# Fractal approach for determining the optimal number of topics in the field of topic modeling.

**V Ignatenko**[1], **S Koltcov**[1], **S Staab**[2] **and Z Boukhers**[2]

[1] National Research University Higher School of Economics, ul. Sedova 55/2, 192148, Saint-Petersburg, Russia

[2] Institute for Web Science and Technologies, University of Koblenz-Landau, Universitaetsstrasse 1, 56070, Koblenz, Germany

E-mail: vignatenko@hse.ru

**Abstract.** In this paper we apply multifractal formalism to the analysis of statistical behaviour of topic models under condition of varying number of topics. Our analysis reveals the existence of two self-similar regions and one transition region in the function of density-of-states depending on the number of topics. As earlier a function that can be expressed through density-of-states was successfully used to determine the optimal number of topics, we test the applicability of the density-of-states function for the same purpose. We provide numerical results for three topic models (PLSA, ARTM, and LDA Gibbs sampling) on two marked-up collections containing texts in two different languages. Our experiments show that the "true" number of topics, as determined by the human mark-up, occurs in the transition region.

## 1. Introduction

Modern information systems generate a huge number of texts such as news, blogs and comments. Analysis of big data is impossible without the construction of formalized mathematical models, and the latter gain a lot from using statistical physics. One of such a model is topic modelling. Briefly speaking, topic modeling (TM) is a family of mathematical algorithms based on the following assumptions [1]:

1. Let $D$ be a collection of textual documents, $\tilde{W}$ be a set (vocabulary) of all unique words, and the number of elements of vocabulary be denoted by $W$. Each document $d \in D$ is a sequence of terms $w_1, ..., w_{n_d}$ from the vocabulary $\tilde{W}$.
2. It is assumed that there exists a finite set of topics $\tilde{T}$ with a finite number of topics $T$, and each entry of a word $w$ in document $d$ is associated with a certain topic $t \in \tilde{T}$. Topic is understood as a combination of words which often (in statistical sense) occur together in a large number of documents.
3. Collection of documents is considered a random and independent sample of triples $(w_i, d_i, t_i)$, $i = 1, ..., n$, from a discrete distribution $p(w, d, t)$ on a finite probability space $\tilde{W} \times D \times \tilde{T}$. Words $w$ and documents $d$ are observable variables, topic $t \in \tilde{T}$ is a latent (hidden) variable.
4. It is assumed that word order in documents is not important for topic detection ('bag of words' model). Order of documents in a collection does not matter as well.

In TM it is assumed that probability $p(w|d)$ of term $w$ to occur in document $d$ can be expressed by multiplication of conditional probabilities $p(w|t)$ and $p(t|d)$. According to the formula of total probability and the hypothesis of conditional independence, we obtain the following expression [1]:

$$p(w|d) = \sum_{t \in \tilde{T}} p(w|t)p(t|d) = \sum_{t \in \tilde{T}} \phi_{wt}\theta_{td}, \tag{1}$$

where $p(w|t) := \phi_{wt}$ is the probability of word $w$ to belong to topic $t$, $p(t,d) := \theta_{td}$ is the probability of topic $t$ in document $d$.

Thus, to construct a topic model of data means to find the set of latent topics $\tilde{T}$ based on observable variables $d$ and $w$, i.e. to find (1) a set of one-dimensional conditional probabilities $p(w|t) \equiv \phi_{wt}$ for each topic $t$ (they form a word-topic matrix $\Phi \equiv \{\phi_{wt}\}_{w \in \tilde{W}, t \in \tilde{T}}$) and (2) a set of one-dimensional conditional probabilities $p(t|d) \equiv \theta_{td}$ for each document $d$ (they form a topic-document matrix $\Theta \equiv \{\theta_{td}\}_{t \in \tilde{T}, d \in D}$).

To date, a large number of TM models have been proposed, but we focus on the following two types: 1. Models based on Gibbs sampling procedure; 2. Models based on E-M algorithm. However, although a great variety of algorithms has been proposed within these two approaches, they all share the problem of selecting the number of topics.

The rest of the paper proceeds as follows. Section 2 explains the logic of the fractal approach used for the analysis of topic models. Section 3 describes the data used and the results of numerical experiments performed to verify our approach.

## 2. Fractal approach

Our fractal approach is based on the following assumptions: 1. The set of documents and words is considered a mesoscopic system, where the number of elements can reach several millions [2]. 2. Collection of documents contains the finite number of topics which is unknown in advance. Let us note that variation of the number of topics in an algorithm of TM allows to regulate algorithm resolution.

