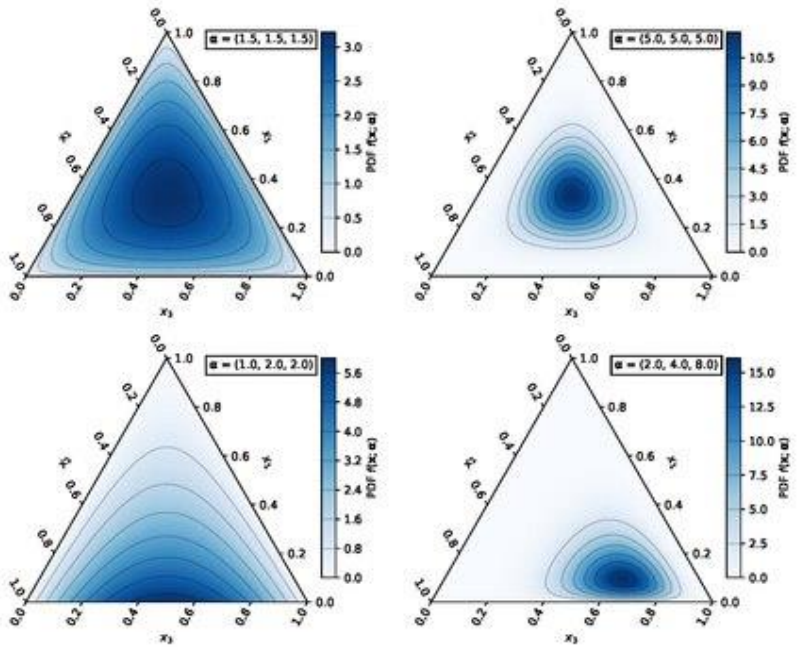


The Dirichlet Distribution and Latent Dirichlet Allocation

CRISTIAN MOJICA

4/12/22

STAT 676

IRP

Dirichlet Distribution Properties

- Dirichlet distribution: $p(\theta \mid \boldsymbol{\alpha}) = \frac{\Gamma(a_1+\cdots+a_k)}{\Gamma(a_1)\cdots\Gamma(a_k)} \theta_1^{a_1-1} \cdots \theta_k^{a_k-1} = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i-1}$
  - $p(\theta \mid \boldsymbol{\alpha}) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j-1}$
  - Continuous; returns a vector of real numbers
  - Specifications: $\theta_1, \ldots, \theta_k \geq 0; \sum_{j=1}^{k} \theta_j = 1$
  - K = 2, θ_1 = x ? Beta distribution!
- What is θ?
  - Vector of probabilities associated with distinct categories
  - Formally, does not have density
- What is α?
  - Concentration parameter, one value per θ_j
  - Pseudocounts
  - Regardless, each value of α is greater than 0
- Shape of distribution: K-dimensional density, with support being a (K-1)-dimensional simplex

The type 1 integrals are given by

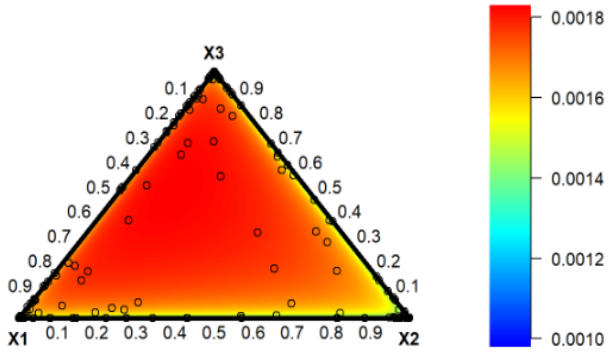
$$I \equiv \iint \cdots \int f(t_1 + t_2 + \ldots + t_n) t_1^{\alpha_1 - 1} t_2^{\alpha_2 - 1} \cdots t_n^{\alpha_n - 1} \, dt_1 \, dt_2 \, dt_n$$

$$= \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)}{\Gamma\left(\sum_n \alpha_n\right)} \int_0^1 f(\tau) \tau^{(\Sigma_n \alpha)-1} \, d\tau,$$

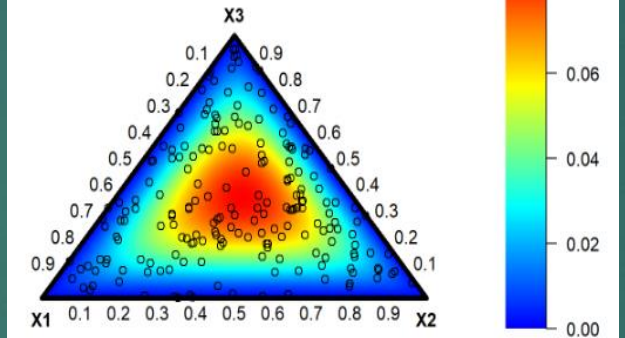
Where did the Dirichlet distribution come from?

