13103057 Bayesian Statistics 1

UNDERGRADUATE PROJECT REPORT

Bayesian Statistics

Project By:
Siddhant Garg
Roll No: 150711
Department of Maths
and Statistics

Supervised By :
Prof. Debasis Kundu
Department of Maths
and Statistics

Contents

Abstract

This report presents the various concepts and implementation used in Bayesian analysis. It first introduces the basic Bayesian procedures to obtain posterior distribution of the parameters, given the data and the priors using the Bayes theorem and then obtain their Bayes estimates using different loss functions. Bayes estimates can be difficult to compute analytically in case of high dimensional data. So, we also talk about more efficient methods to obtain the estimates of the posterior distribution like EM-Algorithm, Monte Carlo Sampling, Metropolis Hastings Algorithm and Gibbs Sampling. This report also contains a section on Hypothesis Testing and model selection, which includes about Bayes factors, BIC, P-value, Bayesian Outlier Detection. To give an illustration, we also include a statistical inference on one-dimensional, 3-parameter Geometric Skew Normal Distribution. It is difficult to obtain the Bayes estimates from the GSN distribution analytically, so we use **Metropolis-Hastings Algorithm** to sample and obtain the Bayes estimates. Further, we present the results and conclusions.

1 Bayesian Procedures [4]

1.1 Prior and Posterior Distributions

Consider a random variable X that has a distribution of probability that depends upon the symbol θ, where θ is an element of a well-defined set Ω. Let us now introduce a random variable Θ that has a distribution of probability over the set Ω. We now look upon θ as a possible value of the random variable Θ.

$$X|\Theta \sim f(x|\theta)$$
$$\Theta \sim h(\theta)$$

The pdf $h(\theta)$ is called the **prior** pdf of θ. Moreover, we now denote the pdf of X by $f(x|\theta)$ since we think of it as a conditional pdf of X, given $\Theta = \theta$.

Suppose that $X_1, \ldots, X_n$ is a random sample from the conditional distribution of $\mathbf{X}$ given $\Theta = \theta$ with pdf $f(x|\theta)$. Thus we can write the joint conditional pdf of $\mathbf{X}$, given $\Theta = \theta$, as

$$L(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta)$$

Thus the joint pdf of $\mathbf{X}$ and Θ is

$$g(\mathbf{x}, \theta) = L(\mathbf{x}|\theta)h(\theta)$$

If Θ is a random variable of the continuous type, the joint marginal pdf of $\mathbf{X}$ is given by

$$g_1(\mathbf{x}) = \int_{-\infty}^{\infty} g(\mathbf{x}, \theta)d\theta$$

The conditional pdf of Θ, given the sample $\mathbf{X}$, is

$$k(\theta|\mathbf{x}) = \frac{g(\mathbf{x}, \theta)}{g_1(\mathbf{x})} = \frac{L(\mathbf{x}|\theta)h(\theta)}{g_1(\mathbf{x})}$$

The above pdf is called **posterior pdf**. The prior distribution reflects the subjective belief of Θ before the sample is drawn, while the posterior distribution is the conditional distribution of Θ after the sample is drawn.

1.2 Bayesian Point Estimation

Suppose we want a point estimator of θ. From the Bayesian viewpoint, this really amounts to selecting a decision function δ, so that $\delta(x)$ is a predicted value of θ when both the computed value $\mathbf{x}$ and the conditional pdf $k(\theta|\mathbf{x})$ are known.

The choice of the decision function should depend upon a loss function $L[\theta, \delta(\mathbf{x})]$. A Bayes estimate is a decision function δ that minimizes

$$E\{L[\theta, \delta(x)]|\mathbf{X} = \mathbf{x}\} = \int_{-\infty}^{\infty} L[\theta, \delta(\mathbf{x})]k(\theta|\mathbf{x})d\theta$$

If Θ is a random variable of the continuous type. That is,

$$\delta(\mathbf{X}) = Argmin \int_{-\infty}^{\infty} L[\theta, \delta(\mathbf{x})]k(\theta|\mathbf{x})d\theta$$

If $L[\theta, \delta(\mathbf{x})] = [\theta - \delta(\mathbf{x})]^2$, then $\delta(\mathbf{x}) = \mathbf{E}(\Theta|\mathbf{x})$

If $L[\theta, \delta(\mathbf{x})] = |\theta - \delta(\mathbf{x})|$, then median of the conditional distribution of Θ given $\mathbf{X} = \mathbf{x}$ is the Bayes Solution.

1.3 Bayesian Interval Estimation

1.3.1 Confidence Interval

Theorem 1.1 (Central Limit Theorem) *Let $X_1, \ldots, X_n$ denote the observations of a random sample from a distribution that has mean μ and finite variance σ^2. Then the distribution function of the random variable $W_n = (\bar{X} - \mu)/(\sigma/n)$ converges to Φ, the distribution function of the $N(0, 1)$ distribution, as $n \to \infty$.*

Large Sample Confidence Interval for mean μ

Suppose $X_1, \ldots, X_n$ is a random sample on a random variable X with mean μ and variance σ^2.

$$Z_n = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Distribution of Z_n is approximately $N(0, 1)$.

Let α be such that $\alpha/2 = P(Z_n > z_{\alpha/2})$.

$$1 - \alpha \approx P\Big(- z_{\alpha/2} < Z_n < z_{\alpha/2}\Big)$$

$$1 - \alpha \approx P\Big(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < Z_n < \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\Big)$$

Again, letting $\bar{x}$ and s denote the realized values of the statistics $\bar{X}$ and S, respectively, after the sample is drawn, an approximate $(1 - \alpha)100\%$ confidence interval for μ is given by,

$$(\bar{x} - z_{\alpha/2}s\sqrt{n}, \bar{x} + z_{\alpha/2}s\sqrt{n})$$

This is called a **large sample** confidence interval for μ.

If an interval estimate of θ is desired, we can find two functions $u(\mathbf{x})$ and $v(\mathbf{x})$ so that the conditional probability

$$P[u(\mathbf{x}) < \Theta < v(\mathbf{x})|\mathbf{X} = \mathbf{x}] = \int_{u(\mathbf{x})}^{v(\mathbf{x})} k(\theta|\mathbf{x})$$

These intervals are often called **credible** or **probability intervals**, so as not to confuse them with confidence intervals.

