} \exp(\nu_j)\, g(\underline{\nu}|\underline{x})\, d\underline{\nu}}{\int_{R^k} g(\underline{\nu}|\underline{x})\, d\underline{\nu}}$$

where $g(\underline{\nu}|\underline{x}) =$

$$\exp\Big\{-\sum_{i=1}^{k}\{e^{\nu_i}-\nu_i x_i\} - \frac{1}{2\tau^2}(\underline{\nu}-\mu\underline{1})'((1-\rho)I_k+\rho\underline{1}\underline{1}')^{-1}(\nu-\mu\underline{1})\Big\}$$

This is a ratio of two k-dimensional integrals and as k grows, the integrals become less and less easy to work with. Numerical integration techniques fail to be an efficient technique in this case. In fact, numerical integration techniques are presently not preferred except for single and two-dimensional integrals.

Monte Carlo Sampling

Consider an expectation that is not available in closed ~~from~~ form. An alternative to numerical integration or analytic approximation to compute this is statistical sampling. To estimate a population mean or a population proportion, a natural approach is to gather a large sample from this population and to consider the corresponding sample mean on the sample proportion. The law of large numbers guarantees that the estimates so obtained will be good provided the sample is large enough.

specifically, let f be a probability density function (or a mass function) and suppose the quantity of interest is a finite expectation of the form .

$$E_f h(\underline{x}) = \int_{\mathcal{X}} h(\underline{x}) f(\underline{x})\, d\underline{x}$$

(or the corresponding sum in the discrete case). If i.i.d. observations $X_1, X_2, \ldots$ can be generated from the density f, then

$$\bar{h}_m = \frac{1}{m} \sum_{i=1}^{m} h(\underline{x}_i)$$

converges in probability (or even almost surely) to $E_f\, h(\underline{X})$. This justifies using $\bar{h}_m$ as an approximation for $E_f h(\underline{x})$ for large m. To provide a measure of accuracy or the extent of error in the approximation, we can again use a statistical technique and compute the standar error. If $Var_f\, h(\underline{X})$ is finite, then $Var_f(\bar{h}_m) = Var_f\, h(\underline{X})/m$. Further $Var_f\, h(\underline{x}) = E_f h^2(\underline{x}) - (E_f h(\underline{x}))^2$ can be estimated by

$$S_m^2 = \frac{1}{m} \sum_{i=1}^{m} \left(h(\underline{x}_i) - \bar{h}_m\right)^2$$

and hence the standard error of $\bar{h}_m$ can be estimated by

$$\frac{1}{\sqrt{m}} S_m = \frac{1}{m}\left(\sum_{i=1}^{m}\left(h(\underline{x}_i) - \bar{h}_m\right)^2\right)^{1/2}$$

Confidence intervals for $E_f h(\underline{x})$ can ~~be~~ also be provided using the CLT. Because

$$\frac{\sqrt{m}\left(\bar{h}_m - E_f h(\underline{x})\right)}{S_m} \xrightarrow{d} N(0,1) \quad \text{as } m \to \infty$$

in distribution, $\left(\bar{h}_m - z_{\alpha/2} S_m/\sqrt{m},\ \bar{h}_m + z_{\alpha/2} S_m/\sqrt{m}\right)$ can be used as an approximate $100(1-\alpha)\%$ confidence interval for $E_f h(\underline{x})$, with $z_{\alpha/2}$ denoting the $100(1-\alpha/2)\%$ quantile of standard normal.

The above discussion suggests that if we want ~~to~~ to approximate the posterior mean, we could try to generate i.i.d. observations from the posterior distribution and consider the mean of this sample. This is rarely useful because most often the posterior distribution will be a non-standard distribution which may not easily allow sampling from it.