Contours of Various Dirichlet Distributions

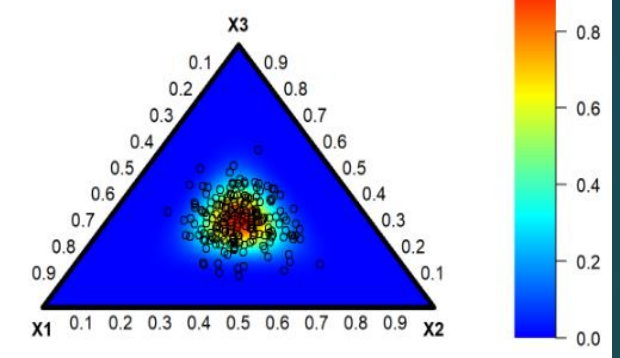
```
draws <- rdirichlet(200, c(.1,.1,.1) )
bivt.contour(draws)
```



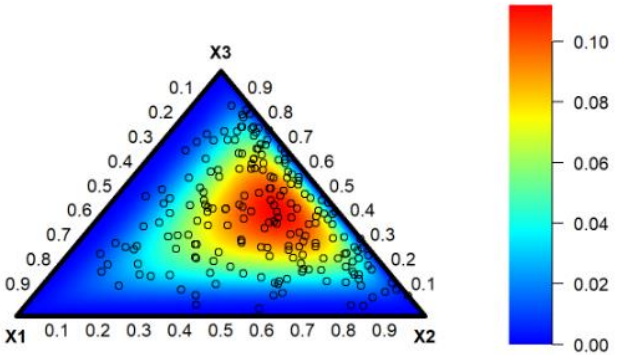

```
draws <- rdirichlet(200, c(1,1,1) )
bivt.contour(draws)
```



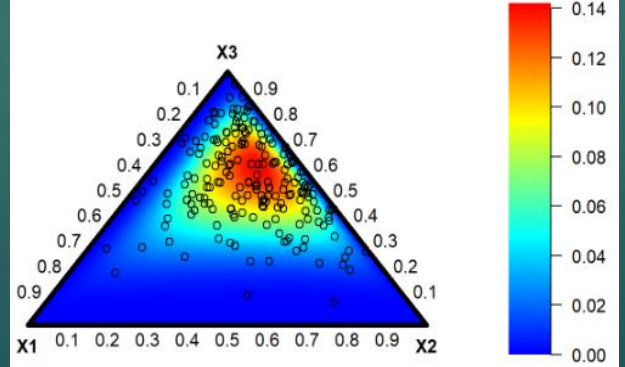

```
draws <- rdirichlet(200, c(10,10,10) )
bivt.contour(draws)
```



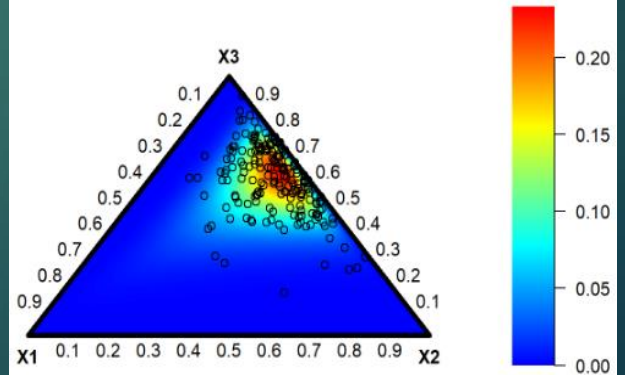

```
draws <- rdirichlet(200, c(1, 2, 2) )
bivt.contour(draws)
```




```
draws <- rdirichlet(200, c(1, 2, 4) )
bivt.contour(draws)
```




```
draws <- rdirichlet(200, c(1, 4, 8) )
bivt.contour(draws)
```



Other Properties

- *Note: a_0 is the sum of the a_i's.*
- Mean: $E(\theta_j) = \frac{a_j}{a_0}$
- Variance: $\mathrm{Var}(\theta_j) = \frac{a_j(a_0 - a_j)}{a_0^2(a_0+1)}$
- Covariance: $\mathrm{cov}(\theta_i, \theta_j) = -\frac{a_i a_j}{a_0^2(a_0+1)}$
- Mode: $\mathrm{mode}(\theta_j) = \frac{a_j - 1}{a_0 - k}$
- Marginal distributions: $X_i \sim Beta(\alpha_i,\ \alpha_0 - \alpha_i)$