2 Bayesian Computations [1]

Bayesian analysis requires 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, quantiles 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 form. Thus special techniques are needed for Bayesian computations.

2.1 The E-M Algorithm

Suppose $Y|\theta$ has density $f(y|\theta)$, and suppose the prior on θ is $\pi(\theta)$, resulting in the posterior density $\pi(\theta|y)$. When $\pi(\theta|y)$ is computationally difficult to handle, as is usually the case, there are some 'data augmentation' methods that can help. The idea is to augment the observed data y with missing or latent data z to obtain the 'complete' data x = (y, z) so that the augmented posterior density $\pi(\theta|x) = \pi(\theta|y, z)$ is computationally easy to handle.

The basic steps in the iterations of the E-M algorithm are the following. Let $p(z|y, \theta)$ be the predictive density of Z given y and an estimate $\hat{\theta}$ of θ.

Find $z^{(i)} = E(Z|y, \hat{\theta}^{(i)})$, where $\hat{\theta}^{(i)}$ is the estimate of θ used in the i^{th} step of the iteration.

Use $z^{(i)}$ to augment y and maximize $\pi(\theta|y, z^{(i)})$ to obtain $\hat{\theta}^{(i+1)}$. Then find $z^{(i+1)}$ using $\hat{\theta}^{(i+1)}$ and continue this iteration.

Implementation of the E-M Algorithm

$$\pi(\theta|y) = \frac{\pi(\theta, z|y)}{p(z|y, \theta)}$$

$$log[\pi(\theta|y)] = log[\pi(\theta, z|y)] - log[p(z|y, \theta)]$$

Taking expectations with respect to $Z|\hat{\theta}^{(i)}, y$

$$\begin{aligned} \mathbf{E}\{log[\pi(\theta|y)]\} &= \int log[\pi(\theta, z|y)p(z|y, \hat{\theta}^{(i)}]dz - \int log[p(z|y, \theta)p(z|y, \hat{\theta}^{(i)})]dz \\ &= Q(\theta, \hat{\theta}^{(i)}) - H(\theta, \hat{\theta}^{(i)}) \end{aligned} \tag{1}$$

Then, the general E-M algorithm involves the following two steps in the i^{th} iteration:

E-Step: Calculate $Q(\theta, \hat{\theta}^{(i)})$;

M-Step: Maximize $Q(\theta, \hat{\theta}^{(i)})$ with respect to θ and obtain $\hat{\theta}^{(i+1)}$ such that

$$\underset{\theta}{max} Q(\theta, \hat{\theta}^{(i)}) = Q(\hat{\theta}^{(i+1)}, \hat{\theta}^{(i)})$$

Note that

$$log\pi(\hat{\theta}^{(i+1)}|y) - log\pi(\hat{\theta}^{(i)}|y) = \Big\{Q(\hat{\theta}^{(i+1)}, \hat{\theta}^{(i)}) - Q(\hat{\theta}^{(i)}, \hat{\theta}^{(i)})\Big\} - \Big\{H(\hat{\theta}^{(i+1)}, \hat{\theta}^{(i)}) - H(\hat{\theta}^{(i)}, \hat{\theta}^{(i)})\Big\}$$

From the E-M Algorithm, we have $Q(\hat{\theta}^{(i+1)}, \hat{\theta}^{(i)}) \geq Q(\hat{\theta}^{(i)}, \hat{\theta}^{(i)})$. Further, for any θ,

$$\begin{aligned} H(\theta, \hat{\theta}^{(i)}) - H(\hat{\theta}^{(i)}, \hat{\theta}^{(i)}) &= \int log[p(z|y,\theta)p(z|y,\hat{\theta}^{(i)})]dz - \int log[p(z|y,\hat{\theta}^{(i)})p(z|y,\hat{\theta}^{(i)})]dz \\ &= \int log\left[\frac{p(z|y,\theta)}{p(z|y,\hat{\theta}^{(i)})}\right]p(z|y,\hat{\theta}^{(i)})dz \\ &= -\int log\left[\frac{p(z|y,\hat{\theta}^{(i)})}{p(z|y,\theta)}\right]p(z|y,\hat{\theta}^{(i)})dz \\ &\leq 0 \end{aligned} \tag{2}$$

because for any 2 densities p_1 and p_2, $\int log(p_1(x)/p_2(x))p_1(x)dx$ is the **Kullback-Leibler** distance between p_1 and p_2, which is atleast 0. Therefore,

$$H(\hat{\theta}^{(i+1)}, \hat{\theta}^{(i)}) - H(\hat{\theta}^{(i)}, \hat{\theta}^{(i)}) \leq 0$$

and hence,

$$\pi(\hat{\theta}^{(i+1)}|y) \geq \pi(\hat{\theta}^{(i)}|y)$$

for any iteration i. Therefore, starting from any point, the E-M algorithm can usually be expected to converge to a local maximum.

2.2 Monte Carlo Sampling

Consider an expectation that is not available in closed form. An alternative to numerical integration or analytic approximation to compute this is statistical sampling. This probabilistic technique is a familiar tool in statistical inference. 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 or the sample proportion. The law of large numbers guarantees that the estimates so obtained will be good provided the sample is large enough.

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

$$E[h(X)] = \int h(x)f(x)dx$$

If i.i.d. observations $X_1, \ldots, X_n$ can be generated from the density f, then

$$\bar{h}_m = \frac{1}{m}\sum_{i=1}^{m} h(X_i) \xrightarrow{p} E[h(X)]$$

Monte Carlo Importance Sampling

Suppose that it is difficult or expensive to sample directly from j, but there exists a probability density u that is very close to f from which it is easy to sample. Then,

$$\begin{aligned} E[h(\mathbf{X})] &= \int h(x)f(x)dx \\ &= \int h(x)\frac{f(x)}{u(x)}u(x)dx \\ &= \int \{h(x)w(x)\}u(x)dx \\ &= E_u[h(\mathbf{X})w(\mathbf{X})] \end{aligned} \tag{3}$$

Now generate $X_1, \ldots, X_m$ from density u and compute

$$\bar{h}_m = \frac{1}{m} \sum_{i=1}^{m} h(\mathbf{X}_i) w(\mathbf{X}_i)$$

The sampling density u is called the **importance function**.

2.3 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 implementation. Situations where posterior distributions are incompletely specified or are specified indirectly 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.

Convergence of a random sequence with the Markov property is being utilized in this procedure.

