Bayesian Framework for Geometric Skew Normal Distribution

Peshal Agarwal

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Supervisor: Prof. Debasis Kundu

Outline

Plan

Introduction

- $N \sim \mathrm{GE}(p)$ and $X_1, X_2, \ldots, X_N$ be *i.i.d.* $N(\mu, \sigma^2)$
- Let $X = \sum_{i=1}^{N} X_i$, then

$$X \sim GSN(\mu, \sigma, p)$$

- The pdf takes the following form

$$f(x; \mu, \sigma, p) = \sum_{k=1}^{\infty} \frac{p}{\sigma\sqrt{k}} \phi\Big(\frac{x - k\mu}{\sigma\sqrt{k}}\Big)(1-p)^{k-1}$$

- However, we would be working with joint distribution of X and N

$$f_{X,N}(x, n) = \frac{1}{\sigma\sqrt{2\pi n}} exp\Big(-\frac{1}{2n\sigma^2}(x - n\mu)^2\Big) p(1-p)^{n-1}$$

Prior Assumptions and Posterior

- We take the following prior assumptions

$$P(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$P(\mu) = \frac{1}{\sqrt{2\pi\omega^2}} \exp\left(-\frac{1}{2\omega^2}(\mu-\theta)^2\right)$$

$$P(\sigma^2) = \frac{\delta^\gamma}{\Gamma(\gamma)}(\sigma^2)^{(-\gamma-1)} \exp\left(\frac{-\delta}{\sigma^2}\right)$$

- Analysis gives us the following posteriors

$$\Pi(\mu \mid \sigma, \{x_i\}, \{m_i\}) \sim Normal(\hat{\theta}, \hat{\omega}^2)$$

$$\Pi(\sigma^2 \mid \mu, \{x_i\}, \{m_i\}) \sim InvGamma(\gamma + n/2, \delta + K/2)$$

$$\Pi(p \mid \{x_i\}, \{m_i\}) \sim Beta(n+\alpha, \sum_{i=1}^{n} m_i - n + \beta)$$

Plan

Algorithm

- We implement the following iterative algorithm to infer the posterior
  1. Initialize $\mu^{(0)}, \sigma^{(0)}, p^{(0)}$
  2. For $t = 0, 1, 2 \dots$
     1. $m_{i,t} = E(N \mid X = x_i, \mu^{(t)}, \sigma^{(t)}, p^{(t)})$ for each $i \in [n]$
     2. $l_{i,t} = E(N^{-1} \mid X = x_i, \mu^{(t)}, \sigma^{(t)}, p^{(t)})$ for each $i \in [n]$
     3. $\mu^{(t+1)} \sim N(\hat{\theta}, \hat{\omega}^2)$
     4. $(\sigma^{(t+1)})^2 \sim InvGamma(\gamma + n/2, \delta + K^t/2)$
     5. $p^{(t+1)} \sim Beta(n + \alpha, \sum_{i=1}^{n} m_{i,t} - n + \beta)$
  3. Repeat for k iterations (burning period)
  4. Find mean and confidence interval using future iterations

Plan

Data Generation

All the below data analysis is performed under the following setup:

- Hyperparameters are set so that priors are close to uniform. $\alpha = \beta = 1$, $\theta = 0$, $\omega = 100$,$\gamma = 0.1$ and $\delta = 0.1$
- We always set initial values of parameters as $\mu = 1.4$, $\sigma = 0.8$ and $p = 0.4$
- Buring period $k = 1500$
- While estimating $\mathbb{E}[N|x, \mu, \sigma, p]$ and $\mathbb{E}[N^{-1}|x, \mu, \sigma, p]$, the infinite sums of both numerator and denominator are approximated by summing just the first 20 terms

Data Analysis

Variation with respect to sample size

Parameter	Avg. Estimates	Cover Fraction	Avg. 95% Length	MSE
$\mu = 2$	2.08	0.88	0.91	0.10
$\sigma = 1$	1.06	0.89	0.61	0.04
$p = 0.5$	0.52	0.91	0.28	0.01

Table: Estimation from sampling performed on simulated data with 1,000 samples with each sample having **sample size 100**, replicated 800 times

Parameter	Avg. Estimates	Cover Fraction	Avg. 95% Length	MSE
$\mu = 2$	2.15	0.90	1.27	0.19
$\sigma = 1$	1.12	0.90	0.88	0.09
$p = 0.5$	0.54	0.94	0.38	0.01