Use in Bayesian Inference

- Recall: Multinomial distribution pmf: $p(\theta) = \binom{n}{\theta_1 \theta_2 \cdots \theta_k} p_1^{\theta_1} \cdots p_k^{\theta_k}$
  - Specifications: $\theta_j = 0, 1, 2, \cdots, n; \sum_{j=1}^{k} \theta_j = n$
- Dirichlet-Multinomial Model
  - $p(\theta|a) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$
  - The Dirichlet distribution is conjugate for the multinomial sampling model. This also works for the categorical distribution (multinomial: k > 2, one trial).

$$\pi(p_1, \cdots, p_k) \propto \prod_{i=1}^{k} p_i^{x_i} \prod_{i=1}^{k} {p_i}^{\alpha_i - 1} \propto \prod_{i=1}^{k} p_i^{\alpha_i + x_i - 1}$$

Latent Dirichlet Allocation (LDA)

- LDA, not to be confused with linear discriminant analysis (a dimension reduction technique).
- In text classification, LDA is an algorithm utilized to sort through documents in order to determine the latent topics of said documents.
- Part of the algorithm is a hierarchical model, and certain variables within it have Dirichlet priors.

- Consider a collection, or corpus, of documents. Each document is about some topic, although sometimes a document may discuss multiple topics.
- Assume we don't have access to the documents' topics. What are the topics of the documents? Assume that a document can only have one topic.
- How will a computer determine topics for these documents?

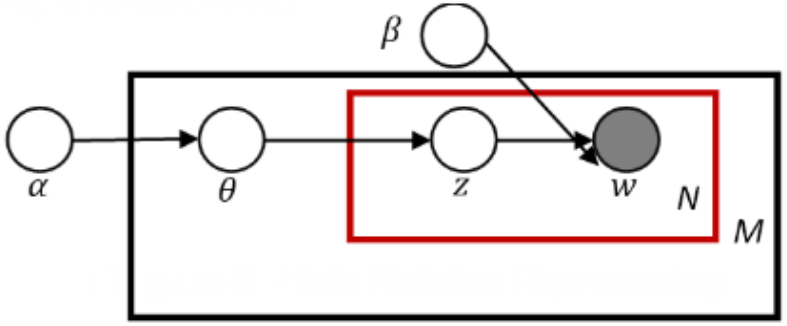


Figure 3: Plate Notation Representing LDA

How Does It Work?: The LDA Algorithm

With plane notation, the dependencies among the many variables can be captured concisely. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document. Thus:

α is the parameter of the Dirichlet prior on the per-document topic distributions,

β is the parament of the Dirichlet prior on the per-topic word distribution,

θ_m is the topic distribution for document m.

φ_k is the word distributed for the k,

z_{mn} is the topic for the nth word in document m, and

w_{mn} is the specific word.

The w_{ij} are the only observable variables, and the other variables are latent variables. Mostly, the basic LDA model will be extended to a smoothed version to gain better results.

The generative process (algorithm):

1. Choose $\theta_i \sim Dir(\alpha)$, where $i \in \{1, \ldots, M\}$ and $Dir(\alpha)$ is the Dirichlet distribution for parameter α
2. Choose $\varphi_k \sim Dir(\beta)$, where $k \in \{1, \ldots, K$
3. For each of the word position i, j, where $j \in \{1, \ldots, N_i$, and $i \in \{1, \ldots, M\}$
   a. Choose a topic $z_{ij} \sim Multinomial\ (\theta_i)$
   b. Choose a word $w_{ij} \sim Multinomial\ (\varphi_{z_{ij}})$

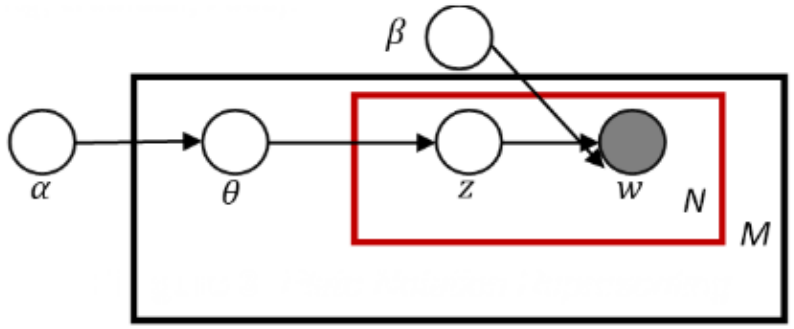


Figure 3: Plate Notation Representing LDA

Rommel Verecio: “Applications of Latent Dirichlet Allocation Algorithm of Published Articles on Cyberbullying”