Theorem : (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 $\pi \equiv (\pi_i : i \in S)$. Then, for any bounded function $h : S \to R$ and for any initial distribution of X_0

$$\frac{1}{n} \sum_{i=0}^{n} h(X_i) \xrightarrow{p} \sum_{j} h(j) \pi_j$$

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.

2.3.1 Metropolis-Hastings Algorithm

The idea here is not to directly simulate from the given target density (which may be computationally 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

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

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

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

$$\rho(i,j) = min\Big\{1, \frac{\pi_i}{\pi_j} \frac{q_{ji}}{q_{ij}}\Big\}$$

for all (i,j) such that $\pi_i q_{ij} > 0$. Note that $\{X_n\}$ is a Markov Chain with transition probability matrix $P = ((p_{ij}))$ given by

$$p_{ij} = \left\{ \begin{array}{ll} q_{ij}\rho_{ij} & j \neq i \\ 1 - \sum_{k \neq i} p_{ik}, & j = i \end{array} \right\}$$

Q is called the "propositional transition probability" and ρ the "acceptance probability". A Markov process is uniquely defined by its transition probabilities, p_{ij} the probability of transitioning from any given state, i, to any other given state, j. It has a unique stationary distribution $\pi(x)$ when the following two conditions are met:

1. For every pair of states i, j, the probability of being in state i and transitioning to state j must be equal to the probability of being in state j and transitioning to state i, $\pi_i p_{ij} = \pi_j p_{ji}$.

2. The stationary distribution π_i must be unique.

We have,

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

$$\frac{p_{ji}}{p_{ij}} = \frac{\pi_j}{\pi_i}$$

The approach is to separate the transition in two sub-steps; the proposal and the acceptance-rejection. The proposal distribution q_{ij} is the conditional probability of proposing a state j given i, and the acceptance distribution ρ_{ij} the conditional probability to accept the proposed state j. The transition probability can be written as the product of them:

$$p_{ij} = q_{ij}\rho_{ij}$$

Inserting this relation in the previous equation, we have

$$\frac{\rho_{ij}}{\rho_{ji}} = \frac{\pi_j}{\pi_i}\frac{q_{ji}}{q_{ij}}$$

The next step in the derivation is to choose an acceptance that fulfills the condition above. One common choice is the Metropolis choice:

$$\rho(i,j) = min\Big\{1, \frac{\pi_i}{\pi_j}\frac{q_{ji}}{q_{ij}}\Big\}$$

i.e., we always accept when the acceptance is bigger than 1, and we reject accordingly when the acceptance is smaller than 1. This is the required quantity for the algorithm.

2.3.2 Gibbs Sampling

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.

Suppose we have a joint density $f(x, y_1 \dots y_k)$ and we are interested in some feature of $f(x)$ (like E(X))

The Gibbs Algorithm for computing this expectation

Assume we can sample the k+1-many univariate conditional densities:

$f(X|y_1, \dots, y_k)$

$f(Y_1|x, y_2, \dots, y_k)$

$f(Y_2|x, y_1, y_3, \dots, y_k)$

...

$f(Y_k|x, y_1, y_3, \dots, y_{k-1})$

Choose, arbitrarily, k initial values: $Y_1 = y_1^0, Y_2 = y_2^0, \ldots, Y_k = y_0^k$.

Draw samples:

$$x^1 \sim f(X|y_1, \ldots, y_k)$$

$$y_1^1 \sim f(Y_1|x, y_2, \ldots, y_k)$$

$$\ldots$$

$$y_k^1 \sim f(Y_k|x, y_1, y_3, \ldots, y_{k-1})$$

This constitutes one Gibbs "pass" through the k+1 conditional distributions, yielding values: $(x^1, y_1^1, \ldots, y_k^1)$

Iterate the sampling to form the i^{th} "pass" $(x^2, y_1^i, \ldots, y_k^i)$ and so on

As $i \to \infty$, $x^i \sim f(X)$.

Theorem: (Hammersley-Clifford) Under the positivity condition, the joint density p satisfies

$$p(y_1, \ldots, y_k) \propto \prod_{j=1}^{k} \frac{p_j(y_j|y_1, \ldots, y_{j-1}, y'_{j+1}, \ldots, y'_k)}{p_j(y'_j|y_1, \ldots, y_{j-1}, y'_{j+1}, \ldots, y'_k}$$

for every **y** and **y'** in the support of p.

2.4 More Bayesian Methods

2.4.1 Hierarchical Bayes

The prior pdf has an important influence in Bayesian inference. One way of having more control over the prior is to model the prior in terms of another random variable. This is called the **hierarchical Bayes** model, and it is of the form

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

$$\Theta|\gamma \sim h(\theta|\gamma)$$

$$\Gamma \sim \psi(\gamma)$$

With this model we can exert control over the prior $h(\theta|\gamma)$ by modifying the pdf of the random variable Γ.

The parameter γ can be thought of a nuisance parameter. It is often called a **hyperparameter**. As with regular Bayes, the inference focuses on the parameter θ; hence, the posterior pdf of interest remains the conditional pdf $k(\theta|\mathbf{x})$.

$$\begin{aligned} g(\theta, \gamma|\mathbf{x}) &= \frac{g(\mathbf{x}, \theta, \gamma)}{g(\mathbf{x})} \\ &= \frac{g(\mathbf{x}|\theta, \gamma)g(\theta, \gamma)}{g(\mathbf{x})} \\ &= \frac{f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)}{g(\mathbf{x})} \end{aligned} \tag{4}$$

Therefore the posterior pdf is given by,

$$k(\theta|\mathbf{x}) = \frac{\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)d\theta}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)d\theta d\psi}$$

Furthermore, assuming squared-error loss, the Bayes estimate of $W(\theta)$ is

$$\delta_W(\mathbf{x}) = \frac{\int_{\infty}^{\infty}\int_{-\infty}^{\infty} W(\theta)f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)d\theta}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)d\theta d\psi}$$

To obtain the Bayes Estimate of $W(\theta)$, we refer to Gibbs Sampler Algorithm. For $i = 1, 2, \ldots, m$, at the i^{th} step

$$\Theta_i|\mathbf{x}, \gamma_{i-1} \sim g(\theta|\mathbf{x}, \gamma_{i-1})$$

$$\Gamma_i|\mathbf{x}, \theta_i \sim g(\gamma|\mathbf{x}, \theta_i)$$

as $i \to \infty$

$$\Theta_i \xrightarrow{\text{D}} k(\theta|\mathbf{x})$$

$$\Gamma_i \xrightarrow{\text{D}} g(\gamma|\mathbf{x})$$

Furthermore the arithmetic average

$$\frac{1}{m-n}\sum_{i=n+1}^{m} W(\Theta_i) \xrightarrow{\text{P}} E[W(\Theta|\mathbf{x})] = \delta_w(\mathbf{x}) \text{ as } m \to \infty$$

Because of the **Monte Carlo** generation these procedures are often called **MCMC**, for **Markov Chain Monte Carlo** procedures.