Example Recall the example 1.

$$E^{\pi}(\theta|x) = \frac{\int_{-\infty}^{\infty} \theta \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\left(\tau^2+(\theta-\mu)^2\right)^{-1} d\theta}{\int_{-\infty}^{\infty} \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\left(\tau^2+(\theta-\mu)^2\right)^{-1} d\theta}$$

$$= \frac{\int_{-\infty}^{\infty} \theta \left\{\frac{1}{\sigma}\phi\left(\frac{\theta-x}{\sigma}\right)\right\}\left(\tau^2+(\theta-\mu)^2\right)^{-1} d\theta}{\int_{-\infty}^{\infty} \left\{\frac{1}{\sigma}\cancel{\theta-x}\,\phi\left(\frac{\theta-x}{\sigma}\right)\left(\tau^2+(\theta-\mu)^2\right)^{-1} d\theta}$$

Where ϕ denotes the density of standard normal. Thus $E^{\pi}(\theta|x)$ is the ratio of expectation of ~~h(θ~~ $h(\theta) = \theta/(\tau^2+(\theta-\mu)^2)$ to that of $h(\theta) = 1/(\tau^2+(\theta-\mu)^2)$ both expectation being with respect to the $N(x,\sigma^2)$ distribution. Therefore we simply sample $\theta_1, \theta_2, \ldots$ from $N(x,\sigma^2)$ and use

$$\widehat{E^{\pi}(\theta|x)} = \frac{\sum_{i=1}^{m} \theta_i\left(\tau^2+(\theta_i-\mu)^2\right)^{-1}}{\sum_{i=1}^{m}\left(\tau^2+(\theta_i-\mu)^2\right)^{-1}}$$

as our Monte carlo estimate of $E^{\pi}(\theta|x)$.

But the problem has not been completely solved. The sample of θ's generated from $N(x,\sigma^2)$ will tend to concentrate around x, whereas to satisfactorily account for the contribution of the Cauchy prior to the posterior mean, a significant portion of the θ's should come from the tails of the posterior distribution.

It may therefore appear that it is perhaps better to express the posterior mean in the form

$$E^{\pi}(\theta|x) = \frac{\int_{-\infty}^{\infty} \theta \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\pi(\theta)\,d\theta}{\int_{-\infty}^{\infty} \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\pi(\theta)\,d\theta}$$

then sample θ's from Cauchy (μ, τ) and use the approximation

$$\widehat{E^{\pi}}(\theta|x) = \frac{\sum_{i=1}^{m} \theta_i \exp\left(-\frac{(\theta_i-x)^2}{2\sigma^2}\right)}{\sum_{i=1}^{m} \exp\left(-\frac{(\theta_i-x)^2}{2\sigma^2}\right)}$$

However, this is also not totally satisfactory because the tails of the posterior distribution are not as heavy as those of the Cauchy prior, and hence there will be excess sampling from the tails relative to the center. To implication is that the convergence of the approximation is slower and hence a larger error in approximation (for a fixed m). Ideally, therefore, sampling should be from the posterior distribution itself for a satisfactory apponimation. With this view in mind, a variation of the above theme has been develop. This is called the Monte Carlo importance sampling.

Consider $E_f(h(\underline{x})) = \int_{\mathcal{X}} h(\underline{x}) f(\underline{x}) \, d\underline{x}$ .... $(*)$

~~$= \int_{\mathcal{X}} h(x$~~

Suppose that it is difficult or expensive to sample directly from f, but there exists a probability density u that is very close. to f from ~~wit~~ which it is easy to sample. Then can rewrite $(*)$ as

$$E_f(h(\underline{x})) = \int_{\mathcal{X}} h(\underline{x}) f(\underline{x}) \, d\underline{x}$$

$$= \int_{\mathcal{X}} h(\underline{x}) \frac{f(\underline{x})}{u(\underline{x})} u(\underline{x}) \, d\underline{x}$$

$$= \int_{\mathcal{X}} \{h(\underline{x}) w(\underline{x})\} u(\underline{x}) \, d\underline{x}$$

$$= E_u \{ h(\underline{X}) \, \omega(\underline{X}) \}$$

where $w(\underline{x}) = \frac{f(x)}{u(x)}$. Now generate i.i.d. observations $\underline{X}_1, \underline{X}_2 \ldots$ from the density u and compute $\overline{h\omega_m} = \frac{1}{m} \sum_{i=1}^{m} h(x_i) \, w(x_i)$

The sampling density u called the importance function.

Example Suppose $x_1, x_2, \cdots x_n$ are i.i.d $N(\theta, \sigma^2)$, where θ and σ^2 are unknown. Independent priors are assumed for θ and σ^2, where θ has a double exponential distribution with density $\exp(-|\theta|)/2$ and σ^2 has the prior density of $(1+\sigma^2)^{-2}$.