- Prominence of cyberbullying with advances in Internet communication
- LDA as unsupervised machine learning
- Notice: five latent themes

Abstract

With huge information available regarding traditional forms of bullying, (i.e.,verbal, physical, relational), cyberbullying research is only recently beginning to thrive. To determine cyberbullying incidence, it is vital to dig available documents of the top ten countries which take the most incidence in cyberbullying as shown in Google trends for the past ten years with published papers in google scholar. Hence, this study utilizes a new dimension of research adopting web mining technique and topic modeling mainly unsupervised machine learning with the application of Latent Dirichlet Allocation Algorithm in the content analysis of the published articles available online. Five Latent themes were identified by the researcher points of view which based on the result generated by the R-programming software and supported through various literature, analysis, and discussions.

Specifically, this seeks answers to the following questions:

1. What are the countries that are active on the web related to cyberbullying?
2. What are the topics and its trends related to cyberbullying for the last five years?
3. What are the latent themes generated from the online documents?
4. What recommendations can be derived based on the findings of the study?

Findings

Top Ten Countries with Documents Published Related to Cyberbullying from April 2012 to March 2017

Ranking	Country	# of Documents Published
1	Philippines	100
2	Singapore	46
3	New Zealand	44
4	Australia	41
5	United States	34
6	Canada	27
7	United Kingdom	26
8	Ireland	23
9	South Africa	21
10	Portugal	19

Source: Google Trends

Answering the first question: countries most active in discussing cyberbullying

Table 1: LDAGibbs Documents to Topics

#	Document	Topic
1	BullyingCyberbullyingandSuicide.txt	3
2	CyberbullyingBehaviorsamongMidlleandHighSchoolStudents.txt	2
3	CyberbullyinginSouthAfricaImpactandResponses.txt	2
4	CyberbullyingtheNewFaceofWorkplacebullying.txt	4
5	CyberVictimizationandCyberAggressionamongPortugueseAdolescents.txt	4
6	Exploring traditional and cyberbullying among Irish adolescents.txt	2
7	ExploringtheConsequencesofBullyingVictimizationinaSampleofSingaporeYouth.txt	3
8	SocialMediaasaChannelanditsImplicationsonCyberBullying.txt	1
9	Suicide onlinePortrayalofWebsiteRelatedSuicidebytheNewZealandMedia.txt	1
10	TheEmotionalImpactofBullyingandCyberbullyingonVictimsAEuropeanCrossNationalStudy.txt	5

We have documents! But what are the five topics lurking, latently?

Table 2: Sample Terms and Frequencies

Terms	Freq	Terms	Freq	Terms	Freq	Terms	Freq
Bullying	825	emotional	85	forms	58	another	45
cyberbullying	368	ideation	85	hinduja	58	well	45
Cyber	341	table	85	analysis	57	affected	45
School	262	parents	84	two	57	responses	45
Online	231	results	83	victim	57	first	44
Victims	210	respondents	81	experiences	56	number	44
Internet	201	one	80	behaviours	56	types	44
Suicide	201	among	79	aggression	55	users	44
reported	180	risk	78	age	54	acts	43
Study	168	significant	77	experience	54	environment	43
traditional	166	related	75	email	53	phones	43
victimization	164	used	74	life	53	someone	43
media	155	girls	73	three	53	direct	42
research	143	boys	71	schools	52	important	42
children	140	reporting	71	time	51	low	42
suicidal	139	patchin	70	associated	50	via	42
social	129	behaviour	69	report	50	years	42
students	124	studies	68	variables	50	questions	41
news	112	using	67	due	49	safety	41
family	111	differences	66	findings	49	suicides	41
bullied	105	information	66	data	48	university	41
health	105	support	66	high	48	included	40
mobile	100	young	66	home	48	less	40
negative	96	different	63	access	48	prevention	40
technology	95	others	62	need	48	world	40
sample	89	workplace	62	issue	47	consequences	39
youth	89	phone	61	adolescents	46	involved	39
impact	88	gender	60	experienced	46	messages	39
found	86	however	59	individuals	46	type	39
people	86	participants	59	way	46	communication	39

120 words from the ten documents and their frequencies are shown here. This is a sample from 5,028 terms.