2.4.2 Emperical Bayes

The empirical Bayes model consists of the first two lines of the hierarchical Bayes model; i.e.,

$$\mathbf{X}|\Theta \sim f(\mathbf{x}|\theta)$$

$$\Theta|\gamma \sim h(\theta|\gamma)$$

Instead of attempting to model the parameter γ with a pdf as in hierarchical Bayes, empirical Bayes methodology estimates γ based on the data as follows.

$$\begin{aligned} g(\mathbf{x}, \theta|\gamma) &= \frac{f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)}{\psi(\gamma)} \\ &= f(\mathbf{x}|\theta)h(\theta|\gamma) \end{aligned} \tag{5}$$

Consider, then, the likelihood function

$$m(\mathbf{x}|\gamma) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)d\theta$$

Using the pdf $m(\mathbf{x}|\gamma)$, we obtain an estimate $\widehat{\gamma} = \widehat{\gamma(\mathbf{x})}$, usually by the method of maximum likelihood. For inference on the parameter θ, the empirical Bayes procedure uses the posterior pdf $k(\theta|\mathbf{x}, \widehat{\gamma})$.

3 Hypothesis Testing and Model Selection

For Bayesians, model selection and model criticism are extremely important inference problems. Sometimes these tend to become much more complicated than estimation problems.

3.1 Testing and Bayes Factor

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

$$M_0 : X \text{ has density } f(x|\theta) \text{ where } \theta \in \Theta_0$$

$$M_1 : X \text{ has density } f(x|\theta) \text{ where } \theta \in \Theta_1$$

We want to test

$$M_0 : \theta \in \Theta \text{ versus } M_1 : \theta \in \Theta$$

Let π_0 and $1 - \pi_0$ be the prior probabilities 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 in the previous approach is nothing but

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

We can calculate the posterior probabilities and posterior odds ratio namely,

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

The Bayes rule for 0-1 loss is to choose the hypothesis with higher posterior probability.

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

$$\begin{aligned} m_\pi(x) &= \int_\Theta f(x|\theta)\pi(\theta) d\theta \\ &= \pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta) d\theta + (1 - \pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta) d\theta \end{aligned}$$

and hence the posterior density of $\theta|X = x$ as

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

It follows then that

$$P(M_0|x) = P^\pi(\Theta_0|x) = \frac{\pi_0}{m_\pi(x)} \int_{\Theta_0} f(x|\theta) g_0(\theta)$$

$$P(M_0|x) = P^\pi(\Theta_0|x) = \frac{1 - \pi_1}{m_\pi(x)} \int_{\Theta_1} f(x|\theta) g_1(\theta)$$

Then, to compare models M_0 and M_1 on the basis of a random sample $x = (x_1, \ldots, , x_n)$ one would use the Bayes factor

$$BF_{01} = \frac{m_0(x)}{m_1(x)}$$

where

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

The posterior odds ratio of M_0 relative to M_1 is

$$\Big(\frac{\pi_0}{1-\pi_0}\Big) BF_{01}$$

$$\begin{aligned} P(M_0|x) &= \frac{\pi_0 m_0(x)}{\pi_0 m_0(x) + (1-\pi_0) m_1(x)} \\ &= \Big\{1 + \frac{1-\pi_0}{\pi_0} BF_{01}^{-1}\Big\}^{-1} \end{aligned}$$

BF_{01} is an important evidential measure that is free of π_0. The smaller the value of BF_{01}, the stronger the evidence against M_0

Testing a Point Null Hypothesis

The problem is to test:

$$M_0 : \theta = \theta_0 \text{ versus } M_1 : \theta \neq \theta_0$$

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. Intuitively, this is clear: if the null hypothesis is a priori impossible, it will remain so a posteriori also.

Therefore, a prior probability of $\pi_0 > 0$ needs to be assigned to 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 g_0 to be a point mass at θ_0

Now the prior π is of form

$$\pi(\theta) = \pi_0 I\{\theta = \theta_0\} + (1-\pi_0) g_1(\theta) I\{\theta \neq \theta_0\}$$

3.2 Bayesian Information Criterion

In statistics, the Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function, and it is closely related to **Akaike information criterion** (AIC).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC.

The BIC is an asymptotic result derived under the assumptions that the data distribution is in the exponential family. The formula for BIC is

$$-2 \ln f(x|k) \approx \text{ BIC } = -2. \ln(MLE) + k. \ln(n)$$

k = number of free parameters to be determined

n = sample size

Given any two estimated models, the model with the lower value of BIC is the one to be preferred.

$$m_i(x) \approx f(x|\widehat{\theta}_i) g_i(\widehat{\theta}_i)(2\pi)^{p_i/2} n_{-p_i/2} |H^{-1}_{1,\widehat{\theta}_i}|^{1/2}$$

where $\widehat{\theta}_i$ is the maximum likelihood estimate for model M_i.

$$2\log B_{01} \approx 2\log\Big(\frac{f(x|\widehat{\theta}_0)}{f(x|\widehat{\theta}_1)}\Big) + 2\log\Big(\frac{g_0(\widehat{\theta}_0)}{g_1(\widehat{\theta}_1)}\Big) - (p_0 - p_1)\log\frac{n}{2\pi} + \log\Big(\frac{|H^{-1}_{1,\widehat{\theta}_0}|}{|H^{-1}_{1,\widehat{\theta}_1}|}\Big)$$

$$2\log B_{01} \approx 2\log\Big(\frac{f(x|\widehat{\theta}_0)}{f(x|\widehat{\theta}_1)}\Big) - (p_0 - p_1)\log n$$

This is the approximate Bayes factor based on the Bayesian information crite- rion (BIC) due to Schwarz (1978). The term $(p_0 - p_1)\log n$ can be considered a penalty for using a more complex model.

