Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 5, Issue 1
www.stmjournals.com

Weighted Latent Dirichlet Allocation, an Optimized and Generalized Probabilistic Topic Model for Large Corpus of Data

Muzafar Rasool Bhat**, M. Arif Wani
Post Graduate Department of Computer Science, University of Kashmir, Jammu and Kashmir, India

Abstract

In this manuscript, a new distribution, Weighted Dirichlet distribution (WDD), which acts as generalization to other distributions like Dirichlet, length or size-biased Dirichlet and area biased Dirichlet distributions, is introduced. The objective of this study is to develop an optimal and generalized probabilistic topic model, weighted latent Dirichlet allocation WLDA for extracting the latent themes from textual corpus using weighted Dirichlet distribution. Various characteristic, statistical and the structural properties of the introduced distribution, especially moments about origin (1st, 2nd, 3rd etc.), measures of central tendency and dispersion are comprehensively studied. For experimentation and comparison with existing model LDA, dataset Cora, a collection of 2410 scientific documents in LDA format conforming to probabilistic topic modeling experimentations has been utilized in this study. To choose optimal topic number k to topic model the corpus, four metrics "CaoJuan2009", "Arun2010", "Griffiths2004" and "Deveaud2014" have been used. Generative process of introduced probabilistic topic model WLDA is also included in the manuscript. Comparison between various special cases of the proposed model is performed based on AIC's, AICC's; BIC's and P-values resulted from the models. Special case of the proposed model, volume biased weighted latent Dirichlet allocation out performs in modeling data as compared to other topic modeling techniques.

Keywords: *Dirichlet distribution, LDA, weighted Dirichlet distribution, WLDA, topic modeling*

****Author for Correspondence*** E-mail: muzafarrasool@gmail.com

INTRODUCTION

Topic models provide a simple way to analyse large volumes of unlabelled text. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings. Topic models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words [1–5].

A topic model is a generative model for documents; it specifies a simple probabilistic procedure by which documents can be generated. For a new document to generate, a distribution over topics needs to be chosen. Then, for each word in that document, a topic at random according to this distribution and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents. Documents with different content can be generated by choosing different distributions over topics. This typical generative process for topics at a corpus level and words within a document can be captured using Dirichlet distribution, often denoted as Dir(α), which is a family of multivariate probability distributions parameterized by a vector of positive real numbers.

Dirichlet distribution is a multivariate generalization of the beta distribution [6]. Dirichlet distributions are very often used as prior distributions in Bayesian statistics, and in fact the Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution. The infinite-dimensional generalization of the Dirichlet distribution is the Dirichlet process. Inference over hierarchical Bayesian models is often done using Gibbs sampling, and in such a case, instances of the Dirichlet distribution are typically marginalized out of the model by integrating out the Dirichlet random variable. This causes the various categorical variables drawn from the same Dirichlet random variable to become correlated, and the joint distribution over them assumes a Dirichlet-multinomial distribution, conditioned on the hyper parameters of the Dirichlet distribution.

In some recent and interesting papers, hierarchical models with a Dirichlet prior, i.e. Dirichlet hierarchical models were used in probabilistic classification applied to various fields besides its usage in statistical machine learning, artificial Intelligence and more specifically in probabilistic topic modelling [7–9]. For extension of probabilistic models to other areas of interest like pattern analysis, machine intelligence, computer vision and character recognition, one can refer to the literature [10–13]. For implementation of LDA on a textual corpora and thorough understanding of algebraic models in vogue, Bhat and Wani elaborate the essential steps involved, besides evaluation of said models on small and large datasets and introduction of mixture model for topic modelling of large corpus of data [14–16].

Big data associated with the problem of analyzing large corpus of text data can be projected into sub-spaces or clusters so that the given problem can be divided into sub-problems. Each sub-problem can then be optimized with an appropriate model [17]. One of the approaches of dividing a given problem into sub-problems involves projecting the big data to various sub-space grids, where each sub-space grid represents a sub-problem [18–21]. Each sub-problem can then be represented independently with various models which can be combined by using a rule based system [22].

DIRICHLET DISTRIBUTION: A BREIF OVERVIEW

The Dirichlet distribution of order $K \geq 2$ with parameters α_1, α_2 ..., α_k > 0 has a probability density function with respect to Lebesgue measure on the Euclidean space R^{K-1} given by (Figure 1):

$$f(x_1,x_2,\ldots,x_k;\alpha_1,\alpha_2,\ldots.\alpha_k) = \frac{1}{\beta(\alpha)}\prod_{i=1}^{k}(x_i)^{\alpha_i-1}$$

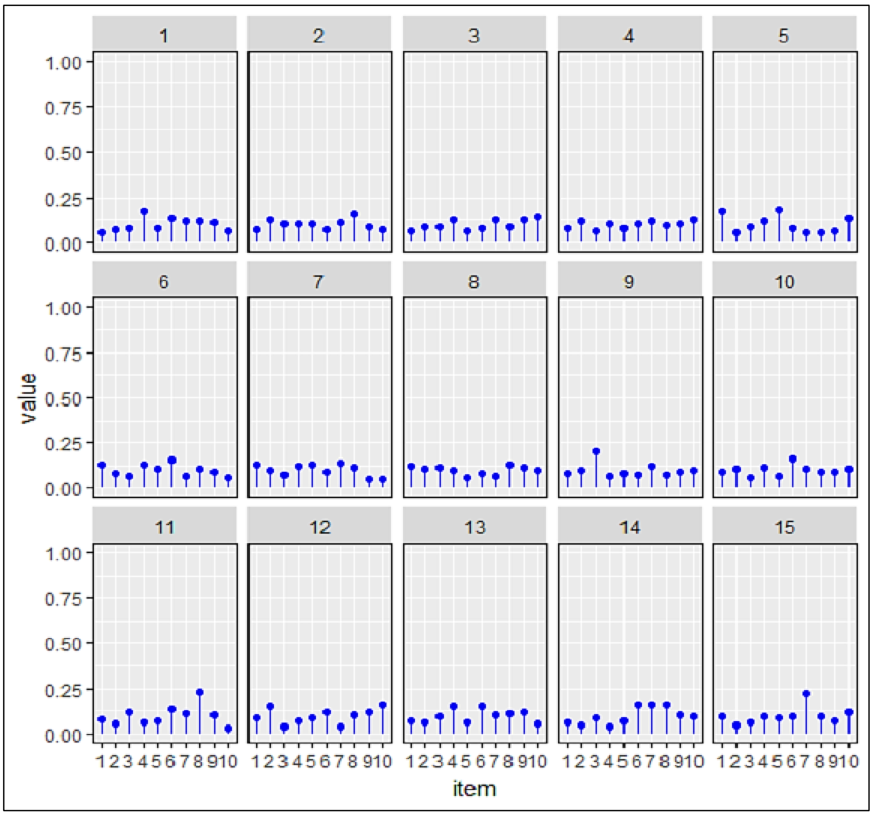

Fig. 1: *PDF Plot of Dirichlet Distribution with* $\alpha = 50/8$.

$$f(x_i;\alpha_i)=\frac{1}{\beta(\alpha)}\prod_{i=1}^{k}(x_i)^{\alpha_i-1} \qquad (1)$$

on the open (K−1)-dimensional simplex defined by:

$\forall x_{i,,i=1,2,3....k}, x_i > 0$,

$\sum_{i=1}^{k} x_i = 1$ and $x_k = 1-\sum_{i=1}^{k-1} x_i$

Where,

$$\beta(\alpha)=\frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma\sum_{i=1}^{k}\alpha_i} \qquad (2)$$