Table 3: LDAGibbs Topics to Terms

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	suicide	cyber	bullying	cyberbullying	bullying
2	media	bullying	school	family	victims
3	news	online	victimization	negative	cyberbullying
4	social	children	suicidal	workplace	emotional
5	internet	school	ideation	respondents	reported
6	reporting	health	youth	support	impact
7	online	study	patchin	bullying	different
8	suicides	risk	bullied	acts	traditional
9	research	young	online	results	found
10	reported	girls	students	study	affected
11	information	internet	research	involvement	direct
12	users	traditional	traditional	rules	internet
13	used	people	hinduja	environment	mobile
14	websites	behaviours	sample	parents	results
15	zealand	life	mobile	email	types
Latent Themes	Role of Social Networking Sites and Media in Reporting Suicidal Issues	Studies on health risks and behaviors of young adults in using the internet	Mobile device as a tool for cyberbullying in school	Family support against cyberbullying	Impact and emotional stress brought by cyberbullying

Word groupings identified by the algorithm per topic; latent themes presented here

Table 4: LDAGibbs Topic Probabilities

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	0.046296296	0.063271605	**0.708333333**	0.085493827	0.096604938
2	0.013213213	**0.557657658**	0.312312312	0.060660661	0.056156156
3	0.049873592	**0.777752241**	0.068490002	0.036543323	0.067340841
4	0.012458074	0.145663632	0.077144226	**0.691423095**	0.073310973
5	0.022667543	0.07260184	0.059461235	**0.74934297**	0.095926413
6	0.013352408	**0.568431092**	0.13018598	0.130662852	0.157367668
7	0.015975336	0.060257848	**0.812780269**	0.034473094	0.076513453
8	**0.573858115**	0.274538387	0.044217687	0.06462585	0.042759961
9	**0.802060738**	0.073752711	0.037689805	0.048806941	0.037689805
10	0.013426037	0.031167586	0.065691681	0.041237113	**0.848477583**

Latent Themes	Role of Social Networking Sites and Media in Reporting Suicidal Issues	Studies on health risks and behaviors of young adults in using the internet	Mobile device as a tool for cyberbullying in school	Family support against cyberbullying	Impact and emotional stress brought by cyberbullying

Consistency and reliability of topics using Gibbs sampler per documents and topics

Summary

- (1) What is it?
  - The Dirichlet distribution is a new distribution for us! It comes from the Dirichlet integral of type I.
  - LDA: Popular algorithm for topic modeling in text analytics
- (2) How does it work?
  - The Dirichlet Distribution is a multivariate generalization of the beta distribution and is used as the prior to a multinomial sampling model.
  - LDA: Sample words and topics, with their hyperparameters being Dirichlet distributed, and words and topics each themselves being multinomial distributed. Bring these together to form documents and bring documents together to form a corpus.
- (3) Why is it cool?
  - This allows us to be able to look at multinomial sampling models and to consider them in Bayesian inference.
  - It sets a prior on unknown distributions. It is a distribution of distributions! This is nonparametric work!
  - LDA is an unsupervised methodology for topic modeling that can allow researchers to identify trends, pertinent/sensitive information in a corpus of documents.
- (4) What are some drawbacks/limitations?
  - The discrete nature of the parameters for certain aspects of hierarchical modeling might make this harder to encode in probabilistic programming software such as Stan.
  - Creating Gibbs samplers and hierarchical models requires the derivation of full conditional distributions, which can be tedious and often requires knowledge of calculus and probability theory.
  - LDA: You need to pre-specify a number of clusters/topics in advance, and humans still need to verify/identify latent themes

References

- Carlin, Bradley P., and Thomas A. Louis. Bayesian methods for data analysis. Third edition. CRC Press, 2009.
- Gelman, Andrew, et al. Bayesian data analysis. Second edition. Chapman and Hall/CRC, 2004.
- Hou, Janpu. “Dirichlet Distribution Example.” RPubs, 27 July 2017, https://rpubs.com/JanpuHou/295096.
- Verecio, Rommel L. "Applications of latent Dirichlet allocation algorithm of published articles on cyberbullying." *International Journal of Applied Engineering Research* 12.21 (2017): 10878-10884. (***)
- https://archive.lib.msu.edu/crcmath/math/math/d/d273.htm
- https://en.wikipedia.org/wiki/Categorical_distribution
- https://en.wikipedia.org/wiki/Dirichlet_distribution#cite_note-devroye-16
- https://www.youtube.com/watch?v=T05t-SqKArY (***)
- https://www.youtube.com/watch?v=BaM1uiCpj_E (***)

Thank you!