A related criterion is

$$2\log\Big(\frac{f(x|\widehat{\theta}_0)}{f(x|\widehat{\theta}_1)}\Big) - (p_0 - p_1)$$

which is based on the **Akaike information criterion (AIC)**, namely,

$$\text{AIC } = 2\log f(x|\widehat{\theta}) - 2p$$

for a model $f(x|\theta)$. The penalty for using a complex model is not as drastic as that in BIC.

3.3 P-Value and Posterior Probabilities of H_0 as Measures 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 also hap- pens to be one of the concepts that is highly misunderstood and misused.

P-Value is the probability under a (simple) 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:

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

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 data $X = x$ is observed, with corresponding $t = T(x)$, the P-value then is

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

To a Bayesian the posterior probability of H_0 summarizes the evidence against H_0. In many of the common cases of testing, the P-value is smaller than the posterior probability by an order of magnitude. The reason for this is that the P-value ignores the likelihood of the data under the alternative and takes into account not only the observed deviation of the data from the null hypothesis as measured by the test statistic but also more significant deviations.

3.4 Bounds on Bayes Factors and Posterior Probabilities

There are irreconcilable differences between the classical P-value and the corresponding Bayesian measures of evidence in many cases. However, one may argue that the differences are perhaps due to the choice of π_0 or g_1 that cannot claim to be really 'objective.' The choice of $\pi_0 = 1/2$ may not be crucial because the Bayes factor, B, which does not need this, seems to be providing the same conclusion, but the choice of g_1 does have substantial effect. To counter this argument, let us consider lower bounds on B and P over wide classes of prior densities. What is surprising is that even these lower bounds that are based on priors 'least favorable' to H 0 are typically an order of magnitude larger than the corresponding P-values for precise null hypotheses.

Thus, in the case of precise null hypotheses, if G is the class of all plausible conditional prior densities g_1 under H_0 , we are then lead to the consideration of the following bounds.

$$B(G, x) = \inf_{g \in G} B_{01} = \frac{f(x|\theta_0)}{\sup_{g \in G} m_g(x)}$$

where $m_g(x) = \int_{\theta \neq \theta_0} f(x|\theta) g(\theta) d\theta$, and

$$P(H_0|G, x) = \inf_{g \in G} P(H_0|x) = \left[1 + \frac{1 - \pi_0}{\pi_0} B(G, x)^{-1}\right]^{-1}$$

Based on evidence the least possible Bayes factor and posterior probability of H_0 are substantially larger than the corresponding P-value.

3.5 Robust Bayesian Outlier Detection

Because a Bayes factor is a weighted likelihood ratio, it can also be used for checking whether an observation should be considered an outlier with respect to a certain target model relative to an alternative model.

X having density $f(x|\theta)$ is observed, and it is of interest to compare two models M_0 and M_1 given by

$$M_0 : X \text{ has density } f(x|\theta) \text{ where } \theta \in \Theta_0$$

$$M_1 : X \text{ has density } f(x|\theta) \text{ where } \theta \in \Theta_1$$

For $i = 1, 2, g_i(\theta)$ is the prior density of θ, conditional on M_i, being the true model. To compare M_0 and M_1 on the basis of a random sample $x = (x_1, \ldots, x_n)$ the Bayes factor is given by

$$B_{01}(x) = \frac{m_0(x)}{m_1(x)}$$

To measure the effect on the Bayes factor of observation x_d one could use the quantity

$$k_d = \log\left(\frac{B(x)}{B(x_{-d})}\right)$$

where $B(x_{-d})$ is the Bayes factor excluding observation x_d. If $k_d < 0$, then when observation x_d is deleted there is an increase of evidence for M_0. Consequently, observation x_d itself favors model M_1 . The extent to which x_d favors M_1 determines whether it can be considered an outlier under model M_0. Similarly, a positive value for k_d implies that x_d favors M_0.

Because k_d, derived from the Bayes factor, is the Bayesian quantity of inferential interest here, upper and lower bounds on k_d over classes of prior densities are required.

We shall illustrate this approach with a precise null hypothesis. Then we have the problem of comparing

$$M_0 : \theta = \theta_0 \text{ versus } \theta \neq \theta_0$$

using a random sample from a population with density $f(x|\theta)$. Under M_1, suppose θ has the prior density g, $g \in \Gamma$. The Bayes factors with all the observations and without the d^th observation, respectively, are

$$B_g(x) = \frac{f(x|\theta)}{\int_{\theta \neq \theta_0} f(x|\theta)g(\theta)d\theta}$$

$$B_g(x_{-d}) = \frac{f(x_{-d}|\theta)}{\int_{\theta \neq \theta_0} f(x_{-d}|\theta)g(\theta)d\theta}$$

Because $f(x|\theta) = f(x_{-d}|\theta)f(x_d|\theta)$, we get

$$\begin{aligned} k_{d,g} &= \log\Big[\frac{f(x|\theta_0)}{f(x_{-d}|\theta_0)}\frac{\int_{\theta \neq \theta_0} f(x_{-d}|\theta)g(\theta)d\theta}{\int_{\theta \neq \theta_0} f(x|\theta)g(\theta)d\theta}\Big] \\ &= \log f(x_d|\theta_0) - \log\Big[\frac{\int_{\theta \neq \theta_0} f(x|\theta)g(\theta)d\theta}{\int_{\theta \neq \theta_0} f(x_{-d}|\theta)g(\theta)d\theta}\Big] \end{aligned}$$

Now note that to find the extreme values of $k_{d,g}$, it is enough to find the extreme values of

$$h_{d,g} = \frac{\int_{\theta \neq \theta_0} f(x|\theta)g(\theta)d\theta}{\int_{\theta \neq \theta_0} f(x_{-d}|\theta)g(\theta)d\theta}$$

over the set Γ. Further, this optimization problem can be rewritten as follows:

$$\begin{aligned} \sup_{g \in G} h_{g,d} &= \sup_{g \in G} \frac{\int_{\theta \neq \theta_0} f(x_d|\theta)f(x_{-d}|\theta)g(\theta)d\theta}{\int_{\theta \neq \theta_0} f(x_{-d}|\theta)g(\theta)d\theta} \\ &= \sup_{g^* \in G^*} \int_{\theta \neq \theta_0} f(x_d|\theta)g^*(\theta)d\theta \end{aligned}$$

$$\begin{aligned} \inf_{g \in G} h_{g,d} &= \inf_{g \in G} \frac{\int_{\theta \neq \theta_0} f(x_d|\theta)f(x_{-d}|\theta)g(\theta)d\theta}{\int_{\theta \neq \theta_0} f(x_{-d}|\theta)g(\theta)d\theta} \\ &= \inf_{g^* \in G^*} \int_{\theta \neq \theta_0} f(x_d|\theta)g^*(\theta)d\theta \end{aligned}$$

where

$$G^* = \left\{ g^* : g^*(\theta) = \frac{g(\theta)f(x_{-d}|\theta)}{\int_{u \neq \theta_0} g(u)f(x_{-d}|\theta)du} \right\}$$