The rth moment about origin of Dirichlet distribution is:

$$E(\prod_{i=1}^{k}x_i^{r_i})=\frac{\prod_{i=1}^{k}\Gamma(\alpha_i+r_i)}{\Gamma\sum_{i=1}^{k}(\alpha_i+r_i)}\frac{\Gamma\sum_{i=1}^{k}(\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}=\frac{\beta(\alpha+r)}{\beta(\alpha)} \qquad (3)$$

WEIGHTED PROBABILITY DISTRIBUTION: A BRIEF OVERVIEW

The concept of weighted distributions can be traced to the work of Fisher; in connection to his studies on how methods of ascertainment can influence the form of distribution [23]. Later it was introduced and formulated in general terms, in connection with modeling statistical data where the usual practice of using standard distributions for the purpose was not found to be appropriate. Rao identified various situations that can be modeled by weighted distributions [24]. These situations refer to instances where the recorded observations cannot be considered as a random sample from the original distributions which may occur due to non-observability of some events. Shaban and Boudrissa introduced length biased version of the Weibull distribution known as Weibull Length-biased (WLB) and examined its shape and other properties [25]. Das and Roy discussed the length-biased Weighted Generalized Rayleigh distribution with its properties [26]. Other contributions have been made by Oluyede and George [27], Ghitany and Al-Mutairi [28], Oluyede and Terbeche [29], Reshi and Ahmed [30] and Reshi *et al*. [31].

Propsed Weighted Dirichlet Distribution

Suppose X is a non-negative random vector with its natural probability density function $f(\prod x_i;\theta)$, where the natural parameter is θ. Suppose a realization x of X under $f(\prod x_i;\theta)$ is recorded with probability proportional to $w(\prod x_i;\lambda)$, so that the weight function $w(\prod x_i;\lambda)$ is a non-negative function with the parameter λ. Clearly, the recorded x is not an observation on X, but on the random vector X_w, having a PDF.

Assume that $E(\prod X_i^{\lambda}) = \int_a^b \left(w(\prod X_i^{\lambda})\right) f(x_i)\, dx_i$ if the moments of w($\prod X_i^{\lambda}$) exists. Therefore a weighted Dirichlet probability distribution is obtained by taking the weights: w($\prod X_i^{\lambda}$) = $\prod_{i=1}^{k} x_i^{\lambda_i}, \lambda_i \geq 0$ to the Dirichlet distribution.

Using Eq. (1) and (3) in Eq. (4), we get,

$$f(x_i;\alpha_i,\lambda_i)=\frac{\int_0^1 \prod_{i=1}^{k} x_i^{\lambda_i}\frac{1}{\beta(\alpha)}\prod_{i=1}^{k}(x_i)^{\alpha_i-1}\, dx_i}{\frac{\beta(\alpha+\lambda)}{\beta(\alpha)}}$$

$$f(x_i;\alpha_i,\lambda_i)=\frac{\int_0^1 \prod_{i=1}^{k}(x_i)^{\alpha_i+\lambda_i-1}\, dx_i}{\beta(\alpha+\lambda)}=1 \qquad (5)$$

From the above Eq. (5), we deduce the weighted Dirichlet distribution and its probability density function is defined as:

$$f(x_i;\alpha_i,\lambda_i)=\frac{\prod_{i=1}^{k}(x_i)^{\alpha_i+\lambda_i-1}}{\beta(\alpha+\lambda)};\alpha_i>0,\lambda_i\geq 0, 0<x_i<1 \qquad (6)$$

Special Cases of Weighted Dirichlet Distribuition

i) If $\lambda_i = 0$, the weighted Dirichlet model reduces to Dirichlet model Eq. (1).

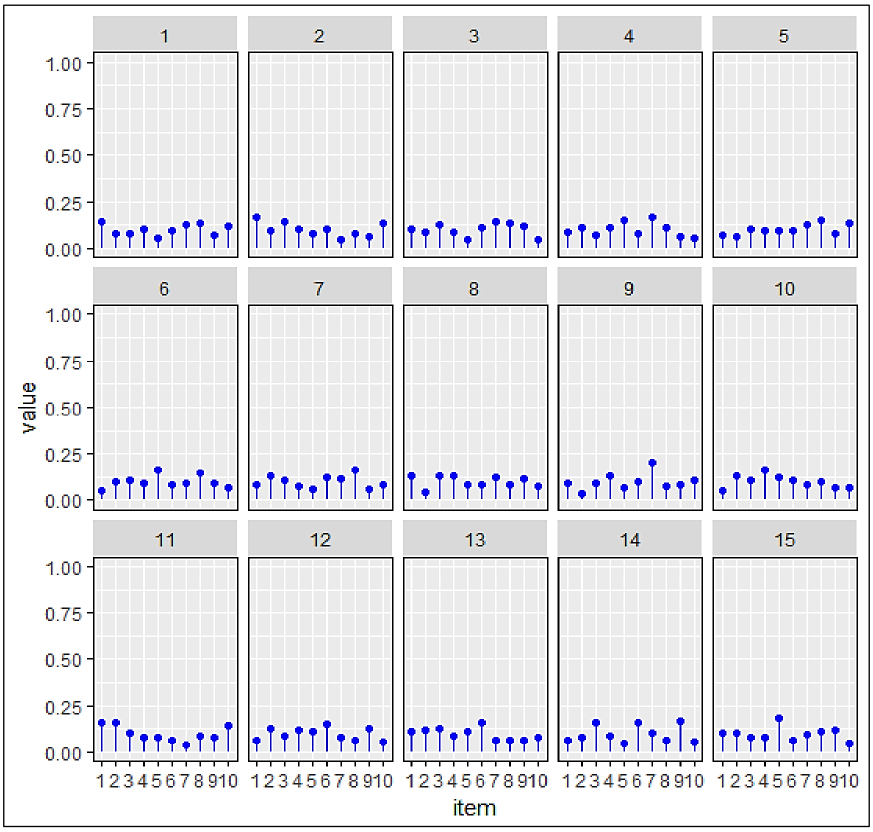

Fig. 2: *PDF Plot of Weighted Dirichlet Distribution (Size Biased) with* $\alpha = 50/8 \text{ and } \lambda = 1$ *where* λ *is a Weight Parameter.*