Recall that under the condition of fixed number of topics, a topic solution is a matrix $\Phi$, where $T \cdot W$ is the number of elements, $T$ is the number of topics (the number of columns of the matrix), $W$ is the number of unique words. Each cell of the matrix contains probability $\phi_{ij}$ of belonging of a word $w_i$ to a topic $t_j$. The multidimensional space of words is covered by a grid of fixed size defined by matrix $\Phi$. The size of each cell of this grid is $\epsilon = 1/(WT)$. Under the condition of fixed size of vocabulary $W$, the size of each cell is defined by the number of topics and if $T \to \infty$ then the size of the cell tends to zero. Let us introduce the density-of-states function which is defined according to the following formula [3]: $\rho = \frac{n}{WT}$, where $n$ is the number of cells of $\Phi$ satisfying $\phi_{ij} > 1/W$ for $i = 1, ..., W$, $j = 1, ..., T$. It was shown in [3] that the behaviour of such function is extremely nonlinear, and the optimal number of topics for a topic model corresponds to the minimum of normalized free energy or the minimum of Massieu function. It is clear that the density-of-states function depends on the number of topics and changes in the process of topic modeling. Thus, the density-of-states function depends on the cell size and on some degree $D(\epsilon)$ [4], [5]: $\rho(\epsilon) \approx \epsilon^{D(\epsilon)}$. The distribution of fractal dimensions $D(\epsilon)$ can be found using 'box counting' algorithm. Application of this algorithm to the calculation of fractal dimensions in topic models consists of the following steps: 1. A certain number of topics is chosen. 2. Multidimensional space of words and topics is covered by a grid of fixed size (matrix $\Phi$). 3. We calculate the number of cells satisfying $\phi_{ij} > 1/W$. 4. We calculate the value of $\rho$ for chosen number of topics $T$. 5. We repeat steps 1 through 4 changing the cell size (i.e. changing the number of topics). 6. We plot a graph showing dependence of $\rho$ in bi-logarithmic coordinates. 7. Using the method of least squares, we estimate the slope of the function. The value of the slope is the value of fractal dimension calculated according to the following formula: $D(\epsilon) = \frac{\ln(\rho(\epsilon))}{\ln(\epsilon)}$.

We assume that the optimal number of topics corresponds to the regions where the fractal dimensions change since for the regions where the fractal dimension is constant, the solution of TM preserves its structure, while, the changes in structure correspond to changes in fractal dimensions.

## 3. Numerical experiments

In this research, the following computer experiments were executed. We used three algorithms of TM, namely, 1. PLSA [1] (E-M algorithm) as implemented in BigARTM package (http://bigartm.org/); 2. ARTM [6] (E-M algorithm) as implemented in BigARTM package; 3. LDA [7] Gibbs sampling as implemented in GibbsLDA++ package (http://gibbslda.sourceforge.net/). Two collections of textual documents were used: 1. English-language dataset ’20 Newsgroups’ (http://qwone.com/ jason/20Newsgroups/) containing 15404 news texts with 50948 unique words. The documents of this dataset are manually labeled with a topic class among 20 topic classes. Since some of these topics are similar and can be united, the dataset can be represented with 15 topics. 2. Russian-language dataset ’Lenta_ru’ consists of 8630 documents (containing 23297 unique words) in Russian language, each of which is manually labeled with a class among 10 topic classes. However, some of these topics can be viewed as parts of other topics, therefore, the documents in this dataset can be represented with 7 distinct topics. When conducting topic modeling on these datasets, the number of topics was varied in the range T=[2;50] in the increments of 1 topic. For each topic solution the value of the density-of-states function was calculated. The obtained curves for two collections were analysed in bi-logarithmic coordinates. These collections were chosen to represent different languages and different dataset sizes.

Fig. 1 shows the distribution of fractal dimensions for the English-language dataset implemented using PLSA algorithm. The figure demonstrates that one can distinguish three regions. The first and the third regions are approximated by linear functions using the method of least squares. Linear regions correspond to the situation when density-of-states function is self-similar, i.e. $\rho(\lambda\frac{1}{WT}) = \lambda^d\rho(\frac{1}{WT})$. Thus, this function contains self-similar regions that possess the properties of fractals. The second region is a transition region, where the first fractal set transforms into another fractal set. Notice that the transition region corresponds to $T = [9, 15]$, the fractal dimension for the first region is equal approximately to 0.805 and for the third region to 0.375.

