On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations

R. Arun, V. Suresh, C.E. Veni Madhavan, and M. Narasimha Murty

Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India
{arun_r,vsuresh,cevm,mnm}@csa.iisc.ernet.in

Abstract. It is important to identify the "correct" number of topics in mechanisms like Latent Dirichlet Allocation(LDA) as they determine the quality of features that are presented as features for classifiers like SVM. In this work we propose a measure to identify the correct number of topics and offer empirical evidence in its favor in terms of classification accuracy and the number of topics that are naturally present in the corpus. We show the merit of the measure by applying it on real-world as well as synthetic data sets(both text and images). In proposing this measure, we view LDA as a matrix factorization mechanism, wherein a given corpus C is split into two matrix factors M_1 and M_2 as given by $C_{d*w} = M1_{d*t}$ x Q_{t*w}. Where d is the number of documents present in the corpus and w is the size of the vocabulary. The quality of the split depends on "t", the right number of topics chosen. The measure is computed in terms of symmetric KL-Divergence of salient distributions that are derived from these matrix factors. We observe that the divergence values are higher for non-optimal number of topics – this is shown by a 'dip' at the right value for 't'.

Keywords: LDA Topic SVD KL-Divergence.

1 Introduction

Topic Modelling is a widely used technique in information retrieval, data mining etc. The idea behind it is the fact that a small number of latent topics are enough to effectively represent a large corpus. As this is often the case with real world corpus such as text which have a large vocabulary, such models have proved to be very effective. However finding the right number of latent topics in a given corpus has remained an open ended question. Almost all previous methods including Latent Semantic Analysis [1], Probabilistic Latent Semantic Analysis [2], Latent Dirichlet Allocation [3], Non-Negative Matrix Factorization [4] which try to model the latent topics either as probability distributions or as a set of basis vectors in the topic space make the implicit assumption that the number of topics is known beforehand. While estimating the right number of topics for a small image or text corpus might seem easy, it becomes unreasonable to guess the same when the corpus size is huge. However the accuracy of all of the above mentioned methods is senstive to the number of topics.

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 391–402, 2010.

In this paper, we consider the Latent Dirichlet Allocation (LDA) [3] model as the basis for our work. We view LDA as a matrix factorization method which factorizes a document-word frequency matrix M into two matrices $M1$ and $M2$ of order $T*W$ and $D*T$ respectively where T is the number of topics and W is the size of the vocabulary of the corpus. We propose a new measure that computes the symmetric Kullback-Leibler divergence of the Singluar value distributions of matrix $M1$ and the distribution of the vector $L*M2$ where L is a $1*D$ vector containing the lengths of each document in the corpus. We show that under certain conditions these distributions are comparable and these conditions are expected to determine the 'right' number of topics. We also present empirical results that indicate that the proposed measure dips down and hits a low for the 'right' number of topics and increases again as the number of topics increase. The number of topics that is considered 'right' is any number in a small range that gives the best accuracy on a held out dataset.

This work is organized into the following sections: In Section 2, we review some related work in topic modelling and some methods proposed to choose the 'right' number of topics. In section 3, we motivate the rationale behind the measure proposed and explain how it is computed. In section 4 , we give experimental evidence to illustrate the robustness of the measure across text and image corpus. Finally we conclude in section 5 with a few points of discussion.

2 Background

2.1 Latent Dirichlet Allocation

LDA is a probabilistic generative model which assumes that every document is a distribution over topics and every topic is a distribution over words. Each word in a document is generated by first sampling a topic from the topic-distribution associated with the document and then sampling a word from the word distribution associated with the topic. Thus, given a corpus, LDA tries to find the right assignment of topic to every word such that the parameters of the generative model are maximized.

Topic Similarity. There have been a couple of approaches in the past which have tried to take advantage of the fact that the topics arising in read world data are correlated. Correlated Topic Models [12] is one such approach which tries to capture relation between topics using a covariance matrix. The Pachinko Allocation Model [14] on the other hand considers an acyclic graph where a topic is a node and is considered as a distribution over not only words but also other topics.