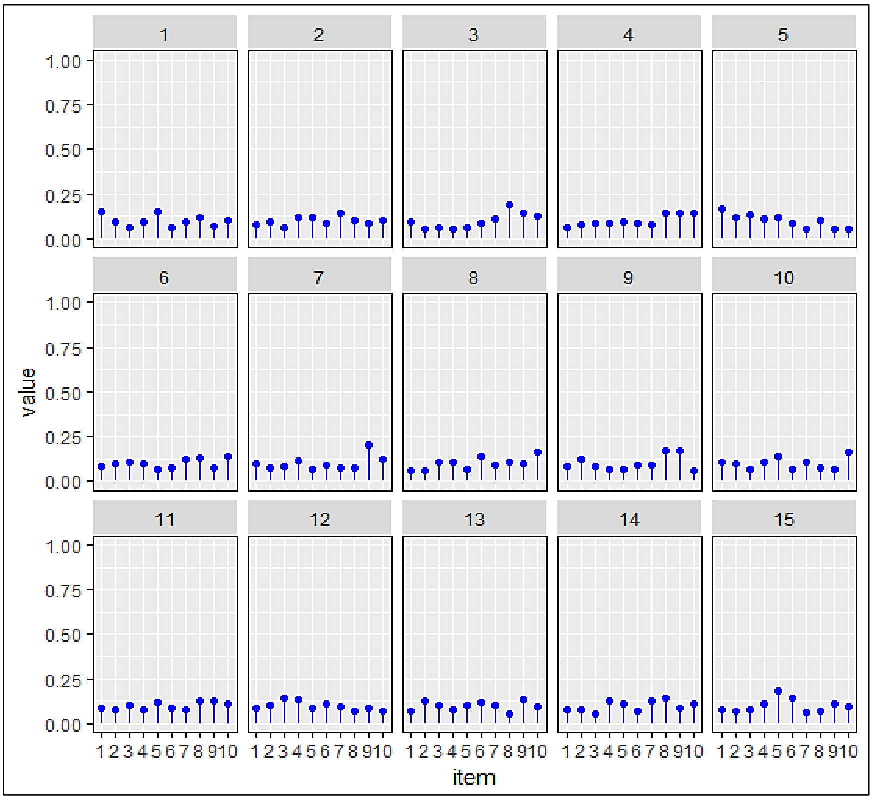

Fig. 3: *PDF Plot of Weighted Dirichlet Distribution (Area Biased) with* $\alpha = 50/8 \text{ and } \lambda = 2$ *where* λ *is a Weight Parameter.*

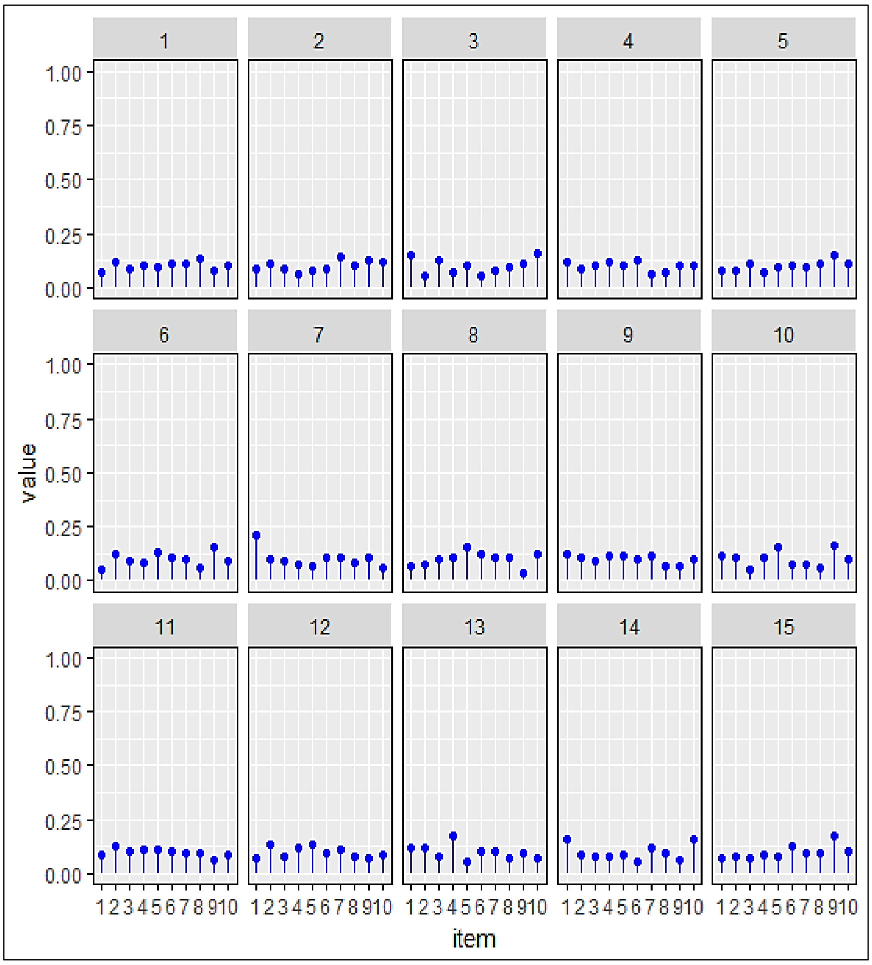

***Fig. 4:** PDF Plot of Weighted Dirichlet Distribution(Volume Biased) with* $\alpha = 50/8$ and $\lambda = 2$ *where* λ *is a Weight Parameter.*

ii) If $\lambda_i = 1$, the weighted Dirichlet model reduces to size or length biased Dirichlet distribution with probabilty density function as (Figure 2):

$$f(x_i;\alpha_i,\ 1) = \frac{\prod_{i=1}^{k} (x_i)^{\alpha_i}}{\beta(\alpha+1)};\alpha_i > 0, 0 < x_i < 1 \tag{7}$$

iii) If $\lambda_i = 2$, the weighted Dirichlet distribution reduces to area biased or even power weighted Dirichlet distribution and its probability density function is given as (Figure 3):

$$f(x_i;\alpha_i,\ 2) = \frac{\prod_{i=1}^{k} (x_i)^{\alpha_i+1}}{\beta(\alpha+2)};\alpha_i > 0, 0 < x_i < 1 \tag{8}$$

iv) If $\lambda_i = 3$, the weighted Dirichlet distribution reduces to volume biased weighted Dirichlet distribution and its probability density function is given as (Figure 4):

$$f(x_i;\alpha_i,\ 3) = \frac{\prod_{i=1}^{k} (x_i)^{\alpha_i+2}}{\beta(\alpha+3)};\alpha_i > 0, 0 < x_i < 1 \tag{9}$$

v) If $\alpha_i = 1$ and $\lambda = 0$, then weighted Dirichlet distribution reduces to uniform distribution.

STRUCTURAL PROPERTIES OF WEIGHTED DIRICHLET DISTRIBUTION

This section deals with some structural properties of the proposed Weighted Dirichlet Distribution. These structural properties include moments about origin, central tendency, and dispersion.

The moments about origin of WDD can be calculated as:

$$E(\prod_{i=1}^{k} x_i^{r_i}) = \int_0^1 \prod_{i=1}^{k} x_i^{r_i} f(x_i;\alpha_i, \lambda_i) dx_i$$

$$E(\prod_{i=1}^{k} x_i^{r_i}) =$$

$$\int_0^1 \prod_{i=1}^{k} x_i^{\lambda_i} \frac{1}{\beta(\alpha+\lambda)} \prod_{i=1}^{k} (x_i)^{\alpha_i+\lambda_i-1} \quad dx_i$$

$$E(\prod_{i=1}^{k} x_i^{r_i}) =$$

$$\frac{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i+\lambda_i)} \int_0^1 \prod_{i=1}^{k} (x_i)^{\alpha_i+\lambda_i+r_i-1} \quad dx_i$$

$$E(\prod_{i=1}^{k} x_i^{r_i}) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i+\lambda_i+r_i)}{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i+r_i)} \frac{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i+\lambda_i)} \tag{10}$$

$$E(\prod_{i=1}^{k} x_i^{r_i}) = \frac{\beta(\alpha+\lambda+r)}{\beta(\alpha+\lambda)} \tag{11}$$