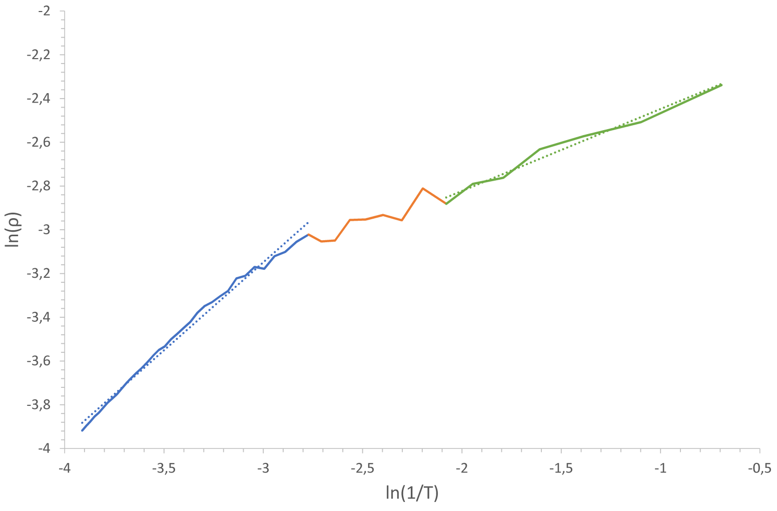

**Figure 1.** Distribution of fractal dimensions for English-language dataset (PLSA model).

In paper [8] a set of cluster algorithms with different types of regularization was investigated. The authors showed that optimal cluster solutions lie in the range of $[15, 20]$ clusters for ’20

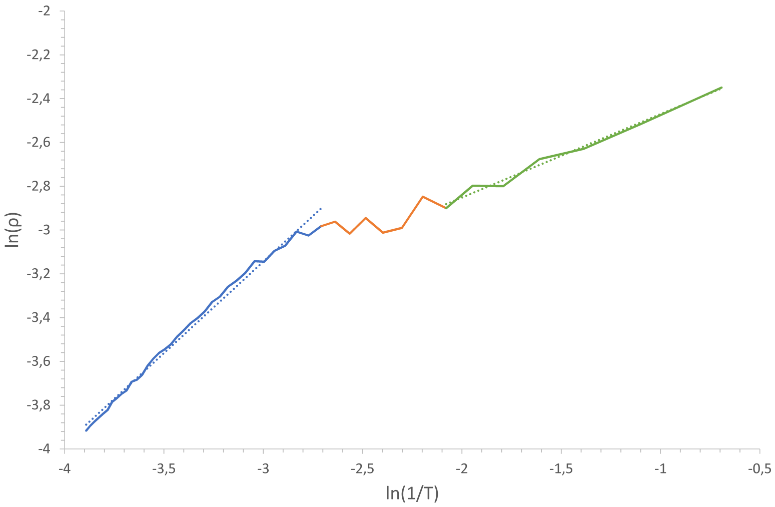

**Figure 2.** Distribution of fractal dimensions for English-language dataset (LDA model, $\alpha = 0.4$, $\beta = 0.5$).

Newsgroups dataset'. In our experiment, restructuring of a topic model occurs in the region of $[9, 15]$ topics which is slightly lower than the "true" number. Fig. 2 shows the distribution of fractal dimensions for the English-language dataset implemented using LDA algorithm with the following values of hyperparameters $\alpha = 0.4$ and $\beta = 0.5$. Let us note that transition region in Fig. 2 corresponds to $T = [8, 15]$. The fractal dimension for the first region of Fig. 2 equals approximately to 0.834 and for the third region to 0.381. Fig. 3 and 4 show the distribution of fractal dimensions for '20 Newsgroups dataset' implemented using ARTM algorithm with different values of regularization coefficients (sparse_phi). One can see that the location of the transition region changes significantly under variation of the values of regularization coefficient. Therefore transition region corresponds to approximately $T = 18$ on Fig. 3 and on Fig. 4 it corresponds to $T = [3, 7]$. Hence, one can conclude that values of regularizing coefficients for the ARTM model influence the structure of the results of TM significantly, and it leads to changing in the optimal number of topics. The fractal dimension for the first region of Fig. 3 equals approximately to 0.853, and for the second region to 0.347. The fractal dimension for the first region on Fig. 4 equals to 0.87, the third region consists only of two points, hence, it makes no sense to speak about fractal dimension for that region.