Neither of these is a standard prior, but robust choice of proper prior all the same. If the posterior mean of θ is of interest, then it is necessary to compute,

$$E^{\pi}(\theta|x) = \int_{-\infty}^{\infty} \int_{0}^{\infty} \theta\, \pi(\theta, \sigma^2 | \underset{\sim}{x})\, d\theta\, d\sigma^2$$

Because $\pi(\theta, \sigma^2 | \underset{\sim}{x})$ is not a standard density, let us look for a standard density close to it.

letting $\bar{x}$ denote the mean of the sample $x_1, x_2, \cdots, x_n$ and $S_n^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n$,

note that, $\pi(\theta, \sigma^2 | \underset{\sim}{x}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sigma}{2\sigma^2}\{(\theta-\bar{x})^2 + s_n^2\}\right) \times$

$$\exp(-|\theta|)(1+\sigma^2)^{-2}$$

$$= \left[s_n^2 + (\theta-\bar{x})^2\right]^{\frac{n}{2}+1} (\sigma^2)^{-(\frac{n}{2}+2)} \exp\left(-\frac{n}{2\sigma^2}\{(\theta-\bar{x})^2 + s_n^2\}\right)$$

$$\times \left\{\left[s_n^2 + (\theta-\bar{x})^2\right]^{-(\frac{n}{2}+1)} \exp(-|\theta|) \left(\frac{\sigma^2}{1+\sigma^2}\right)^2\right.$$

$\therefore \pi(\theta, \sigma^2 | \underset{\sim}{x}) \propto u_1(\sigma^2|\theta)\, u_2(\theta) \exp(-|\theta|) \left(\frac{\sigma^2}{1+\sigma^2}\right)^2$

where $u_1(\sigma^2|\theta)$ is the density of inverse Gamma with shape parameter $\frac{n}{2}+1$ and scale parameter $\frac{n}{2}\left\{(\theta-\bar{x})^2+s_n^2\right\}$ and u_2 is the student's t density with d.f. $n+1$, location $\bar{x}$ and scale a multiple of s_n. It may be noted that the tails of $\exp(-|\theta|)\left(\frac{\sigma^2}{1+\sigma^2}\right)^2$ do not have much of an influence in the presence of $u_1(\sigma^2/\theta)\,u_2(\theta)$. Therefore, $u(\theta,\sigma^2) = u_1(\sigma^2/\theta)\,u_2(\theta)$ may be chosen as a suitable importance function. This involves sampling θ first from the density $u_2(\theta)$ and given this θ, sampling σ^2 from $u_1(\sigma^2/\theta)$. This is repeated to generate further value of (θ,σ^2). Finally, after generating m of these pairs (θ,σ^2), the required posterior mean of θ is approximated by

$$\widehat{E^{\pi}(\theta\,|\,\underset{\sim}{x})} = \frac{\sum_{i=1}^{m} \theta_i\, w(\theta_i,\sigma_i^2)}{\sum_{i=1}^{m} w(\theta_i,\sigma_i^2)}$$

where, $w(\theta,\sigma^2) = f(\underset{\sim}{x}\,|\,\theta,\sigma^2)\,\pi(\theta,\sigma^2)/u(\theta,\sigma^2)$

In some high-dimensional problems, a combination of numerical integration, Laplace approximation and Monte Carlo sampling seem to give appealing results.

Markov Chain Monte Carlo Methods

A severe drawback of the standard Monte Carlo sampling or Monte Carlo importance sampling is that complete determination of the functional form of the posterior density is needed for their implimentaion. Situations where posterior distributions are imcompletely specified or are specified inderectly cannot be handled. One such instance is where the joint posterior distribution of the vector of parameters is specified in terms of several conditional and marginal distributions, but not directly. This actually covers a very large range of Bayesian analysis because a lot of Bayesian modeling is hierarchical so that the joint posterior is difficult to calculate but the conditional posteriors given parameters at different levels of hierarchy are easier to write down and hence sample from.

For instance, consider the normal - Cauchy problem (example 1). ~~As shown~~ This problem can be given a hierarchical structure wherein we have the normal model, the conjugate normal prior in the first stage with a hyperparameter for its variance and this hyperparameter again has the conjugate prior.

