ISSN 2231-346X (Print)
JUSPS-A Vol. 30(5), 283-290 (2018). Periodicity-Monthly
Section A
(Online) ISSN 2319-8044

JOURNAL OF ULTRA SCIENTIST OF PHYSICAL SCIENCES
An International Open Free Access Peer Reviewed Research Journal of Mathematics
website:- www.ultrascientist.org
Estd. 1989

Bayesian analysis of Word frequency distribution in context of Indian literature

VASTOSHPATI SHASTRI[1] RAKESH RANJAN[1] PRAVEEN KUMAR TRIPATHI[2] and S.K. UPADHYAY[12]

[1]DST-Centre of Interdisciplinary Mathematical Sciences Banaras Hindu University, Varanasi (India)
[2]Department of Statistics, Banaras Hindu University, Varanasi (India)
Corresponding Author E-mail: vastoshpati@gmail.com
http://dx.doi.org/10.22147/jusps-A/300503

Acceptance Date 24th April, 2018, Online Publication Date 2nd May, 2018

Abstract

This paper deals with the analysis of words from the text of an Indian author. The text of book is analysed and a frequency of noun words is formed. A suitable statistical model, Sichel distribution is fitted to the data and the fitting is found adequate. We have obtained maximum likelihood (ML) estimators and thereafter using flat prior posterior distribution is obtained. Using Metropolis algorithm we draw the posterior samples from which inference are drawn. In the discussion a European text is also compared and we have found that there is richness in Indian literature.

Key words : Word frequency, Sanskrit text, Sichel distribution, ML estimator, Chi square fitting, prior, posterior, Bayesian Inference, Metropolis.

Mathematics Subject Classification: 62F15, 62G07, 62-07.

1. Introduction

Statistics is being used for obtaining inferences for given data particularly a numerical or categorical data. In recent past, studies for the vocabulary of an author from a given sample of his writings are gaining attention. Generally authors of a particular language or region have habits of using some common words in their texts, which persist over long periods of time and wide ranges of subject matter. Study of Morton[5] helped in taking decision about the authorship of Greek work using stylistic evidence. Estimating an author's vocabulary was also considered by McNeil[4]. Few cases have been reported about statistical analysis which solved dispute related to authorship.

Inferences could also be drawn from sentence-length distributions or word distribution. In sentence length distribution stylistics and patterns provide inferences, on the other hand in word distribution inference is based on word count, uses of nouns, adjectives, and verbs. There are three models for the word frequency distributions namely the lognormal law, the generalized inverse Gauss-Poisson law and the extended generalized Zipf's law (see, Piantadosi[6]). Baayen[1] is a text which deals with Statistical models for word frequency distributions.

If different words in a text are arranged in ascending order of their respective frequencies, a reversed J shaped curve will appear with a long upper tail forming a word frequency distribution. The random variable r denotes the number of times a specific word is used, is discrete with origin at $r = 1$. Yule[14] strongly recommended that only words of one particular kind e.g. noun, verb, adjective, etc. should be considered for drawing inferences. Word frequencies were dealt by Zipf [15], Good[2], Simon[10] but there fitting was not justified by Chi-square test of goodness of fit.

Yule[14] is a text which contains fair number of observed word frequency distributions. Herdan[3] stated that only compound Poisson distribution has property to characterize word frequency distribution. Later it is derived by Sichel[8] by making use of Bessel function integral for $r \geq 0$ as

$$f(r) = \frac{\left((1-\theta)^{\frac{1}{2}}\right)^{\gamma}}{K_{\gamma}\left(\alpha(1-\theta)^{\frac{1}{2}}\right)} \frac{(\alpha\theta)^{r}}{r!} K_{r+\gamma}(\alpha) \tag{1}$$

where $K_{\gamma}(.)$ is the modified Bessel function of the second kind of order γ.