There have also been approaches like Hierarchical Dirichlet Process (HDP) [11] which try to find the right number of topics by assuming that the data has a hierarchical structure to it. Here, both HDP as well as LDA models for the same dataset are built and compared to find the right number of topics.

More recently [10] proposes a method to learn the right size of an ontology by measuring the change in the average cosine distance between topics found as

the number of topics increase. A similar idea is found in [5] where the average correlation between $\binom{n}{2}$ pairs of topics at each stage is considered.

The disadvantage we feel with both [10] and [5] is that they consider only the information in the stochastic topic-word $T*W$ matrix and ignore the document-topic $D*T$ matrix. In the present work, we make use of properties of both these matrices to come up with a robust measure that will help in identifying the right number of topics.

3 Proposed Measure

3.1 Matrix Row Sums

Though LDA is a probabilistic generative model, it can be viewed as a non-negative matrix factorization method that breaks a given Document-Word Frequency Matrix M into a Topic-Word matrix $M1$ of order $T*W$ and a Document-Topic matrix $M2$ of order $D*T$ where D, T and W represent the number of documents, topics and words respectively. Both $M1$ and $M2$ are stochastic matrices where the k th row in $M1$ is a distribution over words in the k^{th} topic and n^{th} row in $M2$ is a distribution of topics in the n^{th} document. If these were not stochastic matrices, but just represented counts i.e if the $(i,j)^{th}$ element in matrix $M1$ indicated the number of the times word j has been assigned topic i and if the (i,j) th element in matrix $M2$ indicated the number of times topic j is assigned to a word in document i, then it is clear that

$$\sum_{v=1}^{W} M1(t,v) = \sum_{d=1}^{D} M2(d,t) \ \forall t = 1 \ to \ T \ .$$

This is nothing but the number of words assigned to each topic looked in two different ways - one as row sum over words and other as column-sum over documents. However, when both these matrices are row normalized (as done by LDA), this equality will not hold anymore.

The idea behind the proposed measure is to take advantage of the simple fact that both these sums represent proportion of topics assigned to the corpus and hence can be compared with each other. However, a mere comparison between these values is useless as they always will be the same irrespective of the number of topics considered. Hence, we seek a measure which while trying to compare similar properties of these matrices will also be low only when the 'right' number of topics is reached.

3.2 Distribution over Singular Values

Singular Value Decomposition (SVD) [8] is a matrix factorization technique that breaks (uniquely) any rectangular matrix M of order $m*n$, $m <= n$ into three matrices U , Σ *and* V or orders $m*m$, $m*n$ and $n*n$ respectively such that

$$M = U * \Sigma * V^{'}$$

where $V^{'}$ denotes the transpose of V. The matrix Σ is diagonal matrix and the matrices U *and* V are unitary. Also U contains the eigenvectors of matrix $M*M^{'}$ and V contains the eigenvectors of the matrix $M^{'}*M$.

The diagonal entries of Σ are called the singular values , denoted by σ_i , $i = 1$ *to* m.

Geometrically, the singular values represent the length of the semi-axes of the hyper-ellipsoid that encloses all the vectors in the matrix. So, a distribution of these singular values will give the distribution of the variance in direction of each axis of the hyper-ellipsoid. Now, if we consider the $T * W$ matrix $M1$, its distribution of singular values will be the distribution of variance in topics.

For simplicity, let us consider a case where the words in the vocabulary set are *well separated* in $M1$. By this we mean that the words in the vocabulary are partitioned into T sets V_i, $i = 1$ *to* T such that $V_i \cap V_j = \emptyset$ when $i \neq j$. Now if each topic T_i (a row in matrix $M1$) contains words only from set V_i, then the following proposition holds for $M1$

Proposition: If the words in the vocabulary are well separated, then the singular value σ_i *is equal to the* L_2 *norm of row-i vector of* $M1$, $\forall\ i = 1$ *to* m

Proof: This is easy to see both geometrically and algebraically. Geometrically, we observe that when the rows are well separated, the row vectors are orthogonal to each other. Thus, the axes of the hyper-ellipsoid will be be row vectors themselves which means the singular values will be their distance from origin, or in other words the L_2 norm.