Similarly, consider example 2 where we have independent observations $X_i \sim$ Poisson (θ_i). Now suppose the prior on the θ_i's is a conjugate mixture. We again see that a hierarchical prior stucture can lead to analytically tractable conditional posterior. It turns out that it is indeed possible in such cases to adopt an irerative Monte Carlo sampling Scheme, Which at the point of convergence will guarantee a random draw from the target joint posterior distribution. These iterative Monte Carlo procedures typically generate a random sequence with the Markov property such that this Markov chain is ergodic with the limiting distribution being the target posterior distribution. This is actually a whole class of such iterative procedures collectively called Markov chain Monte Carlo (MCMC) ~~produ~~ procedures. Different procedure from this class are suitable for different situations.

As mentioned above, convergence of a random sequence with the Markov property is being utilized in this procedure, and hence some basic understanding of Markov chains is required.

Law of large numbers for Markov chains

Let $\{X_n\}_{n\geq 0}$ be a Markov chain with a countable state space S and a transition probability matrix P. Further, suppose it is irreducible and has a stationary probability distribution $\pi = (\pi_i : i \in S)$. Then, for any bounded function $h : S \rightarrow R$ and for any initial distribution of X_0

$$\frac{1}{n}\sum_{i=0}^{n-1} h(x_i) \rightarrow \sum_{j} h(j)\,\pi_j$$

in probability as $n \rightarrow \infty$.

A similar law of large numbers (LLN) holds when the state space S is not countable. The limit value will be the integral of h with respect to the stationary distribution π.

A sufficient condition for the validity of this LLN is that the Markov chain $\{X_n\}$ be Harris irreducible and have a stationary distribution π. Harris irreducibility is the notion of irreducibility in general case.

To see how this is useful, consider the following: Given a probability distribution π on a set S, and a function h on S, suppose it is desired to compute the "integral of h with respect to π", which reduces to $\sum_j h(j)\pi_j$ in the countable case. Look for an irreducible Markov Chain $\{X_n\}$ with state space S and stationary distribution π. Then, starting from some initial value X_0, run the Markov chain $\{X_j\}$ for a period of time, say, $0, 1, 2, \dots n-1$ and consider as an estimate

$$\mu_n = \frac{1}{n} \sum_{0}^{n-1} h(X_j).$$

By LLN, this estimate μ_n will be close to $\sum h(j)\pi_j$ for large n.

This technique is called Markov chain Monte Carlo (MCMC). For example, if one is interested in $\pi(A) = \sum_{j \in A} \pi_j$ for some $A \subset S$ then by LLN this reduces to $\pi_n(A) = \frac{1}{n} \sum_{0}^{n-1} I_A(X_j) \to \pi(A)$ in probability as $n \to \infty$, ~~for any initial distribution~~ of ~~X_0~~. Where $I_A(X_j) = 1$ if $X_j \in A$ and 0 otherwise.

An irreducible Markov chain $\{X_n\}$ with a countable state space S is called aperiodic if for some $i \in S$ the g.c.d $\{ n : p_{ii}^{(n)} > 0 \} = 1$. Then, in addition to the LLN, the following result on the convergence of $P(X_n = j)$ holds.

$$\sum_j | P(X_n = j) - \pi_j | \to 0 \quad \text{as} \quad n \to \infty,$$

for any initial distribution of X_0.

In the other words, for large n the probability distribution of X_n will be close to π. There exists a result similar to above. for the general state space case that asserts that under suitable Conditions, the probability distribution of X_n will be close to π as $n \to \infty$.

This suggests that instead of doing over run of length n, one could do N dependent runs each of length m so that $n = Nm$ and then from the ith run use only the m^{th} observation, say, $X_{m,i}$, and consider the estimate

$$\tilde{\mu}_{N,m} = \frac{1}{N} \sum_{i=1}^{N} h(X_{m,i})$$

Other variations exist as well.

Metropolis - Hasting Algorithm

Here, we discuss a very general MCMC method with wide applications. The idea here is not to directy simulate from the given target density (which may be computationaly very difficult) at all, but to simulate an easy Markov chain that has this target density as the density of its stationary distribution.