Since observed word frequency distributions of text are highly skewed in the extreme, therefore mixing with other high tailed distribution is mandatory. One of the high tailed distribution is inverse Gaussian distribution. The generalized inverse Gaussian distribution has the form

$$f(\lambda) = \frac{\left(\frac{\psi}{\xi}\right)^{\frac{\gamma}{2}}}{2K_{\gamma}\left(\sqrt{\psi\xi}\right)} \lambda^{\gamma-1} e^{-\left(\frac{\psi\lambda}{2} - \frac{\xi}{2\lambda}\right)} \tag{2}$$

Following the reparameterization $\psi = 2\left(\frac{1}{\theta} - 1\right)$ and $\xi = \frac{\alpha^2\theta}{2}$, we get mix distribution suggested by Good(1969) as

$$f(\lambda) = \frac{1}{2} \frac{\left(2(1-\theta)^{\frac{1}{2}}/\alpha\theta\right)^{\gamma}}{K_{\gamma}\left(\alpha(1-\theta)^{\frac{1}{2}}\right)} \lambda^{\gamma-1} exp\left\{-\left(\frac{1}{\theta} - 1\right)\lambda - \frac{\alpha^2\theta}{4\lambda}\right\} \tag{3}$$

On truncating (1) at $r = 1$ and mixing with (3) Sichel[9] suggested most general form of the model for word frequencies as

$$f(r) = \left[\left((1-\theta)^{\frac{1}{2}}\right)^{-\gamma} K_{\gamma}\left(\alpha(1-\theta)^{\frac{1}{2}}\right) - K_{\gamma}(\alpha)\right]^{-1} \frac{(\alpha\theta/2)^{r}}{r!} K_{r+\gamma}(\alpha), \quad for\ r \geq 1 \tag{4}$$

The model known as Sichel distribution has three parameters $-\infty \leq \gamma \leq \infty$, $0 < \theta < 1$ and $\alpha > 0$. The

considered model was analyzed in classical framework by several authors including Stein *et. al.*[12]. But nothing has appeared in the literature for analyzing the above model in Bayes paradigm. This paper considers data containing nouns from Indian text Shastri[7]. In the next section the data is fitted to the model and Chi-squared test was administered. Third section contains parameter estimation of the model, where maximum likelihood estimators have obtained. Using non-informative prior via Bayes theorem posterior distribution is obtained. We have obtained HPD intervals for the parameters, density plots of the parameters and correlations between parameters. A summary is given in the end with recommendation of further research.

2. Estimation of the parameters :

Let $\underline{r} : r_1, r_2, \dots, r_n$ be the number of occurrences of *n* specific words which were used in a text then the likelihood function corresponding to the word frequency distribution (4) can be written as

$$L = \prod_{i=1}^{n} f(r_i) \tag{5}$$

Using (4) one can further rewrite the equation (5) in the form given below

$$L = \left[\left((1-\theta)^{\frac{1}{2}}\right)^{-\gamma} K_\gamma\left(\alpha(1-\theta)^{\frac{1}{2}}\right) - K_\gamma(\alpha)\right]^{-n} \times \frac{\left(\frac{\alpha\theta}{2}\right)^{\sum_{i=1}^{n} r_i}}{\prod_{i=1}^{n} r_i!} \prod_{i=1}^{n} K_{r_i+\gamma}(\alpha) \tag{6}$$

To obtain the maximum likelihood estimates of the parameters, it is more convenient to write the log-likelihood function of word frequency distribution (4) which can be written as

$$\begin{aligned} logL = {} & -n\, log\left[\left((1-\theta)^{\frac{1}{2}}\right)^{-\gamma} K_\gamma\left(\alpha(1-\theta)^{\frac{1}{2}}\right) - K_\gamma(\alpha)\right] + \sum_{i=1}^{n} r_i \log\left(\frac{\alpha\theta}{2}\right) - \\ & \sum_{i=1}^{n} \log r_i! + \sum_{i=1}^{n} \log K_{r_i+\gamma}(\alpha). \end{aligned} \tag{7}$$

The use of log-likelihood, instead of likelihood function, to obtain the ML estimates of the parameters is justified in the sense that the ML estimates are invariant under the logarithmic transformation. ML estimates can be obtained by maximizing (7). Maximization can be done using any non-linear optimization techniques say for example Nelder Mead method, simulated annealing, Newton Raphson method etc. In this article, we have used Newton Raphson method to maximize (7).