Algebraically, we see that for any well separated matrix $M1$, the matrix $(M1) * (M1)^{'}$ will be a diagonal matrix with the standard basis as the eigenvectors. Thus if SVD of $M1 = U * \Sigma * V^{'}$, then as columns of U are the eigenvectors of $(M1) * (M1)^{'}$ i.e the standard basis. Hence the $(i,j)^{th}$ entry $(i, j = 1$ to $m)$ in V is given by

$$(V)_{ij} = (M1)_{ij}/\sigma_i$$

As the matrix V is a unitary matrix, we need $V * V^{'}$ to be I and equating each element of $V * V^{'}$ to elements of I proves the proposition.

Thus if in the ideal case, when the topics are well separated, the Singular values will equal the row L_2 norms. Also note that the distribution over singular values will not change by making the matrix stochastic if and only if the number of words assigned to each topic is the same. This can be proved by breaking the matrix into product of a diagonal matrix and the normalized matrix and considering when the distributions of singular values will be the same. But unfortunately both the notion of 'well-separatedness' and same number of words getting assigned to each topic are not likely to hold in real world datasets. Still we can hope that increasing the number of topics will make the topic vectors more and more orthogonal to each other, and hence the distribution over singular values will be close enough to the distribution over the row L_2 norms when the 'right' number of topics is reached.

3.3 Topic Splitting

We observe that the following phenomenon holds for the vectors got from LDA for well separated datasets. If a dataset contains K well separated topics and

when LDA is run with $K^{'} > K$ topics, the $K^{'}$ word distributions over topics obtained are also orthogonal to each other. This can happen when topics 'split' which means that a set of words assigned (with high probability) to a particular topic gets assigned to the new topic(s). Thus the new set of topics will still remain orthogonal to each other i.e the new set of topic vectors will form a orthogonal basis in the bigger topic space as well . This becomes an issue if we wish to compare the singular value distributions with the L_2 row norms as they will remain close even after the topics get nearly orthogonal to each other. Of course, we could look at the first topic number where the topics get nearly orthogonal to each other, but this is difficult to judge on real world datasets. So, we require a measure which will increase once the 'right' number of topics is identified.

3.4 Norms in Higher Dimension

As mentioned in section 3.1 , the sums of rows of matrix $M1$ is equal to the sums of columns of matrix $M2$. This of course will not hold once the matrices are made stochastic. Let $M1$ and $M2$ be row-stochastic. Now consider the product of vector L which contains the length of each document with matrix $M2$. We get a vector of length T with components indicating the fraction of each topic present in the corpus. Let the normalized version of this vector be called C_{M1}. (to indicate distribution of topics in the corpus C got from the matrix $M1$). If the lengths of all the documents were the same, this would be equal to the L_1

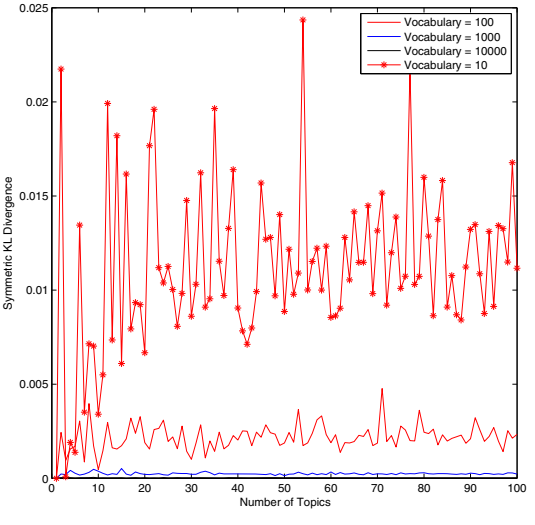

Fig. 1. Plot showing how divergence between L_1 and L_2 norm distributions vary with topics for different vocabulary sizes. As seen, for high vocabulary, the divergence is almost zero.

norm of the row vectors in $M2$. Also let C_{M2} be the distribution over singular values of matrix $M1$.