Let S be a finite or countable set. Let π be a probability distribution on S. We shall call π the target distribution. Let $Q \equiv ((q_{ij}))$ be a transition probability matrix such that for each i, it is computationally easy to generate a sample from the distribution $\{q_{ij} : j \in S\}$. Let us generate a Markov chain $\{X_n\}$ as follows. If $X_n = i$, first sample from the distribution $\{q_{ij} : j \in S\}$ and denote that observation Y_n. Then, choose X_{n+1} from the two values X_n and Y_n according to

$$P(X_{n+1} = Y_n \mid X_n, Y_n) = \rho(X_n, Y_n)$$

$$P(X_{n+1} = X_n \mid X_n, Y_n) = 1 - \rho(X_n, Y_n)$$

where the acceptance probability $\rho(\cdot,\cdot)$ is given by

$$\rho(i,j) = \min\left\{ \frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1 \right\} \quad \text{for all } (i,j) \text{ such that } \pi_i q_{ij} > 0.$$

Note that $\{X_n\}$ is a Markov chain with transition probability matrix $P = ((p_{ij}))$ given by

$$p_{ij} = \begin{cases} q_{ij}\,\rho_{ij} & j \neq i \\ 1 - \sum\limits_{k \neq i} p_{ik} & j = i \end{cases}$$

Q is called the "proposal transition probability" and ρ the "acceptance probability".

A significant feature of this transition mechanism P is that P and π satisfy

$$\pi_i\, p_{ij} = \pi_j\, p_{ji} \qquad \text{for } i, j .$$

This implies that for any j,

$$\sum_i \pi_i\, p_{ij} = \pi_j \sum_i p_{ji} = \pi_j$$

or, π is a stationary probability distribution for P.

Now assume that S is irreducible with respect to Q and $\pi_i > 0$ for all i in S. It can ~~be~~ then be shown that P is irreducible. and because it has a stationary distribution π, LLN is available. This algorithm is thus a very flexible and useful one. The choice of Q is subject only to the condition that S is irreducible with respect to Q. Clearly, it is no loss of generality to assume that $\pi_i > 0$ for all i in S.

A sufficient condition for the aperiodicity of P is that $p_{ii} > 0$ for some i or equivalently

$$\sum_{j \neq i} q_{ij}\, \rho_{ij} < 1.$$

A sufficient condition for this is that there exists a pair (i, j) such that $\pi_i q_{ij} > 0$ and $\pi_j q_{ji} < \pi_i q_{ij}$.

Recall that if P is aperiodic, then both the LLN and $\sum_j |P(X_n = j) - \pi_j| \to 0$ as $n\to\infty$ hold. If S is not finite or countable but is a continum and the target distribution $\pi(\cdot)$ has a density $p(\cdot)$, then one ~~prod~~ proceeds as follows: Let Q be a transition function such that for each x, $Q(x, \cdot)$ has a density $q(x, y)$. Then proceed as in the discrete case but set the acceptance probability $\rho(x, y)$ to be $\rho(x,y) = \min\left\{ \frac{p(y)\, q(y,x)}{p(x)\, q(x,y)}, 1 \right\}$ for all (x, y) such that $p(x)\, q(x,y) > 0$.

A particularly useful feature of the above algorithm is that it is enough to know $p(\cdot)$ upto a multiplicative constant. This assures us that in Bayesian applications it is not necessary to have the normalizing constant of the posterior density.

Gibbs Sampling

Most of the new problems that Bayesians are asked to solve are high-dimensional. Application areas such as micro-arrays and image processing are some examples. Bayesian analysis of such problems invariably involve target (posterior) distributions that are high-dimensional multivariate distributions. In image processing, for example, typically one has N×N grid of pixels with N = 256 and each pixel has $k \geq 2$ possible values. Thus each configuration has $(256)^2$ components and the state space S has $k^{(256)^2}$ configurations. To simulate a random configuration from a target distribution over such a large S is not an easy task. The Gibbs sampler is a technique especially suitable for generating an irreducible aperiodic Markov chain that has as its stationary distribution a target distribution in a high-dimensional space but having some special structure. The most interesting aspect of this technique is that to run this Markov chain, it suffices to generate observations from univariate distributions.

The Gibbs sampler in the context of a bivariate probability distribution can be described as follows. Let π be a target probability distribution of a bivariate random vector (X, Y). For each x, let $P(x, \cdot)$ be the conditional probability distribution of Y given $X = x$. Similarly, let $Q(y, \cdot)$ be the distribution X given $Y = y$.