Table: Estimation from sampling performed on simulated data with 1,000 samples with each sample having **sample size 50**, replicated 800 times

Data Analysis

Here we look at effect of the value of p (keeping sample size fixed to 100)

Parameter	Avg. Estimates	Cover Fraction	Avg. 95% Length	MSE
$\mu = 0$	-0.01	0.92	0.49	0.007
$\sigma = 1$	0.98	0.94	0.55	0.01
$p = 0.75$	0.70	0.95	0.61	0.013

Parameter	Avg. Estimates	Cover Fraction	Avg. 95% Length	MSE
$\mu = 0$	0.00	0.96	0.51	0.006
$\sigma = 1$	1.07	0.98	0.66	0.021
$p = 0.5$	0.59	1	0.59	0.022

Parameter	Avg. Estimates	Cover Fraction	Avg. 95% Length	MSE
$\mu = 0$	-0.03	0.93	0.65	0.011
$\sigma = 1$	1.39	0.9	0.88	0.194
$p = 0.25$	0.50	0.88	0.53	0.077

Data Analysis

We also analyze effect of σ (keeping sample size fixed to 100)

Parameter	Avg. Estimates	Cover Fraction	Avg. 95% Length	MSE
$\mu = 0$	-0.01	0.92	0.49	0.007
$\sigma = 1$	0.98	0.94	0.55	0.01
$p = 0.75$	0.70	0.95	0.61	0.013

Parameter	Avg. Estimates	Cover Fraction	Avg. 95% Length	MSE
$\mu = 0$	-0.03	0.91	2.35	0.197
$\sigma = 5$	4.72	0.89	2.77	0.391
$p = 0.75$	0.67	0.92	0.62	0.019

Data Analysis

We find similar patterns in the ML estimates

Parameter	Avg. MLE	True Value	MSE
μ	0.00	0	0.004
σ	1.17	1	0.043
p	0.35	0.25	0.013

Parameter	Avg. MLE	True Value	MSE
μ	0.00	0	0.005
σ	1.05	1	0.014
p	0.56	0.5	0.012

Parameter	Avg. MLE	True Value	MSE
μ	-0.01	0	0.008
σ	1.00	1	0.007
p	0.78	0.75	0.012

Plan

Guinea Pig data

- We study the survival times of guinea pigs injected with different doses of tubercle bacilli. It contains a total of 72 observations. We obtain the following estimates

Parameter	Posterior Mean	MLE Estimates
μ	1.17	1.13
σ^2	0.26	0.35
p	0.58	0.57

Table: Analysis on guinea pig dataset

- The KS test statistics turns out to be **0.09** and the corresponding p-value 0.58.

Guinea Pig data

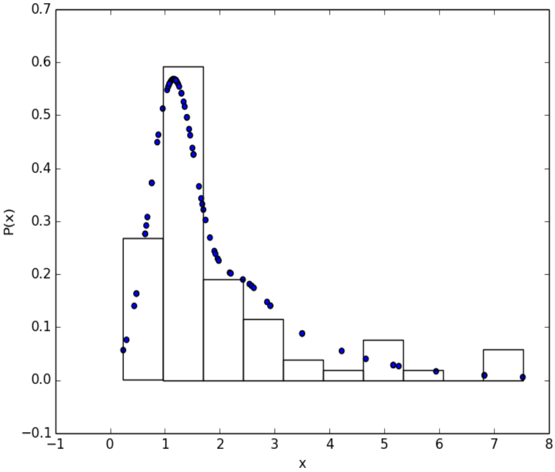
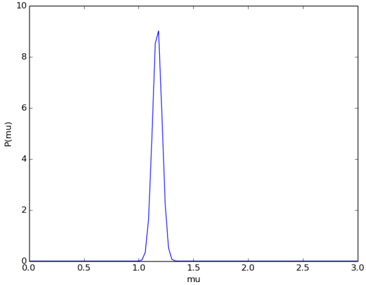
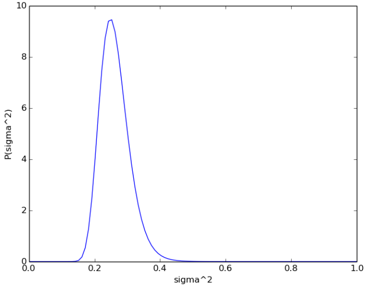