3.6 Nonsubjective Bayes Factors

When subjective specification of prior distributions is not possible, which is frequently the case, one would look for automatic method that uses standard noninformative priors for calculating Bayes Factors. There are, however, difficulties with noninformative priors that are typically improper.

A solution to the above problem with improper priors is to use part of the data as a training sample. The data are divided into two parts, $X = (X_1, X_2)$. The first part X_1 is used as a training sample to

obtain proper posterior distributions for the parameters (given X_1) starting from the noninformative priors.

$$G_i(\theta_i|X_1) = \frac{f_i(X_1|\theta_i)g(\theta_i)}{\int f_i(X_1|\theta_i)g(\theta_i)d\theta_i}$$

These proper posteriors are then used as priors to compute the Bayes factor with the remainder of the data (X_2). This conditional Bayes factor, condi- tioned on X_1 , can be expressed as

$$\begin{aligned} B_{10}(X_1) &= \frac{\int f_1(X_2|\theta_1)g(\theta_1|X_1)d\theta_1}{\int f_0(X_2|\theta_0)g(\theta_0|X_1)d\theta_0} \\ &= \frac{m_1(X)}{m_0(X)} \frac{\int f_0(X_1|\theta_0)g(\theta_0)d\theta_0}{\int f_1(X_1|\theta_1)g(\theta_1)d\theta_1} \\ &= B_{10}\frac{m_0(X_1)}{m_1(X_2)} \end{aligned}$$

A part of the data, X_1 , may be used as a training sample as described above if the corresponding posteriors $g_i(\theta_i|X_i)$, i = 0,1 are proper or, equivalently, the marginal densities $m_i(X_1)$ of X_1 under M_i, i = 0,1 are finite. One would naturally use minimal amount of data as such a training sample leaving most part of the data for model comparison. A training sample X 1 may be called proper if $0 < m_i(X_1) < \infty$, i = 0, 1 and minimal if it is proper and no subset of it is proper.

3.7 The Intrinsic Bayes Factor

As described above, a solution to the problem with improper priors is obtained using a conditional Bayes factor $B_{10}(X1)$, conditioned on a training sample X_1. However, this conditional Bayes factor depends on the choice the training sample X_1 . Let X(l), l = 1,2, ... L be the list of all possible minimal training samples. Berger and Pericchi (1996a) suggest considering all these minimal training samples and taking average of the corresponding L conditional Bayes factors B_{10} (X(l))'s to obtain what is called the intrinsic Bayes factor (IBF). For example, taking an arithmetic average leads to the arithmetric intrinsic Bayes factor (AIBF)

$$AIBF_{10} = B_{10}\frac{1}{L}\sum_{l=1}^{L}\frac{m_0(X(l))}{m_1(X(l))}$$

and the geometric average gives the geometric intrinsic Bayes factor (GIBF)

$$GIBF_{10} = B_{10}\left(\prod_{l=1}^{L}\frac{m_0(X(l))}{m_1(X(l))}\right)$$

the sum and product in being taken over the L possible training samples X(l), $l = 1, \ldots, L$.

4 Geometric Skew Normal Distribution [2]

It is a new three parameter skewed distribution introduced by Prof. Debasis Kundu of which normal distribution is a special case. This distribution is obtained by using geometric sum of independent identically distributed normal random variables. We call this distribution as the geometric skew normal distribution. Different properties of this new distribution have been investigated. The probability density function of geometric skew normal distribution can be unimodal or multimodal, and it always has an increasing hazard rate function.

Definition : Suppose $N \sim GE(p)$, $\{X_i : i = 1, \ldots\}$ are i.i.d. $N(\mu, \sigma^2)$ random variables, and N and X_i 's are independently distributed. Define

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{N} X_i$$

.

Then X is said to be GSN random variable with parameters μ, σ and p. It will be denoted as $GSN(\mu, \sigma, p)$.

The joint PDF, $f_{X,N}(x, n)$ of (X, N) is given by

$$f_{X,N}(x,n) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi n}} e^{-\frac{1}{2n\sigma^2}(x-n\mu)^2} p(1-p)^{n-1} & 0 < p < 1 \\ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} & p = 1 \end{cases}$$

for $-\infty < x < \infty$, $\sigma > 0$ and for any positive integer n.

If $p \neq 1$ the PDF of X becomes,

$$\begin{aligned} f_X(x) &= \sum_{n=1}^{\infty} f_{X,N}(x,n) \\ &= \sum_{n=1}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2n\sigma^2}(x-n\mu)^2} p(1-p)^{n-1} \end{aligned}$$

Generation from GSN

- Step 1 : Generate from GE(p)
- Step 2 : Generate x from $N(m\mu, m\sigma^2)$,and x is the required sample.