Mean of Weighted Dirichlet Distribution

If we put r_i =1, then Eq. (10) reduces to:

$$E(\prod_{i=1}^{k} x_i) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i+\lambda_i+1)}{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i+1)} \frac{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i+\lambda_i)}$$

$$E(\prod_{i=1}^{k} x_i) = \frac{\alpha_i}{\sum_{i=1}^{k}(\alpha} \tag{12}$$

Variance of Weighted Dirichlet Distribution

If we put r_i =2, then Eq. (10) reduces to:

$$E(\prod_{i=1}^{k} x_i^2) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i+\lambda_i+2)}{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i+2)} \frac{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i+\lambda_i)}$$

After simplification, we get second moment about origin as:

$$E(\prod_{i=1}^{k} x_i^2) = \frac{(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1)}{\sum_{i=1}^{k}(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1)} \tag{13}$$

The variance of WDD can be obtained as:

$$V(\prod_{i=1}^{k} x_i) = E(\prod_{i=1}^{k} x_i^2) - \{E(\prod_{i=1}^{k} x_i)\}^2$$

$$V(\prod_{i=1}^{k} x_i) = \frac{(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1)}{\sum_{i=1}^{k}(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1)} - \left(\frac{\alpha_i+\lambda_i}{\sum_{i=1}^{k}(\alpha_i+\lambda_i)}\right)^2$$

After simplifying the above equation, we have:

$$V(\prod_{i=1}^{k} x_i) = \left(\frac{(\alpha_i+\lambda_i)\sum_{i=1}^{k}(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1) - (\alpha_i+\lambda_i)^2\sum_{i=1}^{k}(\alpha_i+\lambda_i+1)}{\sum_{i=1}^{k}(\alpha_i+\lambda_i)^2(\alpha_i+\lambda_i+1)}\right) \tag{14}$$

Standard Deviation and Coefficient of Variation of Weighted Dirichlet Distribution

The standard deviation of weighted Dirichlet distribution is given as:

$$SD(\prod_{i=1}^{k} x_i) = \sqrt{\left(\frac{(\alpha_i+\lambda_i)\sum_{i=1}^{k}(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1) - (\alpha_i+\lambda_i)^2\sum_{i=1}^{k}(\alpha_i+\lambda_i+1)}{\sum_{i=1}^{k}(\alpha_i+\lambda_i)^2(\alpha_i+\lambda_i+1)}\right)} \tag{15}$$

Using Eqs. (12) and (15), we get coefficient of variation of weighted Dirichlet distribution and is given as:

$$CV(\prod_{i=1}^{k} x_i) = \frac{\sqrt{\left(\frac{(\alpha_i+\lambda_i)\sum_{i=1}^{k}(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1) - (\alpha_i+\lambda_i)^2\sum_{i=1}^{k}(\alpha_i+\lambda_i+1)}{\sum_{i=1}^{k}(\alpha_i+\lambda_i)^2(\alpha_i+\lambda_i+1)}\right)}}{\frac{\alpha_i+\lambda_i}{\sum_{i=1}^{k}(\alpha_i+\lambda_i)}} \tag{16}$$

Harmonic Mean for Weighted Dirichlet Distribution

The harmonic mean for weighted Dirichlet distribution can be calculated as:

$$HM(\prod_{i=1}^{k} x_i) = \int_0^1 \frac{1}{\prod_{i=1}^{k} x_i} f(x_i;\alpha_i, \lambda_i) dx_i$$

$$HM(\prod_{i=1}^{k} x_i) = \int_0^1 \frac{1}{\beta(\alpha+\lambda)} \prod_{i=1}^{k} (x_i)^{\alpha_i+\lambda_i-2} \quad dx_i$$

$$HM(\prod_{i=1}^{k} x_i) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i + \lambda_i - 1)}{\Gamma\sum_{i=1}^{k}(\alpha_i + \lambda_i - 1)} \frac{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i + \lambda_i)} \tag{17}$$

Remark(s)

I. If we put $\lambda_i = 0$, we get the measures of central tendency and dispersion of Dirichlet distribution.
II. If we put $\lambda_i = 1$, we get the measures of central tendency and dispersion of size biased or length biased Dirichlet distribution.
III. If we substitute $\lambda_i = 2$, we get measures of central tendency and dispersion of area biased or even power weighted Dirichlet distribution.

Higher Order Moments about Origin of Weighted Dirichlet Distribution

From Eq. (10), we get the first four moments about origin of weighted Dirichlet Distribution. The 3rd and 4th order moments about origin are calculated as:

$$E(\prod_{i=1}^{k} x_i^3) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i + \lambda_i + 3)}{\Gamma\sum_{i=1}^{k}(\alpha_i + \lambda_i + 3)} \frac{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i + \lambda_i)} \tag{18}$$

and

$$E(\prod_{i=1}^{k} x_i^4) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i + \lambda_i + 4)}{\Gamma\sum_{i=1}^{k}(\alpha_i + \lambda_i + 4)} \frac{\Gamma\sum_{i=1}^{k}(\alpha_i+\lambda_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i + \lambda_i)} \tag{19}$$

The other higher order moments can be obtained from Eq. (10) by substituting different values for r_i.

CHARACTERIZATION OF WEIGHTED DIRICHLET DISTRIBUTION

In this research, we propose a simple procedure to obtain n estimators of weighted Dirichlet distribution by using its characterization and moment estimation approach.

Theorem 1

If $n \geq 3$, let X_1, $X_2,\ldots,X_k$ be a set of K dimensional independent and identically distributed random vectors having the multi-nominal probability density function as:

$$f(x_i;\alpha_i,\lambda_i) = \frac{\prod_{i=1}^{k}(x_i)^{\alpha_i+\lambda_i-1}\, dx_i}{\beta(\alpha+\lambda)} ;\alpha_i > 0, \lambda_i \geq 0, 0 < x_i < 1$$

Then,

$$E(\prod_{i=1}^{k}\bar{x}_{in}^2) = \frac{\begin{pmatrix} (\alpha_i+\lambda_i)\sum_{i=1}^{k}(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1) \\ +(n-1)(\alpha_i+\lambda_i)^2\sum_{i=1}^{k}(\alpha_i+\lambda_i+1) \end{pmatrix}}{n\left(\left(\sum_{i=1}^{k}(\alpha_i+\lambda_i)\right)^2(\alpha_i+\lambda_i+1)\right)} \tag{20}$$

Proof

The above theorem can be proved considering the following step wise procedure:

Step 1: Consider the rth moments about origin of weighted Dirichlet distribution.

Step 2: Calculate the value of $E(\prod_{i=1}^{k}\bar{X}_{in})$ & $V(\prod_{i=1}^{k}\bar{X}_{in})$.