Note that for each x, $P(x, \cdot)$ is a univariate distribution, and for each y, $Q(y, \cdot)$ is also a univariate distribution. Now generate a bivariate Markov chain $Z_n = (X_n, Y_n)$ as follows:

Start with some $X = x_0$. Generate an observation Y_0 from the distribution $P(x_0, \cdot)$. Then generate an observation X_1 from $Q(Y_0, \cdot)$. Next generate an Y_1 from $P(X_1, \cdot)$ and so on.

If π is a discrete distribution concentrated on $\{(x_i, y_i) : 1 \leq i \leq K, 1 \leq j \leq L\}$ and if $\pi_{ij} = \pi(x_i, y_j)$ then $P(x_i, y_j) = \frac{\pi_{ij}}{\pi_{i\cdot}}$ and

$$Q(y_j, x_i) = \frac{\pi_{ij}}{\pi_{\cdot j}},$$

Where $\pi_{i\cdot} = \sum_j \pi_{ij}$ and $\pi_{\cdot j} = \sum_i \pi_{ij}$

Thus the transition probability matrix $R = ((r_{(ij),(kl)}))$ for the $\{Z_n\}$ chain is given by $r_{(ij)(kl)} = Q(y_j, x_k)\,P(x_k, y_l)$

$$= \frac{\pi_{kj}}{\pi_{\cdot j}} \; \frac{\pi_{kl}}{\pi_{k\cdot}} .$$

It can be verified that this chain is irreducible, aperiodic, and has π as its stationary distribution. Thus LLN holds in this case.. Thus for large n, Z_n can be viewed as a sample from a distribution that is close to π and one can approximate $\sum_{i,j} h(i,j)\,\pi_{ij}$ by $\sum_{i=1}^{n} h(X_i, Y_i)/n$

Example Consider sampling from $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \left[\begin{matrix}1, \\ \rho \end{matrix}\right.\right.$

Note that the conditional distribution of X given $Y=y$ and that of Y given $X=x$ are

$$X \mid Y=y \sim N(\rho y, 1-\rho^2) \text{ and } Y \mid X=x \sim N(\rho x, 1-\rho^2)$$

Using this property, Gibbs sampling proceeds as described below to generate (X_n, Y_n), $n = 0, 1, 2 \ldots$ by starting from an arbitrary value x_0 for X_0 and repeating the following steps for $i=0,1,2 \ldots n$.

1. Given x_i for X, draw a random deviate from $N(\rho x_i, 1-\rho^2)$ and denote it by Y_i.

2. Given y_i for Y, draw a random deviate from $N(\rho y_i, 1-\rho^2)$ and denote it by X_{i+1}.

The theory of Gibbs sampling tells us that n is large, then (x_n, y_n) is a random draw from a distribution that is close to $N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$.

To see why Gibbs sampler works here, recall that 'a sufficient condition for the LLN and the limit result is that an appropriate irreducibility condition holds and a stationary distribution exists. From steps 1 and 2 above and using the conditional distributions $X|Y=y$ and $Y|X=x$, one has

$$Y_i = \rho X_i + \sqrt{1-\rho^2}\,\eta_i$$

and $$X_{i+1} = \rho Y_i + \sqrt{1-\rho^2}\,\xi_i,$$

Where η_i and ξ_i are independent standard normal random variables independent of X_i. Thus the sequence $\{X_i\}$ satisfies the stochastic difference equation $X_{i+1} = \rho^2 X_i + U_{i+1}$,

Where $U_{i+1} = \rho\sqrt{1-\rho^2}\,\eta_i + \sqrt{1-\rho^2}\,\xi_i$.

Because η_i, ξ_i are independent $N(0,1)$ random variables, U_{i+1} is also a normally distributed random variables with mean 0 and variance $\rho^2(1-\rho^2) + (1-\rho^2) = 1-\rho^4$. Also $\{U_i\}_{i\geq 1}$ being i.i.d. makes $\{X_i\}_{i\geq 0}$ a Markov chain. It turns out that the irreducibility condition holds here. Also the standard $N(0,1)$ distribution a stationary distribution for $\{X_n\}$.