As the best use of topic models such as LDA arise when the dimension (i.e vocabulary size) is large, we shall assume that the datasets that we deal with are vectors in high dimension. In such cases, we observe that for a random matrix R of order $T * W$, the vector R_{l1} which is the distribution of row L_1 norms and the vector R_{l2} , the distribution over row L_2 norms look very similar component-wise when W is large enough. An example is given in Figure 1 where, as the topics are varied from $T = 1$ to 100, the Symmetric Kullback-Leibler (KL) divergence [9] is calculated for four different values of W. It can be seen that as W becomes large enough, the Symmetric KL divergence goes towards zero. This will happen when the components of both the vectors are very close to each other so that every term in the Symmetric KL divergence of R_{l1} and R_{l2} which is defined as $KL(R_{l1}||R_{l2}) = \sum_{i=1}^{T} R_{l1}(i) * log(R_{l1}(i)/R_{l2}(i)) + \sum_{i=1}^{T} R_{l2}(i) * log(R_{l2}(i)/R_{l1}(i))$ goes to zero.

However, this behavior need not hold true when we consider divergence between distributions generated from non-random matrices such as $M1$ and $M2$. In fact what we observe empirically is that this divergence between the L_1 distribution and the singular value distribution (which is close to the L_2 norm distribution if topics are orthogonal) start to increase once the right number of topics is reached. The reason we believe is the fact that once topic-splitting happens after topics become nearly orthogonal, the L_1 norm acts as a penalty term in the sense that more and more noise gets added in terms of low probability values for words not belonging to a topic which contribute to the increase in divergence value.

3.5 Divergence Measure

We summarize the proposed divergence measure here. For a given corpus C and a given topic T, LDA outputs two stochastic matrices $M1$ and $M2$. The proposed measure is the following:

$$ProposedMeasure(M1,M2) = KL(C_{M1}||C_{M2}) + KL(C_{M2}||C_{M1})$$

where ,
C_{M1} is the distribution of singular values of Topic-Word matrix$M1$,
C_{M2} is the distribution obtained by normalizing the vector $L * M2$ (where L is $1 * D$ vector of lengths of each document in the corpus and $M2$ is the Document-Topic matrix).

A point to be mentioned here is that both the distributions C_{M1} and C_{M2} are in sorted order so that the corresponding topic components are expected to match.

In the next section, we give results for various experiments conducted on both image and text data on toy as well as real world datasets that illustrate the efficacy of the proposed measure in finding the right number of topics.

4 Experiments

4.1 Image Data

Color Palette. We conducted a simple toy experiment on an palette image containing various colors as shown in the left panel of Figure 2. A document in this case was a row of pixel values in the image. Each pixel had three component values for Red, Green and Blue, in the range 0-255. These three component values were concatenated together to form a single value. (For example RGB values of 220,245,230 was made into a single number 220245230) .The image was in *jpg* format and hence the pixels values are not all the same even in a palette with a single perceived color. The number of dimensions (or unique pixel values) for this image is 27777. But as we can see, the number of latent topics is smaller by a huge magnitude. The values of the proposed measure is plotted against various topic values. As we observe from the right pane of Figure 2, the measure dips down close to zero in the range 30 - 40 which in fact is the number of perceived colors ('latent topics') in the palette image.

This simple experiment where the intuitive number of topics is the number of palettes (36 in this case) clearly demonstrates the efficacy of the proposed measure.