Step 3: Calculate the value of $E(\prod_{i=1}^{k}\bar{X}_{in}^2)$

Using following relation:

$$E(\prod_{i=1}^{k}\bar{X}_{in}^2) = V(\prod_{i=1}^{k}\bar{X}_{in}) + \left(E(\prod_{i=1}^{k}\bar{X}_{in})\right)^2 \tag{21}$$

After solving the above Eq. (21), we get:

$$E(\prod_{i=1}^{k}\bar{x}_{in}^2) = \frac{\begin{pmatrix} (\alpha_i+\lambda_i)\sum_{i=1}^{k}(\alpha_i+\lambda_i)(\alpha_i+\lambda_i+1) \\ +(n-1)(\alpha_i+\lambda_i)^2\sum_{i=1}^{k}(\alpha_i+\lambda_i+1) \end{pmatrix}}{n\left(\left(\sum_{i=1}^{k}(\alpha_i+\lambda_i)\right)^2(\alpha_i+\lambda_i+1)\right)}$$

Theorem 2

If $n \geq 3$, let X_1, $X_2,\ldots,X_k$ be a set of n dimensional independent and identically distributed random vectors having the multi-nominal probability density function as:

$$f(x_i;\alpha_i,\lambda_i) = \frac{\prod_{i=1}^{k}(x_i)^{\alpha_i+\lambda_i-1}\, dx_i}{\beta(\alpha+\lambda)} ;\alpha_i > 0, \lambda_i \geq 0, 0 < x_i < 1$$

Then

$$E\left(\frac{\prod_{i=1}^{k} S_n^2}{\prod_{i=1}^{k} \bar{x}_{in}^2}\right) \rightarrow \left(\frac{\sigma^2}{\mu^2}\right) \quad as\, n \rightarrow \infty \qquad (22)$$

Proof

We know that:

$$E\left(\prod_{i=1}^{k} S_n^2\right) = n\left(V(\prod_{i=1}^{k} \bar{X}_{in})\right)$$

$$E\left(\prod_{i=1}^{k} S_n^2\right) = \frac{\begin{pmatrix}(\alpha_i + \lambda_i)\sum_{i=1}^{k}(\alpha_i + \lambda_i)(\alpha_i + \lambda_i + 1) \\ -(\alpha_i + \lambda_i)^2 \sum_{i=1}^{k} (\alpha_i + \lambda_i + 1)\end{pmatrix}}{\left(\left(\sum_{i=1}^{k}(\alpha_i + \lambda_i)\right)^2 (\alpha_i + \lambda_i + 1)\right)} \qquad (23)$$

Since the random vectors are independent and identically distributed, therefore:

$$E\left(\frac{\prod_{i=1}^{k} S_n^2}{\prod_{i=1}^{k} \bar{x}_{in}^2}\right) = \left(\frac{E(\prod_{i=1}^{k} S_n^2)}{E(\prod_{i=1}^{k} \bar{x}_{in}^2)}\right) \qquad (24)$$

Using Eqs. (20) and (23) in Eq. (24), we get:

$$E\left(\frac{\prod_{i=1}^{k} S_n^2}{\prod_{i=1}^{k} \bar{x}_{in}^2}\right) = \frac{n\begin{pmatrix}(\alpha_i + \lambda_i)\sum_{i=1}^{k}(\alpha_i + \lambda_i)(\alpha_i + \lambda_i + 1) \\ -(\alpha_i + \lambda_i)^2 \sum_{i=1}^{k} (\alpha_i + \lambda_i + 1)\end{pmatrix}}{\begin{pmatrix}(\alpha_i + \lambda_i)\sum_{i=1}^{k}(\alpha_i + \lambda_i)(\alpha_i + \lambda_i + 1) \\ +(n-1)(\alpha_i + \lambda_i)^2 \sum_{i=1}^{k} (\alpha_i + \lambda_i + 1)\end{pmatrix}} \qquad (25)$$

As $n \rightarrow \infty$ in Eq. (25), we get:

$$E\left(\frac{\prod_{i=1}^{k} S_n^2}{\prod_{i=1}^{k} \bar{x}_{in}^2}\right) \rightarrow \left(\frac{\sigma^2}{\mu^2}\right) \quad as\, n \rightarrow \infty$$

ESTIMATION OF PARAMETERS

In this section, we obtain the estimates of parameters of weighted Dirichlet distribution by using method of maximum likelihood estimate.

Let $X_1, X_2, \ldots, X_n$ be an n dimensional random vectors drawn from weighted Dirichlet distribution. Then the likelihood function of weighted Dirichlet distribution is given as:

$$L(x_i;\alpha_i,\ \lambda_i) = \prod_{i}^{n}\left(\frac{\prod_{i=1}^{k}(x_i)^{\alpha_i+\lambda_i-1}}{\beta(\alpha+\lambda)}\right)$$

$$L(x_i;\alpha_i,\ \lambda_i) = \prod_{i}^{n}\left(\frac{\Gamma\sum_{i=1}^{k}\alpha_i+\lambda_i}{\prod_{i=1}^{n}\Gamma(\alpha_i+\lambda_i)}\right)\prod_{i}^{n}\left(\prod_{i=1}^{k}(x_i)^{\alpha_i+\lambda_i-1}\right)$$

Taking log on both sides, we get:

$$\log(L(x_i;\alpha_i,\ \lambda_i)) = \log\left(\prod_{i}^{n}\left(\frac{\Gamma\sum_{i=1}^{k}\alpha_i+\lambda_i}{\prod_{i=1}^{n}\Gamma(\alpha_i+\lambda_i)}\right)\prod_{i}^{n}\left(\prod_{i=1}^{k}(x_i)^{\alpha_i+\lambda_i-1}\right)\right) \qquad (26)$$

After simplifying the above Eq. (26), we have:

$$\log(L(x_i;\alpha_i,\ \lambda_i)) = n\begin{pmatrix}\left(\log(\Gamma\sum_{i=1}^{k}\alpha_i+\lambda_i)\right) - \sum_{i=1}^{k}\log(\Gamma(\alpha_i+\lambda_i)) \\ +(\alpha_i+\lambda_i-1)\dfrac{(\sum_{j=1}^{n}\sum_{i=1}^{k}\log(x_{ji}))}{n})\end{pmatrix}$$

$$\log(L(x_i;\alpha_i,\ \lambda_i)) = n\begin{pmatrix}\left(\log(\Gamma\sum_{i=1}^{k}\alpha_i+\lambda_i)\right) - \sum_{i=1}^{k}\log(\Gamma(\alpha_i+\lambda_i)) \\ +(\alpha_i+\lambda_i-1)\log(\bar{x}_k))\end{pmatrix} \qquad (27)$$

Where, $\log(\bar{x}_k) = \dfrac{(\sum_{j=1}^{n}\sum_{i=1}^{k}\log(x_{ji}))}{n}$ is called sufficient statistics. The maximum likelihood estimate can be obtained by differentiating Eq. (27) with respect to parameters and subsequently equating to zero. These equations cannot be solved analytically. But computer programs can be used to estimate the parameters numerically.

Remark(s)

If we put λ_i =0, 1, 2, the log likelihood function of weighted Dirichlet distribution reduces to log likelihood function of Dirichlet distribution, size biased or length biased Dirichlet distribution and area biased Dirichlet distribution respectively.

$$E\left(\frac{\prod_{i=1}^{k} S_n^2}{\prod_{i=1}^{k} \bar{x}_{in}^2}\right) \to \frac{\left((\alpha_i + \lambda_i)\sum_{i=1}^{k}(\alpha_i + \lambda_i)(\alpha_i + \lambda_i + 1)\right)}{\left((\alpha_i + \lambda_i)^2 \sum_{i=1}^{k}(\alpha_i + \lambda_i + 1)\right)} - 1 \quad (28)$$

From weighted Dirichlet distribution Eq. (6), the square of population co-efficient of variation can be obtained as:

$$\left(\frac{\sigma^2}{\mu^2}\right) = \frac{\left((\alpha_i + \lambda_i)\sum_{i=1}^{k}(\alpha_i + \lambda_i)(\alpha_i + \lambda_i + 1)\right)}{\left((\alpha_i + \lambda_i)^2 \sum_{i=1}^{k}(\alpha_i + \lambda_i + 1)\right)} - 1 \quad (29)$$

Comparing Eqs. (28) and (29), we have:

$$E\left(\frac{\prod_{i=1}^{k} S_n^2}{\prod_{i=1}^{k} \bar{x}_{in}^2}\right) \to \left(\frac{\sigma^2}{\mu^2}\right) \quad as\, n \to \infty$$

Thus the square of sample co-efficient of variation is asymptotically unbiased estimator of square of the population co-efficient of variation.