In order to perform Bayesian analysis, first we need to specify prior distributions corresponding to each parameter. If one has strong prior information about the parameters, specific distributional form of the prior can be used. Otherwise, one should go for non informative prior. The major problem with non-informative prior is that they are usually improper and may lead to improper posterior. In this case, we do not have any strong information about the parameters, so we will proceed with vague uniform prior. The following vague priors are considered to form the posterior distribution

$$g_1(\alpha) \propto U[0, M_1] \tag{8}$$

$$g_2(\theta) \propto U[0, 1] \tag{9}$$

$$g_3(\gamma) \propto U[-M_2, M_3] \tag{10}$$

Since parameter θ lies in the interval [0,1], we have consider uniform prior over same interval. The value M_1, M_2, M_3 should be large enough to make the priors vague.

Combining the prior distributions (8), (9), (10) to the likelihood function (6), the resulting posterior distribution up to proportionality can be expressed as

$$p(\alpha,\theta,\gamma|\underline{r}) \propto A^n \frac{\left(\frac{\alpha\theta}{2}\right)^{\sum_{i=1}^n r_i}}{\prod_{i=1}^n r_i!} \prod_{i=1}^n K_{r_i+\gamma}(\alpha) \times I_{[0,M_1]}(\alpha) \times I_{[0,1]}(\theta) \times I_{[M_2,M_3]}(\gamma) \quad (11)$$

After looking at posterior (11), one can easily say that the posterior is analytically not tractable. Hence, sample based approach is only alternative (see, for example, Upadhyay *et al.*[13]). In this study, we will use Metropolis algorithm to draw posterior sample. Metropolis algorithm simulate a candidate observation (α',θ', γ') from symmetric proposal distribution q(α,θ,γ|α^*,θ^*,γ^*) where (α^*,θ^*,γ^*) is the previous realization of parameters. The candidate observation is then accepted with probability α which can be written as

$$\alpha = min\left(1, \frac{p(\alpha',\theta',\gamma'|\underline{r})}{p(\alpha^*,\theta^*,\gamma^*|\underline{r})}\right)$$

If the candidate observation is rejected Metropolis algorithm sets the previous realization as current realization. We, therefore, in our study considered a multivariate normal proposal density. The initial value of mean vector is taken as ML estimates of parameters and variance-covariance matrix was taken based on Hessian based approximation of asymptotic variance covariance matrix of ML estimates. For more details of the scheme one may refer to Smith and Roberts[11].

It is to be noted multivariate normal kernel proposes a candidate observation on real scale but our parameters α and θ lies on positive [0,1] scale. So, for α we take the logarithmic transformation $\phi_1 = log(\alpha)$ and for θ, we consider the logit transformation $\phi_2 = log\left(\frac{\theta}{1-\theta}\right)$ to draw the posterior samples with their support.

3. Primary data and its modelling :

For analyzing word frequency a sample of text is taken from the book "Why Sanskrit" of an eminent Indian author, who received one of highest civilian award in the field of literature. Table 1 shows number of nouns in a text containing 10290 occurrences from a book named "Why Sanskrit". The text contained English and Sanskrit words, after thorough filtering, punctuation and articles were removed, thereafter other non-noun words were removed. The word "Sanskrit" has been repeated 622 times whereas; the percentage of number of nouns occurring only once is 63.6. Use of variety of nouns shows the writing capability of the author. A text reported in Yule[14] have 48.3 percent of nouns used in the sample text. The percentage of noun occurring once is 63.6 percent which is higher that of corresponding European sample texts reported by Sichel[8], McNeil[4], Sichel[9] among others.

Table 1 : Word frequency of text from the book authored by Shastri[7].

r	f_x	r	f_x	r	f_x	r	f_x
1	2395	18	9	37	1	70	1
2	526	19	10	39	1	81	1
3	231	20	6	41	1	82	1
4	143	21	5	42	2	85	1
5	94	22	3	43	1	88	1
6	76	23	5	47	2	95	1
7	48	24	4	50	1	96	1
8	31	25	3	51	2	102	1
9	25	26	3	52	1	114	1
10	18	27	3	53	1	169	1
11	17	28	1	57	1	183	1
12	11	29	3	58	1	213	1
13	12	30	3	59	1	374	1
14	11	31	3	60	1	622	1
15	10	32	3	62	1		
16	7	33	1	66	1		
17	7	34	2	67	2	Total	3764

Here $S_0 = 3764$, $S_1 = 12685$, $S_2 = 840020$, Mean=2.734, $\sigma = 14.68$.