Gray-Scale Palette. The immediate idea was to check the effect of dimensionality on the same image. To this end, we converted the image to a gray scale and hence reduced the number of dimensions from 27777 to 255. The same measure for the gray scale image is plotted in Figure 3 . As we see, the dip and rise though not wrong, is not as indicative as it was in the previous case. This suggests that

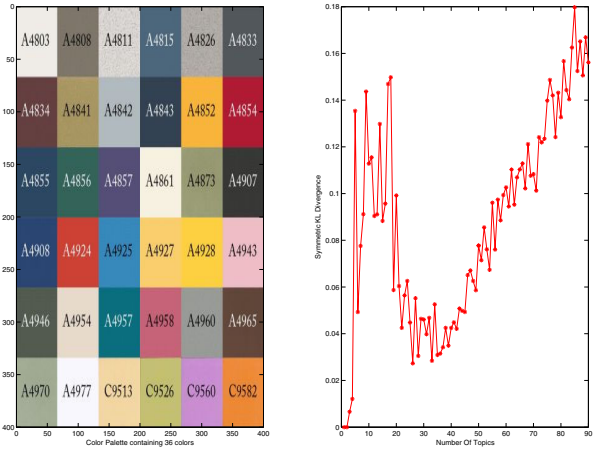

Fig. 2. Number of Topics Vs Symmetric KL Divergence for Color Palette. The dip in divergence is obtained when the number of topics is in the range of 30-40 which is same as the perceived number of 36. (to be viewed in color).

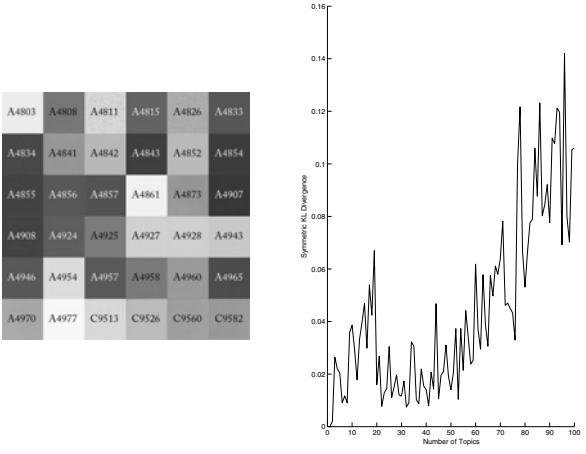

Fig. 3. Number of Topics Vs Symmetric KL Divergence for Gray scale Palette. Note that the range for right number of topics is correct, but not as indicative as in the Color Palette case.

the proposed measure is a good indicator of right number of topics for datasets involving high dimensions.

4.2 Text Data

We conducted several experiments on toy as well as real world text data sets. Each of these is explained below.

Toy Dataset. In this experiment, 12 documents of average length of 500 words were considered. The number of dimensions (vocabulary) was 1525. The documents were wikipedia articles on 3 broad topics (Science , Dance and Dostoyevsky) with 4 documents in each. Each of these topics had sub-topics (such as Quantum Mechanics, Probability, Electromagnetism etc under Science - Salsa, Mambo, Tango etc under Dance - Crime and Punishment, Brothers of Karamazov,Prison Experiences etc under Dostoyevsky). The Measure values are plotted in the left pane of Figure 4. We see a dip starting at around 20 , which is roughly a reasonable number of low level noise free topics to expect from this dataset.

Authorship Dataset. Usually, the top few words in every topic is indicative whether the topics are well split or not. If the right number of topics is reached, the words belonging to every topic are expected to be semantically close to each other. While this might help in deciding the right number of topics for small datasets, it becomes tough to decide the same on abstract datasets. To verify this, we built a dataset containing literary works of 12 authors. Each work was broken up into 5000 word per document and the total number of documents was 834. The details of the dataset are given in table 1. We then stripped off