GENERATIVE PROCESS OF PROPOSED WEIGHTED LATENT DIRICHLET ALLOCATION

If in existing model Latent Dirichlet allocation, we change prior from simple Dirichlet distribution Dir(τ) to weighted Dirichlet prior Dir($\tau + \gamma$), then resulting model will assume following generative process to topic model a corpus D consisting of L documents each of length N based on an assumption that these documents are essentially characterized as random mixtures over hidden topics, where each topic is characterized as a multinomial distribution over vocabulary generated from the corpus.

1. Choose $\theta_l \sim Dir(\tau)$ where $l \in \{1,2,..L\}$ and $Dir(\tau)$ is a Dirichlet distribution.
2. Choose $\varphi_p \sim Dir(\tau + \gamma_2)$ where $p \in \{1,2,..p\}$ and $Dir(\tau + \lambda)$ is a Weighted Dirichlet distribution with weight $\lambda = 0,1,2,3$. Symbol p represents number of topics used to topic model corpus D.
3. For each word position (m,n), where $n \in \{1,2,..N\}$ and $m \in \{1,2,..L\}$

a. Choose a topic $z_{m,n} \sim multinomial(\theta)$

b. Choose a word $w_{m,n} \sim multinomial(\phi_{m,n})$

Weight λ depends on corpus. A proper selection of value for weight λ to topic model using weighted latent Dirichlet allocation needs to be thoroughly studied.

This generative process can is explained precisely using plate notation in Figure 5.

EXPERIMENTAL SETUP AND EVALUATION

A collection of 2410 scientific documents in LDA format suitable for experimentation vis-à-vis probabilistic topic modeling i.e. cora.documents and cora.vocab have been used in this study. cora.titles is a character vector of titles for each document (i.e., each entry of cora.documents). cora.cites is a list representing the citations between the documents in the collection.

The suitable number of topics to topic model a corpus needs to be chosen carefully. As a standard practice, corpus is iteratively modeled using varying number of topics and using metrics like CaoJuan2009, Arun2010, Griffiths2004 and Deveaud2014, optimal number of topics (k) is chosen. Figure 6 is a glimpse of the cora dataset vocabulary as a word cloud representation of the said dataset.

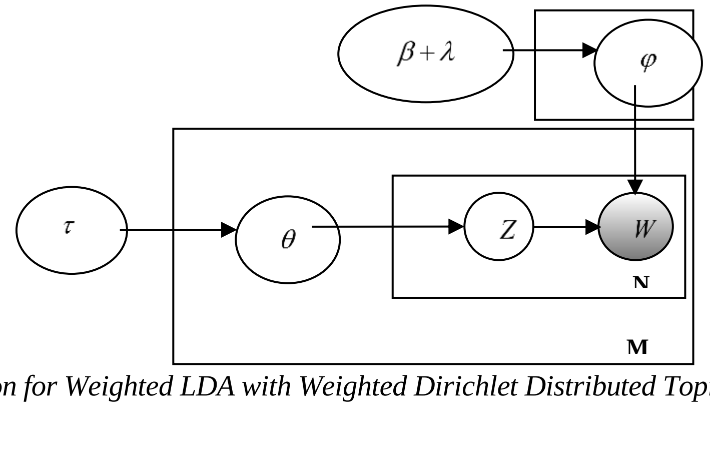

Fig. 5: *Plate Notation for Weighted LDA with Weighted Dirichlet Distributed Topic-Word Distributions*

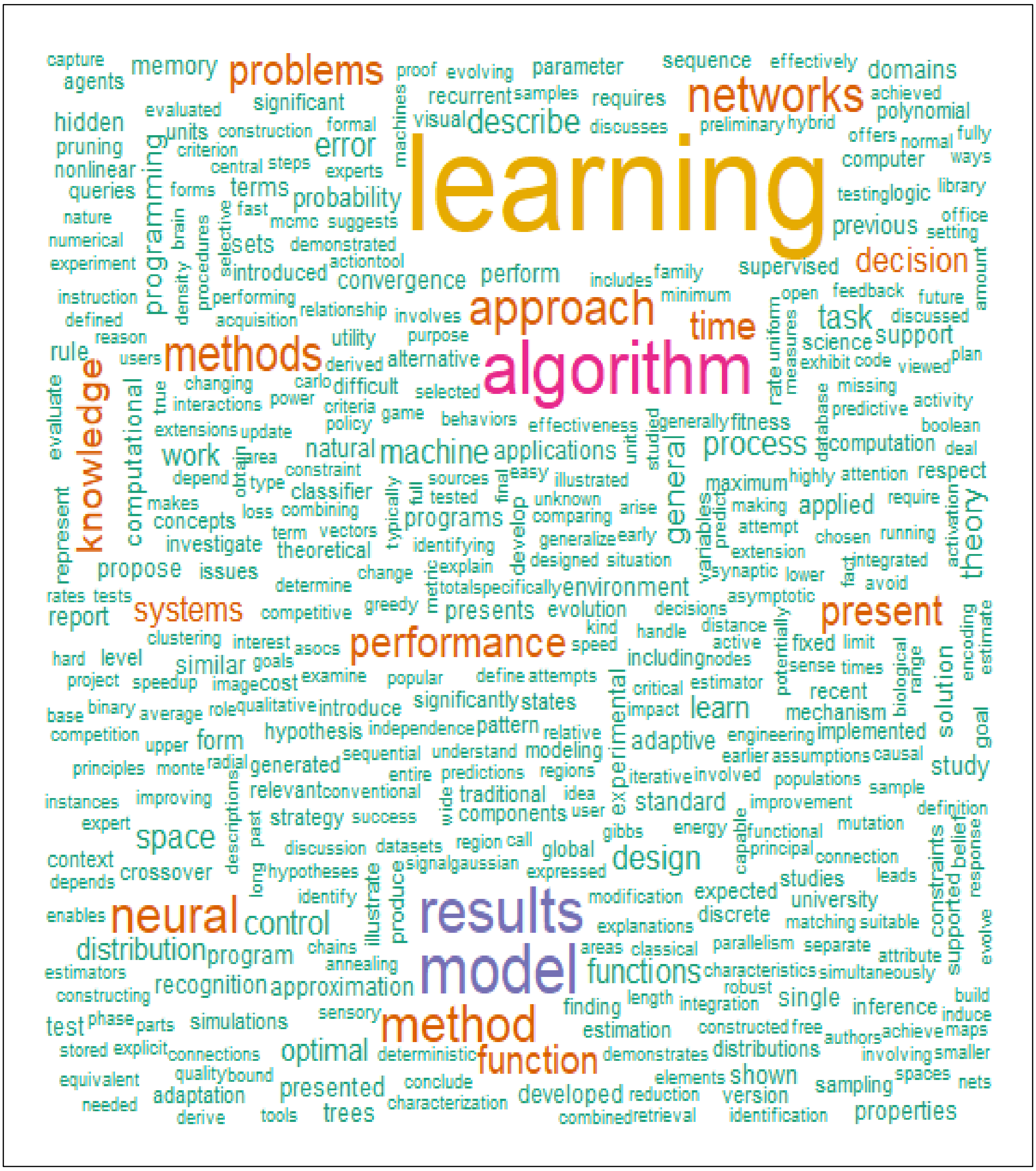

Fig. 6: *Word Cloud Representation of Vocabulary Generated from Cora Dataset.*

Table 1: *Top Five Words of Cora Dataset Topic Modeled with WLDA with No. of Topics* $k = 8$ *and hyperparameter* $\alpha = 50/8$.