Moment Generating Function

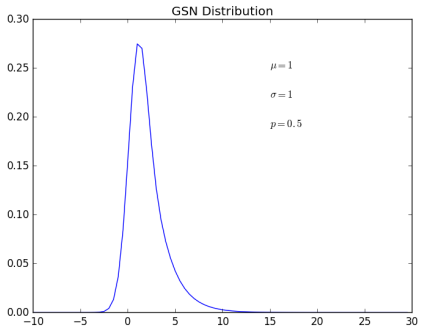
If $X \sim GSN(\mu, \sigma, p)$, then the moment generating function of X becomes,

$$M_X(t) = Ee^{tX} = E[E(e^{tx}|N)] = E[e^{N\mu t + \frac{N\sigma^2 t^2}{2}}] = \frac{pe^{\mu t + \frac{\sigma^2 t^2}{2}}}{1-(1-p)e^{\mu t + \frac{\sigma^2 t^2}{2}}}, t \in \mathbb{R}$$

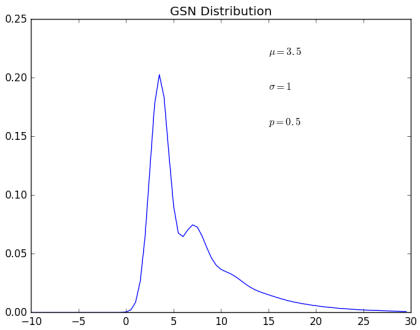
Suppose $Y = X_1 + \cdots + X_n$, where $X_i \sim GSN(\mu, \sigma, p)$ and $X_i's$ are iid $\forall i = 1, \ldots, n$. Then

$$M_Y(t) = (M_X(t))^n = \left(\frac{pe^{\mu t + \frac{\sigma^2 t^2}{2}}}{1-(1-p)e^{\mu t + \frac{\sigma^2 t^2}{2}}} \right)^n$$

.



(a) Unimodel GSN



(b) Bimodal GSN

4.1 Statistical Inference

In this section, we will compute the Bayes Estimates of GSN distribution given some data. We will observe that the estimates cannot be obtained in explicit form. So we will emply **Metropolis-Hastings Algorithm** to sample from the posterior distribution function and then compute the Bayes estimates. Suppose we take $X = \{x_1 \ldots, x_n\}$ as a random sample of size n from $GSN(\mu, \sigma, p)$. Then the likelihood function $L(X; \mu, \sigma, p)$ is given by:

$$L(X;\mu,\sigma,p) = \prod_{i=1}^{n} \left\{ \sum_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\Big[- \frac{(x_i - t\mu)^2}{2t\sigma^2} \Big] \right\}$$

Further, suppose that μ, σ and p are unknown. Then independent prior are chosen for them. Let the prior of μ be a gaussian, for σ is inverse gamma and for p is beta distribution.

Let $\Theta = (\mu, \sigma, p)$ .Therefore, the complete prior probability $P(\Theta)$ can be given by

$$P(\Theta) = f(\mu)g(\sigma)h(p)$$

where

$$\begin{aligned}
f(\mu) &= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \\
g(\sigma) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\Big(- \frac{\beta}{\sigma} \Big), \ \alpha > 0 \text{ shape}, \ \beta > 0 \text{ scale} \\
h(p) &= \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)}, \ \text{where } Beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}
\end{aligned}$$

Thus, the posterior probability is given by

$$\begin{aligned}
k(\Theta|X) &= \frac{L(X|\Theta)P(\Theta)}{\int_\Theta L(X|\Theta)P(\Theta)} \\
&= \frac{\prod_{i=1}^{n} \left\{ \sum_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\Big[- \frac{(x_i - t\mu)^2}{2t\sigma^2} \Big] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\Big(- \frac{\beta}{\sigma} \Big) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)}}{\iiint_{\mu\sigma p} \prod_{i=1}^{n} \left\{ \sum_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\Big[- \frac{(x_i - t\mu)^2}{2t\sigma^2} \Big] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\Big(- \frac{\beta}{\sigma} \Big) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)} \, dp \, d\sigma \, d\mu}
\end{aligned}$$

4.1.1 Bayes Estimates of GSN

In this section, we will try to find the bayes estimates of the posterior density function computed above under squared error loss function. In that case the bayes estimate is given by the mean of the posterior distribution. The bayes estimates of the indivisual parameters can be computed using

$$\delta_\mu(X) = \frac{\iiint_{\mu\sigma p} \mu \prod_{i=1}^{n} \left\{ \sum_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\Big[- \frac{(x_i - t\mu)^2}{2t\sigma^2} \Big] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\Big(- \frac{\beta}{\sigma} \Big) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)} \, dp \, d\sigma \, d\mu}{\iiint_{\mu\sigma p} \prod_{i=1}^{n} \left\{ \sum_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\Big[- \frac{(x_i - t\mu)^2}{2t\sigma^2} \Big] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\Big(- \frac{\beta}{\sigma} \Big) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)} \, dp \, d\sigma \, d\mu}$$

$$\delta_\sigma(X) = \frac{\iiint_{\mu\sigma p} \sigma \prod_{i=1}^{n} \left\{ \sum_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\Big[- \frac{(x_i - t\mu)^2}{2t\sigma^2} \Big] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\Big(- \frac{\beta}{\sigma} \Big) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)} \, dp \, d\sigma \, d\mu}{\iiint_{\mu\sigma p} \prod_{i=1}^{n} \left\{ \sum_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\Big[- \frac{(x_i - t\mu)^2}{2t\sigma^2} \Big] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\Big(- \frac{\beta}{\sigma} \Big) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)} \, dp \, d\sigma \, d\mu}$$

$$\delta_p(X) = \frac{\iiint\limits_{\mu\sigma p} p \prod\limits_{i=1}^{n} \left\{ \sum\limits_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\left[-\frac{(x_i - t\mu)^2}{2t\sigma^2} \right] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma} \right) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)} \, dp \, d\sigma \, d\mu}{\iiint\limits_{\mu\sigma p} \prod\limits_{i=1}^{n} \left\{ \sum\limits_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\left[-\frac{(x_i - t\mu)^2}{2t\sigma^2} \right] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma} \right) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)} \, dp \, d\sigma \, d\mu}$$

Note that the bayes estimates are in the form of ratio of integrals. These cannot be solved analytically. However, there are various approximation methods introduced to solve such type of integrals. One of the methods to solve the ratio of integrals is given by Lindley [3]. But here we will not consider that approach.

We note that the normalising constant of the posterior distribution cannot be obtained analytically. In this case the best procedure to generate samples from the posterior distribution is by using **Metropolis-Hastings Algorithm**.

4.1.2 Metropolis-Hastings Algorithm

The algorithm gives samples from the **target distribution**. It requires a **proposal distribution** function from which we will generate the samples and we accept those samples using the **acceptance probability**.