For applying chi-square test of fit, the data has been pooled for frequencies of words greater than 20. The table 2 shows the fitting and value of chi square with its p-value.

Table 2 : Fitting of sample Indian literature data to Sichel distribution

Frequency of Words r	Number of Words observed	Number of Words expected		Frequency of Words r	Number of Words observed	Number of Words expected
1	2395	2389.2		14	11	11.5
2	526	559.1		15	10	10.1
3	231	234.2		16	7	9.0
4	143	129.6		17	7	8.0
5	94	82.9		18	9	7.2
6	76	57.9		19	10	6.5
7	48	42.9		20	6	6.0
8	31	33.2		21-30	33	39.9
9	25	26.5		31-45	23	26.9
10	18	21.7		>45	21	14.9
11	17	18.1		Total	3764	3764
12	11	15.4		χ^2	-	20.487
13	12	13.2		df	-	22.0
				p-value	-	0.553

The p-value is 0.553 which shows the fit is good.

4. Discussion

To get the posterior samples, we used the multivariate normal kernel with mean vector having the elements of ML estimates, that is, -0.8119, 0.3619 and 0.9942 of the parameters γ, α and θ respectively. The variance-covariance matrix can be obtained numerically as hessian based approximation at these values of the corresponding parameters. Of course, to get the efficient posterior samples one has to choose the tuning parameter c carefully and, therefore, a careful investigation suggests to fix $c = 0.7$ is a good choice for our study.

We consider a single long run of the chain of iteration about 200K. We examine the convergence of the chain, based on ergodic average, at about 100K iteration after the initial transient behaviour of the chain. We pick up the samples of size 1K at a gap of 100 to draw the posterior based inferences. Normally, the gap was so chosen to make the negligible correlation among the generating variates. Finally, we obtain the posterior summaries, based on 1K posterior samples, in the form of posterior means, posterior medians, posterior modes and HPD intervals as given below in the table.

Table 2 : Posterior summary of the parameters.

Parameter	Posterior Mean	Posterior Median	Posterior Mode	Posterior HPD Interval
γ	-0.8217	-0.8208	-0.8211	(-0.8808, -0.7648)
α	0.3766	0.3750	0.3773	(0.2682, 0.4949)
θ	0.9946	0.9947	0.9947	(0.9943, 0.9950)

The correlation between the parameters of the model is also obtained and the same is provided in the table given below.

Table 3 : Correlations among the parameters.

Correlation	γ	α	θ
γ	1.000	-0.906	-0.052
α	-0.906	1.000	0.013
θ	-0.052	0.013	1.000

The table simply interprets that the parameters γ and α are highly negatively correlated while, the correlations between (γ, θ) and (θ, α) are comparatively very less. This obviously shows the significance of the use of Bessel function involved in the model.

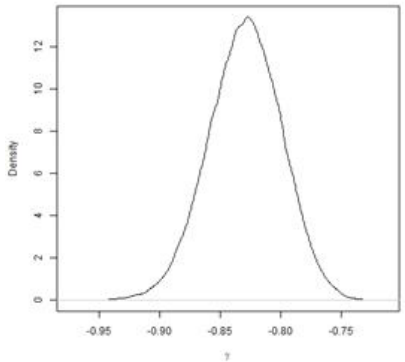

Figure 1: Marginal posterior density plot for γ.

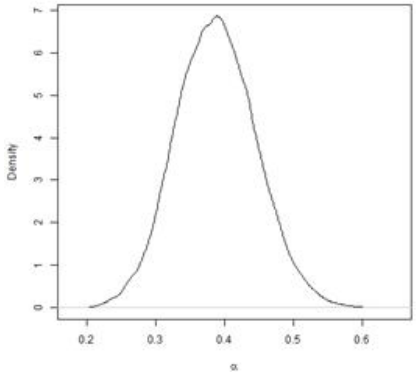

Figure 2: Marginal posterior density plot for α.