Topic 1	Topic 2	Topic 3	Topic 4
"function"	"system"	"learning"	"method"
"functions"	"knowledge"	"paper"	"methods"
"algorithm"	"design"	"control"	"decision"
"optimal"	"case"	"reinforcement"	"training"
"error"	"reasoning"	"task"	"classification"
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
"genetic"	"neural"	"models"	"learning"
"search"	"network"	"research"	"algorithm"
"problem"	"networks"	"model"	"algorithms"
"programming"	"input"	"distribution"	"number"
"problems"	"model"	"analysis"	"examples"

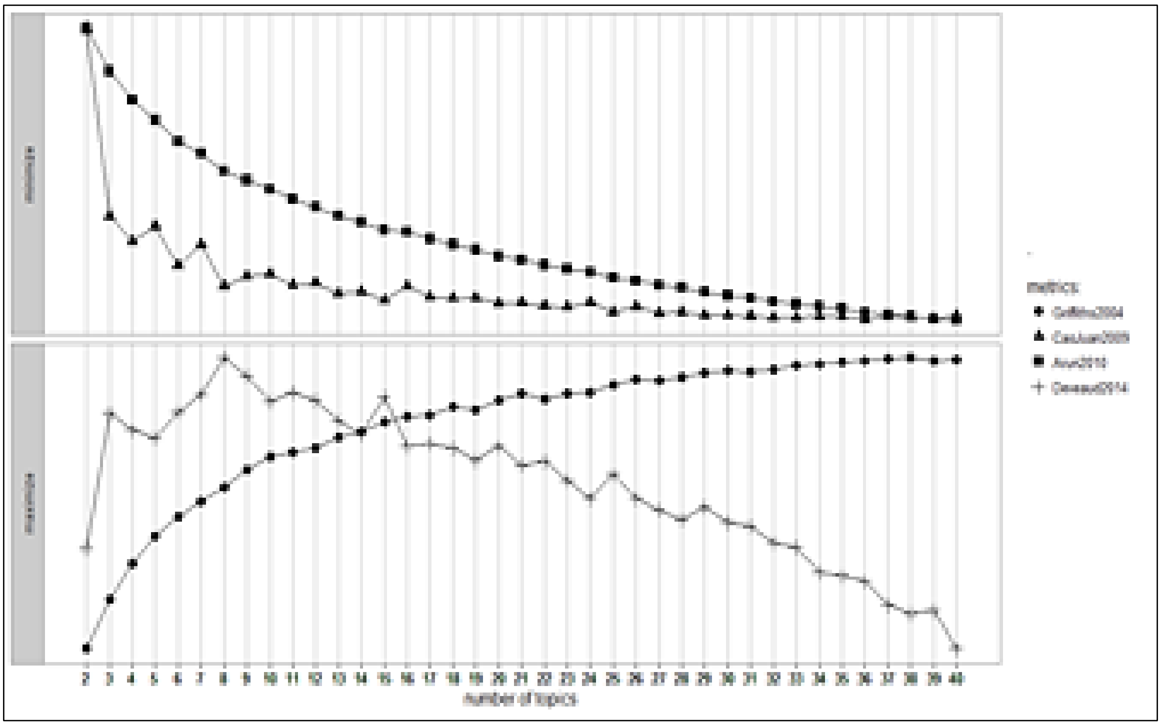

Fig. 7: *Selection of Parameter k as Number of Topics to Topic Model Corpus (Cora Dataset) Varying Number of Topics k between 2 and 40.*

In this study, topic modeling of corpus (cora dataset) was done for number of topics ranging between 2 to 40. This process took 7 h for execution on i7 machine. Among above mentioned four metrics, "CaoJuan2009" and "Arun2010" need to be evaluated for the minimum value while as "Griffiths2004" and "Deveaud2014" are evaluated for the maximum. Number of topics chosen to topic model cora dataset was chosen eight (8) (k=8) topic numbers =8 correspond to the maximum on Deveaud2014 curve with almost minimum on CaoJuan2009 for the same choice (as is evident from Figure 7). The other two metrics give a wide choice which is practically unacceptable.

Standard techniques were used to compare the weighted Dirichlet distribution with classical Dirichlet distribution. These techniques are:

i) AIC (Akaike Information Criterion) and Entropy estimation, where, $\text{AIC} = 2\text{p} - 2\log(\text{l})$ and $\hat{\text{H}} = \frac{-\log \text{l}}{\text{n}}$
ii) AICC (Akaike Information Criterion Corrected), Where $AICC = AIC + \frac{2p(p+1)}{n-p-1}$
iii) BIC (Bayesian information criterion) of a model having P parameters. $BIC = p\log(n) - 2\log(l)$

Table 1: represents five most probable words to each eight topics resulted from topic modeling of cora datset using proposed Weighted Latent Dirichlet allocation (WLDA).

RESULTS AND CONCLUSION

Table 2 acts as an acronym table for models resulted from different special cases of weighted Dirichlet distribution to topic model the corpus using different distributions which in turn result in many different topic models. These acronyms are used in Table 3 as precise reference to the models.

This study revealed that from many proposed models resulted from using weighted Dirichlet distribution to topic model cora dataset (as is evident from Tables 2 and 3), volume biased weighted Dirichlet model provides the least values of AIC, AICC and BIC values on its fitting to the said dataset. Furthermore, it is evident that AIC, AICC and BIC values decrease quite significantly, when weight parameter λ varies from 0 to 3 with parameters $\alpha = \beta = 0.3$. It is therefore concluded that the weighted volume biased Dirichlet model will be treated an improved way to topic model text repositories as is illustrated on cora dataset in this study.

Table 2: *Acronym and Their Full Forms.*

S. No	Symbol	Full Form
01	DM	Dirichlet Model
02	SBDM	Size Biased Dirichlet Model
03	ABDM	Area Biased Dirichlet Model
04	VBDM	Volume Biased Dirichlet Model

Table 3: *Comparison of Different Models of Weighted Dirichlet Distributions with Dirichlet Model.*