Target Distribution is given by

$$\pi(\mu, \sigma, p) = \frac{\prod\limits_{i=1}^{n} \left\{ \sum\limits_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\left[-\frac{(x_i - t\mu)^2}{2t\sigma^2} \right] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma} \right) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)}}{\iiint\limits_{\mu\sigma p} \prod\limits_{i=1}^{n} \left\{ \sum\limits_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sigma\sqrt{2\pi t}} \exp\left[-\frac{(x_i - t\mu)^2}{2t\sigma^2} \right] \right\} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(\mu-\mu_1)}{2{\sigma_1}^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma} \right) \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)} \, dp \, d\sigma \, d\mu}$$

Proposal Distribution is given by

$$q(\mu, \sigma, p) = \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\mu - \mu')^2}{2\sigma_2^2} \right] \right) \left(\frac{\beta'^{\alpha'} \sigma^{\alpha'-1} e^{-\beta'\sigma}}{\Gamma(\alpha')} \right) \left(\frac{p^{a'-1}(1-p)^{b'-1}}{Beta(a', b')} \right)$$

Generation from proposal distribution: Since μ, σ, and p are independent we can generate them separately from $\mathcal{N}(\mu', \sigma'^2)$, $Gamma(\alpha', \beta')$ and $Beta(a', b')$ respectively. We will chose the parameters of the proposal distribution in a such a way that the mean of the distribution is close to the actual value and variance of the distribution to be very low for effective sampling.

Since μ, σ and p are independent, we can generate separately from normal distribution, gamma distribution and beta distribution.

Let $x = (x_1, x_2, x_3)$ be a current state in the markov chain simulated by q. To generate the next transition state $y = (y_1, y_2, y_3)$ using the current state, simulate

$$\begin{aligned} y_1 &\sim \mathcal{N}(x_1, \sigma'^2) \\ y_2 &\sim \text{Gamma}(x_2/\beta', \beta') \\ y_3 &\sim \text{Beta}(\lambda x_3, \lambda(1 - x_3)) \end{aligned}$$

The expected values of the $\mathcal{N}(x_1, \sigma'^2)$, Gamma$(x_2/\beta', \beta')$ and Beta$(\lambda x_3, \lambda(1 - x_3))$ are x_1, x_2 and x_3 respectively. σ', β' and λ are chosen in such a way that the variance is low for effective sampling. So, the new sample generated will be close to the mean if the variance is low.

Acceptance Probability

$$\rho(y,x) = min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{q(x|y)}{q(y|x)}\right\}$$

The state $y = (y_1, y_2, y_3)$ simulated, is accepted as the next transition state with probability $\rho(y,x)$

Algorithm 1 Metropolis-Hastings Algorithm

0: **Initialization** : $x^{(0)} \sim q(\mu, \sigma, p)$
0: **for** $i = 1, 2, \dots$ **do**
0: $y \sim q(x|x^{(i-1)})$
0: Take $X^{(i)} = \begin{cases} y & \text{with probability } \rho(y, x^{i-1}) \\ x^{(i-1)} & \text{with probability } 1 - \rho(y, x^{i-1}) \end{cases}$
0:

The markov chain $\{X_i = (\mu_i, \sigma_i, p_i), \forall i = 1, 2, \dots\}$ has the stationary distribution as the target probability density function which is same as our posterior distribution. Therefore, the Bayes estimates of the posterior parameters can be obtained as

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\mu_i \qquad \hat{\sigma} = \frac{1}{m}\sum_{i=1}^{m}\sigma_i \qquad \hat{p} = \frac{1}{m}\sum_{i=1}^{m}p_i$$

5 Data Analysis

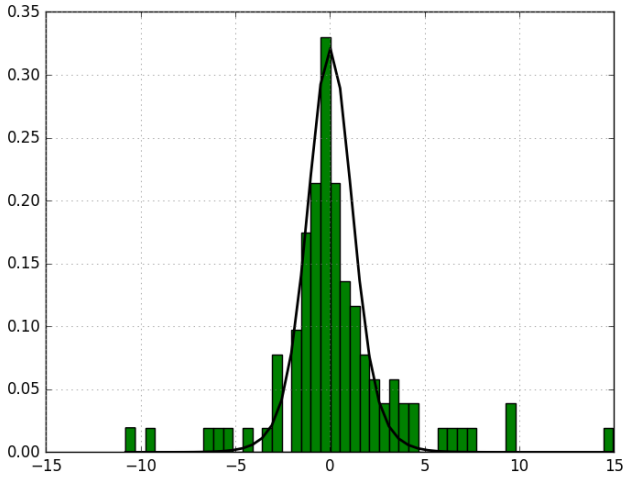
In this section we analyze one data set to see the effectiveness of the proposed model. The data set consist of 100 samples points and we will compute the bayes estimates of the parameters of the Geometric Skew Normal Distribution that fit the the given data.

Following the procedure using Metropolis-Hastings Algorithm, we obtain the following Bayes Estimates

$$\hat{\mu} = 0.00096 \qquad \hat{\sigma} = 1.0450 \qquad \hat{p} = 0.569274859$$

The corresponding **95% Confidence Interval** becomes for μ, σ and p are (0.00096 ± 0.0041), (1.0450 ± 0.03443306) and (0.5692 ± 0.02502) respectively.

We also provide a histogram with the fitted PDF



(a) Histogram of the data with the fitted PDF

6 Conclusions

We explored the different methods to sample from the posterior distribution and obtain the desired estimates from it. We used Monte Carlo Sampling for generating the samples and calculating the Bayes estimates of the desired parameters. Monte Carlo sampling based approaches for inference make use of limit theorems such as the law of large numbers and the central limit theorem to justify their validity. It may appear at first that this procedure necessarily depends on waiting until the Markov chain converges to the target invariant distribution, and sampling from this distribution. In other words, one needs to start a large number of chains beginning with different starting points, and pick the draws after letting these chains run sufficiently long. This is certainly an option, but the law of large numbers for dependent chains, says also that this is unnecessary, and one could just use a single long chain. It may, however, be a good idea to use many different chains to ensure that convergence indeed takes place.

References

[1] Jayanta K Ghosh, Mohan Delampady, and Tapas Samanta. *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media, 2007.

[2] Debasis Kundu. Geometric skew normal distribution. *Sankhya B*, 76(2):167–189, 2014.

[3] Dennis V Lindley. Approximate bayesian methods. *Trabajos de estadística y de investigación operativa*, 31(1):223–245, 1980.

[4] Dana Sylvan. Introduction to mathematical statistics: by rv hogg, j. mckean, and at craig, boston, ma: Pearson, 2012, isbn 978-0-321-795434, x+ 694 pp., 110.67., 2013.