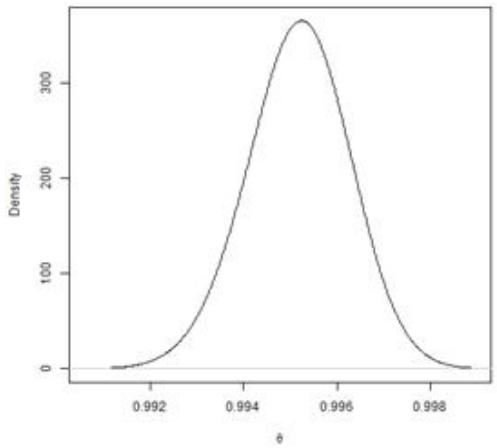

Figure 3: Marginal posterior density plot for θ.

Moving up to posterior inferences one can easily interpret the results reported in the table 2. Just for the sake of concluding the table we can say that the marginal posterior densities of the parameters are almost symmetrical in nature. Also, the posterior estimates are quite close to the ML estimates of the three parameters. The HPD intervals are shorter enough to cover the true estimates of the parameters with maximum probability. The marginal posterior density of the three parameters of the model is provided in the figure 1 to 3.

5. Conclusion

The primary data has obtained and a word frequency distribution was fitted to the data. Using fitting test, it was found that the fitted data is good. We have analyzed the data and found that single occurrence of noun in an Indian sample text of literature is found at more places as compared to a European sample text of literature. For the parameters of word frequency distribution Bayesian analysis was performed and posterior based result has obtained. For comparison purpose ML estimators of the parameters of the word frequency distribution is also obtained. Both the methods yield compromise result. In order to observe writing styles of authors, one may analyze verbs, adjectives and compare with relative data so as to visualize richness of the languages.

References

1. Baayen, H., Statistical models for word frequency distributions: A linguistic evaluation. Computers and the Humanities, *26(5-6)*, 347-363, (1992).
2. Good, I.J., Statistics of Language," The Encyclopaedia of Linguistic Information and Control, Oxford: Pergamon Press, (1969).
3. Herdan, G., Type-token mathematics: A textbook of mathematical linguistics. Mouton, (1960).
4. McNeil, D. R., Estimating an author's vocabulary. Journal of the American Statistical Association, *68(341)*, 92-96 (1973).
5. Morton, A. Q., The authorship of Greek prose. Journal of the Royal Statistical Society. Series A (General), 128(2), 169-233, (1965).

6. Piantadosi, S. T., Zipf 's word frequency law in natural language: A critical review and future directions. Psychonomic bulletin & review, *21(5)*, 1112-1130, (2014).
7. Shastri, V., Why Sanskrit, Vagyoga Chetna Publication, (2016).
8. Sichel, H. S., On a family of discrete distributions particularly suited to represent long-tailed frequency data. In Proceedings of the Third Symposium on Mathematical Statistics.SACSIR. (1971).
9. Sichel, H. S., On a distribution law for word frequencies. Journal of the American Statistical Association, *70(351a)*, 542-547, (1975).
10. Simon, H.A., "On a Class of Skew Distribution Functions," Biometrika, *42*, Nos. 3 & 4, 425-40, (1955).
11. Smith, A. F., & Roberts, G. O., Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. Journal of the Royal Statistical Society. Series B (Methodological), 3-23, (1993).
12. Stein, G. Z., Zucchini, W., &Juritz, J. M., Parameter estimation for the Sichel distribution and its multivariate extension. Journal of the American Statistical Association, *82(399)*, 938-944, (1987).
13. Upadhyay, S. K., Vasishta, N., & Smith, A. F. M., Bayes inference in life testing and reliability via Markov chain Monte Carlo simulation. Sankhy?: The Indian Journal of Statistics, Series A (1961-2002), *63(1)*, 15-40, (2001).
14. Yule G.U., A Statistical Study of Vocabulary, Cambridge, England: Cambridge University Press, (1944).
15. Zipf, G. K., Human Behaviour and the Principle of Least Effort. Cambridge, MA: Addison-Wesley, (1949).