Table 1. Statistics of the Authorship dataset

Author	Genre[1]	Timeline	Stop-words	Content Words
Daniel Defoe	A,F,Hi	1808-1894	477201	210550
Jane Austen	F,Hu,Ps	1811-1818	474305	248676
Allen Grant	B,F,Sc,Ph	1848-1899	215001	148670
George Elliot	F,Ps	1859-1871	520661	311030
Harold Bindloss	Ro,F	1866-1945	525724	321902
James Otis	A,C,F,Hi	1883-1899	252518	136089
George Bernard Shaw	D,F,Hi,Hu,W	1885-1912	145385	85476
Hamlin Garland	A,F,Ps,Ro,Sp,T	1897-1921	349296	229164
Captain Ralph Bonehill	A	1902	175659	105169
Phillips Oppenheim	F,M,Po	1902-1920	415892	243386
G K Chesterton	M,Ph,Ps,Re	1905-1916	186409	111290
Baronness Orczy	A,Hi,Po,Ro	1905-1921	392127	261019
Total	18	1808-1921	4130178	2307252

the content words and retained words which fell in a small set of 555 stop-words (words such as 'and', 'the' , 'it' etc). We ran LDA on this dataset to test its efficacy in classifying the authors. As the vocabulary contains only stop words, it is not easy to identify the cohesion in a topic by looking at the top few words. Though, the implementation details and other stylometric properties studied from this dataset are explained elsewhere, we just mention here that the accuracy on a held out testset was highest when the number of topics was 15-25. The plot in the right pane of Figure 4 shows the variation of the proposed measure with topics. As we see, the graph seems to linearly increase with a very small dip at around the right number of topics. The reason for the dip not being significant can again be attributed to the number of dimensions being not very large (555 in this case). However, it is easy to infer a broad range for the right number of topics from the plot.

NIPS Dataset/ AP corpus. We now present our results on two real world text datasets. The first one is a standard collection of bag-of-words from NIPS corpus [15] containing 1500 documents with a total of 1932365 words and 12419 dimensions (vocabulary). The plot of our measure for this dataset is shown in left pane of Figure 5 has a dip in the range 100 - 120. [13] compares LDA performance on the same dataset using different number of processors. They consider topics till 80 and observe that irrespective of the processor count,LDA perplexity value goes down from 20 topics to 80 topics on a held out dataset.

[1] A - Adventure, Au - Autobiography, B - Biography, C - Children, D - Drama, F - Fiction Hi - History, Hu - Humour, M - Mystery, Ph - Philosophy, Po - Politics, Ps - Psychology R - Religion, Ro - Romance, Sc - Science, Sp - Spirituality, T - Travel, W - War.

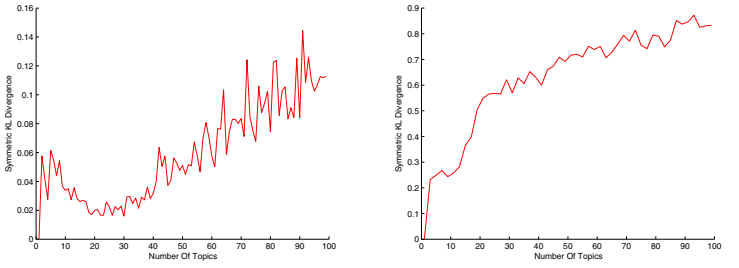

Fig. 4. Plot showing how the proposed measure varies with number of topics for Toy Dataset (Left) and Authorship Dataset(Right)

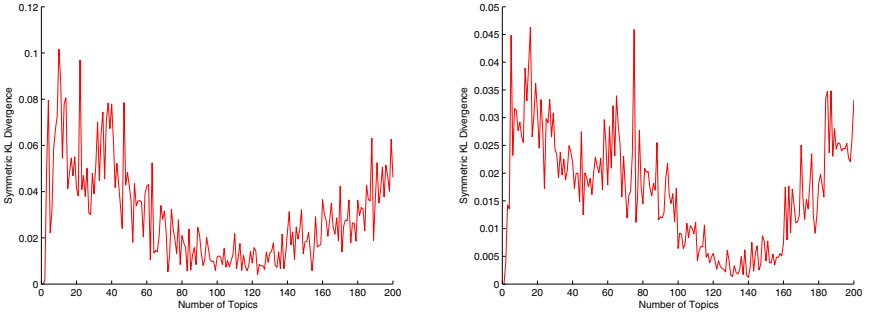

Fig. 5. Plot showing how the proposed measure varies with number of topics for NIPS abstract Dataset (Left) and Associated Press corpus Dataset(Right). The dip is seen the right number of topics for which lowest perplexity is reported.