Figure: Plot of GSN (pdf) and histogram for Guinea pig data

Guinea Pig data

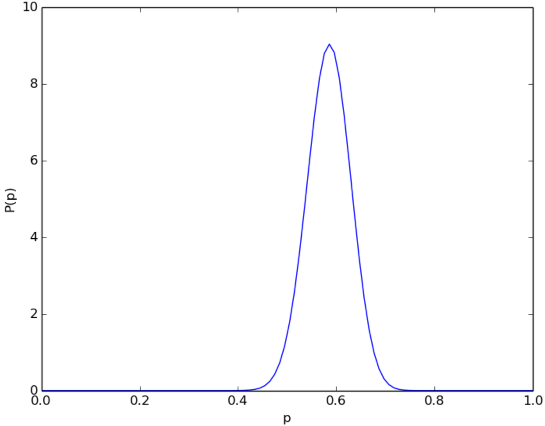


(a) Posterior of $\mu \sim N(1.17, 0.002)$



(b) Posterior of $\sigma^2 \sim \mathit{InvGamma}(\alpha = 36.1, \beta = 9.2)$

Guinea Pig data



(c) Posterior of $p \sim Beta(\alpha = 73.0, \beta = 51.9)$

JEE Data

- We also fit GSN to the marks obtained by students in JEE in Physics, Chemistry and Mathematics.
- We find that all the three follow a Gaussain distrbution ($p \approx 1$)
- Thus, we show that GSN also fits *nicely*, even if the data is not skewed.
- The KS statistcs are 0.04, 0.06 and 0.07 for Physics, Chemistry and Mathematics respectively

Fit to JEE Data

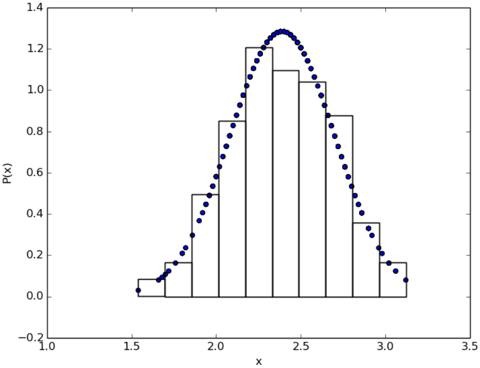


Figure: Plot of GSN (pdf) and histogram for Physics marks

Fit to JEE Data

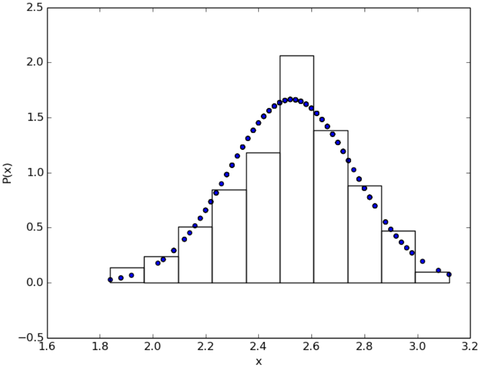


Figure: Plot of GSN (pdf) and histogram for Chemistry marks

Fit to JEE Data

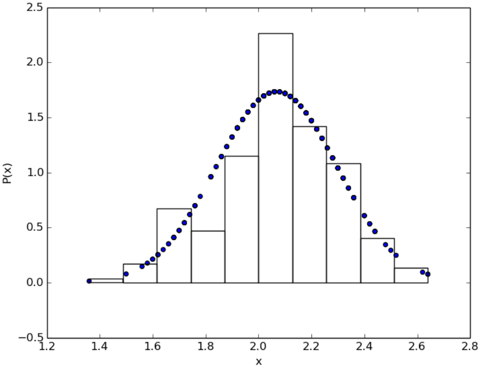


Figure: Plot of GSN (pdf) and histogram for Mathematics marks

Plan

Conclusion

- Developed a fully Bayesian framework for a skewed distribution
- Obtain a complete distribution over parameters instead of a point estimate
- Validated the robustness for a wide range for values
- Show that it fits well even for non-skewed data
- One can learn hyper-parameters from data for a better fit
- One can also extend this to a multivariate setting

References I

Kundu, D.
Geometric Skew Normal Distribution.
Sankhya, 72(2):167–189, 2014.

Azzalini, A.
A Class of Distributions Which Includes the Normal Ones.
Scandinavian Journal of Statistics, 12(2):171–178, 1985.

Bjerkedal, T.
Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli
American Journal of Hygiene, vol. 72, 130-148 1960.

Ghosh, Jayanta K., **Delampady**, Mohan, **Samanta**, Tapas
An Introduction to Bayesian Analysis.
Springer Text in Statistics, 2006.

THANK YOU