Model	α	β	λ	Log likelihood	AIC	BIC	CAIC	H^
DM	0.5	0.3	0	–972674.4	1945354.8	1945359.214	1945354.804	328.495238
SBDM	0.5	0.3	1	–970894.9	1941795.8	1941800.214	1941795.804	327.894259
ABDM	0.5	0.3	2	–970909.1	1941824.2	1941828.614	1941824.204	327.899054
VBDM	0.5	0.3	3	–967706.9	1935419.8	1935424.214	1935419.804	326.817595
DM	1	0.3	0	–972873.8	1945753.6	1945758.014	1945753.604	328.56258
SBDM	1	0.3	1	–971015.5	1942037	1942041.414	1942037.004	327.934988
ABDM	1	0.3	2	–968420.2	1936846.4	1936850.814	1936846.404	327.058494
VBDM	1	0.3	3	–965741.5	1931489	1931493.414	1931489.004	326.153833
DM	1.5	0.3	0	–970616	1941238	1941242.414	1941238.004	327.800068
SBDM	1.5	0.3	1	–969651.8	1939309.6	1939314.014	1939309.604	327.474434
ABDM	1.5	0.3	2	–970049.6	1940105.2	1940109.614	1940105.204	327.608781
VBDM	1.5	0.3	3	–969845.4	1939696.8	1939701.214	1939696.804	327.539818
DM	2	0.3	0	–971794.8	1943595.6	1943600.014	1943595.604	328.198176
SBDM	2	0.3	1	–968373	1936752	1936756.414	1936752.004	327.042553
ABDM	2	0.3	2	–966237.2	1932480.4	1932484.814	1932480.404	326.321243
VBDM	2	0.3	3	–967452.3	1934910.6	1934915.014	1934910.604	326.731611
DM	2.5	0.3	0	–970729.7	1941465.4	1941469.814	1941465.404	327.838467
SBDM	2.5	0.3	1	–968373.9	1936753.8	1936758.214	1936753.804	327.042857
ABDM	2.5	0.3	2	–967138.7	1934283.4	1934287.814	1934283.404	326.625701
VBDM	2.5	0.3	3	–966244	1932494	1932498.414	1932494.004	326.323539
DM	3	0.3	0	–969174.2	1938354.4	1938358.814	1938354.404	327.313137
SBDM	3	0.3	1	–966632	1933270	1933274.414	1933270.004	326.454576
ABDM	3	0.3	2	–967053	1934112	1934116.414	1934112.004	326.596758
VBDM	3	0.3	3	–968117.4	1936240.8	1936245.214	1936240.804	326.956231

REFERENCES

1. Griffiths TL, Steyvers M. A Probabilistic Approach to Semantic Representation. *Proc 24th Annu Conf Cogn Sci Soc.* 2002; 381–386p.
2. Griffiths TL, Steyvers M, Blei DM, *et al.* Integrating Topics and Syntax. *Adv Neural Inf Process Syst.* 2005; 17: 537–544p.
3. Griffiths TL, Steyvers M. Finding Scientific Topics. *Proc Natl Acad Sci.* 2004; 101(Supplement 1): 5228–5235p.
4. Hofmann T. Probabilistic Latent Semantic Analysis. *Proc 15th Conf Uncertain Artif Intell.* 1999; 289–296p.
5. Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Mach Learn.* 2001; 42(1–2): 177–196p.
6. Johnson Norman L, Samuel Kotz, Balakrishnan N. *Continuous Multivariate Distributions, Volume 1, Models and Applications*. Vol. 59. New York: John Wiley & Sons; 2002.

7. Blei David M, Ng Andrew Y, Jordan Michael I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. Jan (2003): 993-1022.
8. Dahl David B. Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. *Bayesian Inference for Gene Expression and Proteomics*. 2006; 201–218p.
9. Ishwaran Hemant, James Lancelot F. Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information. *J Comput Graph Stat*. 11.3 (2002): 508-532.
10. Arif Wani M, Batchelor Bruce G. Edge-Region-Based Segmentation of Range Images. *IEEE Trans Pattern Anal Mach Intell*. 1994; 16(3): 314–319.
11. Arif Wani M, Batchelor Bruce G. Heuristic Segmentation of Range Images. *Intelligent Robots and Computer Vision X: Algorithms and Techniques*. Vol. 1607. International Society for Optics and Photonics; 1992.
12. Arif Wani M, Batchelor Bruce G. Two-Dimensional Boundary Inspection Using Autoregressive Model. *High-Speed Inspection Architectures, Barcoding, and Character Recognition*. Vol. 1384. International Society for Optics and Photonics; 1991.
13. Bhat Farooq Ahmad, Arif Wani M. Performance Comparison of Major Classical Face Recognition Techniques. *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE; 2014.
14. Bhat Muzafar Rasool, Arif Wani M. Evaluating Algebraic Model Based Information Retrieval Algorithms for Small and Large Data Set. *Computing for Sustainable Global Development (INDIACom), 2017 4th International Conference on*. IEEE; 2017.
15. Bhat Muzafar Rasool, Arif Wani M. Selecting Appropriate Number of Singular Values for Latent Semantic Indexing in Information Retrieval. *Recent Trends and Advancements in Engineering and Technology 2016, 4th international conference on. ICRTAET*. 2016.
16. Rasool Bhat Muzafar, Arif Wani M. Mixture Weighted Latent Dirichlet Allocation, an Optimized and Generalized Probabilistic Model for Large Corpus of Data. *Artificial Intelligent Systems and Machine Learning* 10.1 (2018): 8-17.
17. Wani MA, Afzal S. A New Framework for Fine Tuning of Deep Networks. *16th IEEE International Conference on Machine Learning and Applications*. 2017; 359–363p.
18. Wani MA. Incremental Hybrid Approach for Microarray Classification. *Proceedings of the Seventh International Conference on Machine Learning and Applications*. 2008; 514–520p.
19. Wani MA. Microarray Classification Using Sub-Space Grids. *Proceedings of the Tenth International Conference on Machine Learning and Applications*. 2011; 1: 389–394p.
20. Wani MA. Introducing Subspace Grids to Recognise Patterns in Multidimensinal Data. *International Conference on Machine Learning and Applications*. 2012; 1: 33–39p.
21. Wani MA, Yesilbudak M. Recognition of Wind Speed Patterns Using Multi-Scale Subspace Grids with Decision Trees. *International Journal of Renewable Energy Research (IJRER)*. 2013; 3(2): 458–462p.
22. Wani MA. SAFARI: A Structured Approach for Automatic Rule. *IEEE Trans Syst Man Cybern B (Cybernetics)*. 2001; 31(4): 650–657p.
23. Fisher RA. The Effect of Methods of Ascertainment Upon the Estimation of Frequencies. *Ann Eugen.* 1934; 6(1): 13–25p.
24. Radhakrishna Rao C. On Discrete Distributions Arising Out of Methods of Ascertainment. *Sankhyā: The Indian Journal of Statistics, Ser A*. 1965; 27(2–4): 311–324p.
25. Shaban SA, Naima Ahmed Boudrissa. The Weibull Length Biased Distribution Properties and Estimation. MSC Primary: 60E05, Secondary: 62E15, 62F10 and 62F15. 2000.
26. Das Kishore K, Tanusree Deb Roy. On Some Length-Biased Weighted Weibull Distribution. *Pelagia Research Library, Adv Appl Sci Res*. 2011; 2(5): 465–475p.
27. Oluyede Broderick O, Olusegun George E. On Stochastic Inequalities and Comparisons of Reliability Measures for Weighted Distributions. *Math Probl Eng.* 2002; 8(1): 1–13p.

28. Ghitany ME, Al-Mutairi DK. Size-biased Poisson-Lindley Distribution and its Application. *Metron-International Journal of Statistics*. 2008; 66(3): 299–311p.
29. Oluyede Broderick O, Mekki Terbeche. On Energy and Expected Uncertainty Measures in Weighted Distributions. *Int Math Forum*. 2007; 2(20): 947–956p.
30. Reshi JA, Ahmed A. Characterization, Reliability and Information Measures of Even-Power Weighted Generalized Gamma Distribution. *J Stat Appl Probab Lett (JSAPL)*. 2015; 2(3): 247–262p.
31. Reshi JA, Ahmed A. Characterization and Estimations of Weighted Generalized Beta Probability Distributions. *J Stat Appl Probab*. 2015; 4(3): 513p.

Cite this Article

Muzafar Rasool Bhat, Arif Wani M. Weighted Latent Dirichlet Allocation, an Optimized and Generalized Probabilistic Topic Model for Large Corpus of Data. *Journal of Artificial Intelligence Research & Advances*. 2018; 5(1): 18–31p.