The second is a collection of articles from the Associated Press dataset [15] containing 2246 documents with 435839 words and 10473 dimensions. The proposed measure plot is shown in the right pane of Figure 5. The best number of topics is seen at around 140. [3] reports that the perplexity reduces until the number of topics is 100 and then there seems to be very little change in the range 100 to 200. But in our case, we observe a steady increase after 130 to 140 topics which is a smaller range for fixing the topic number than what the perplexity values indicate.

5 Discussions and Conclusion

While it might seem that in using the proposed measure, LDA has to be run once for every topic to get to the right number of topics. But this need not be the case always. In most real world datasets, we observe that the variance between divergence values decreases significantly when the right number of topics is reached. This could be taken as a cue to jump appropriately to get nearer to the correct topic number.

Also we mention two points for clarification and emphasis. The first is that comparing the Singular value distribution with the row L_2 norm distribution is not very useful as the measure will not increase once the right number of topics is reached and so it becomes tough to guess the right range of topics. The second point is that Singular value distribution cannot be compared directly with the row L_1 norm of the $M1$ matrix because of its stochasticity. Hence we need to reconstruct the corpus topic distribution from the matrix $M2$. Also it is not correct to compare the singular value distribution (or the row L_2 norm) with the column L_2 norm of $M2$ as the components in a column vector of $M2$ represent distribution over individual documents which will not be same as the distribution over words.

To summarize, we have proposed a new measure for identifying the right number of topics in a give corpus by looking at distributions generated from Topic-Word and Document-Topic matrix outputs of LDA. We showed that the distribution over singular values is close to the distribution over row L_2 norm when the topics become orthogonal . Further, in high dimension the distribution over L_1 and L_2 norms tend to converge for random matrices and hence become candidates for comparison. The measure proposed combines these two facts and compares the singular value distribution of Topic-Word matrix with the row L_1 norm of the Document-Topic matrix. to We illustrated the efficacy of the measure by testing it on several real world and synthetic datasets and on both text and images.

In the future, We hope to explore further in the direction of arriving at more robust theoretical justifications and possible worst case bounds for the proposed measure.

References

1. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by Latent Semantic Analysis. JASIS 41(6), 391–407 (1990)
2. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: SIGIR 1999, pp. 50–57 (1999)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Jordan: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
4. Lee, D.D., Sebastian Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
5. Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S.: A density-based method for adaptive LDA model selection. Neurocomputing 72(7-9), 1775–1781 (2009)
6. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg (2000)
7. Gaussier, E., Goutte, C.: Relation between PLSA and NMF and Implications. In: Proc. 28th international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2005), pp. 601–602 (2005)
8. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika 1(3), 211–218 (1936)

9. Kullback, S., Leibler, R.A.: On Information and Sufficiency. Annals of Mathematical Statistics 22(1), 79–86 (1951)
10. Zavitsanos, E., Petridis, S., Paliouras, G., Vouros, G.A.: Determining Automatically the Size of Learned Ontologies. In: ECAI 2008, pp. 775–776 (2008)
11. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In: NIPS 2004 (2004)
12. Blei, D.M., Lafferty, J.D.: Correlated Topic Models. In: NIPS 2005 (2005)
13. Smyth, P., Welling, M.: Asynchronous Distributed Learning of Topic Models. In: NIPS 2008, pp. 81–88 (2008) (bibliographical record in XML Arthur Asuncion)
14. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: ICML 2006, pp. 577–584 (2007)
15. `http://archive.ics.uci.edu/ml/datasets/Bag+of+Words`, `http://www.cs.princeton.edu/~blei/lda-c/`