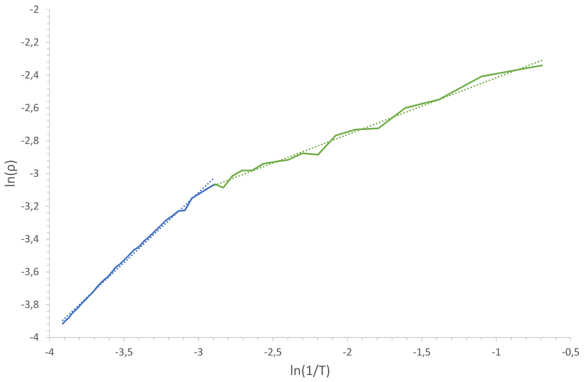

**Figure 3.** Distribution of fractal dimensions for English-language dataset (ARTM, sparse_phi=0.01).

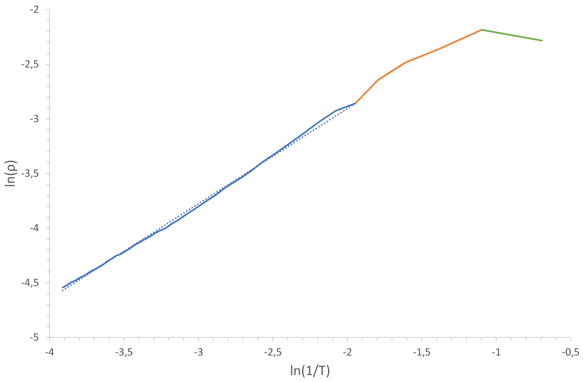

**Figure 4.** Distribution of fractal dimensions for English-language dataset (ARTM, sparse_phi=-10).

Let us consider numerical results for the Russian-language dataset. Fig. 5 shows the distribution of fractal dimensions for topic model based on the PLSA algorithm. The transition region in this case corresponds to $T \approx 7$. The fractal dimension for the first region corresponds to $D \approx 0.688$ and for the second $D \approx 0.448$. Fig. 6 shows the distribution of fractal dimensions for LDA Gibbs sampling model with the following values of hyper-parameters $\alpha = 0.4$ and $\beta = 0.5$. Note that the transition region on Fig. 6 corresponds to $T = [6, 11]$. The fractal dimension for the first region is $D \approx 0.708$ and for the second $D \approx 0.49$. Fig. 7 and 8 demonstrate the distribution of fractal dimensions for the ARTM model with different values of "sparse_phi" coefficient. Transition region on Fig. 7 corresponds to $T = [6, 9]$, and on Fig. 8 it corresponds to $T = [3, 4]$. For the first region of Fig. 7 the fractal dimension is $D \approx 0.73$ and for the third it is $D \approx 0.447$. Finally, on Fig. 8 the fractal dimensions for the first and for the third regions are $D \approx 1.006$ and $D \approx 0.744$, correspondingly.

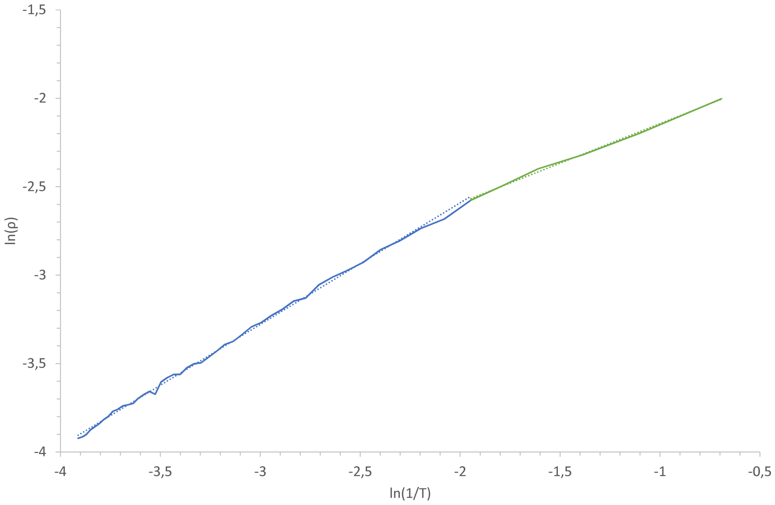

**Figure 5.** Distribution of fractal dimensions for Russian-language dataset (PLSA model).

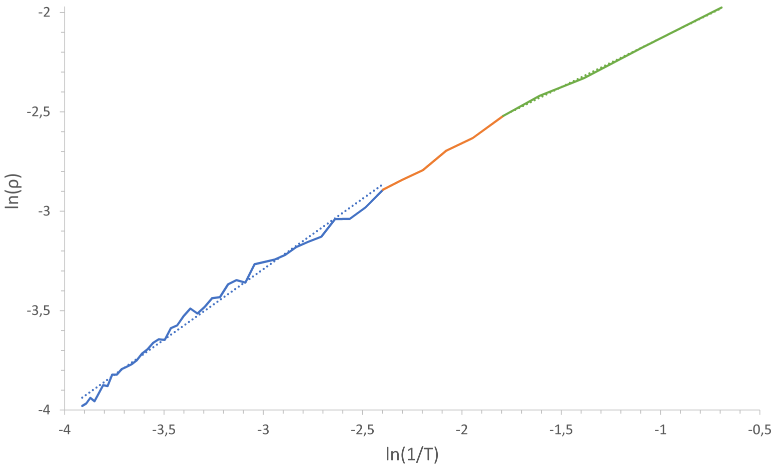

**Figure 6.** Distribution of fractal dimensions for Russian-language dataset (LDA, $\alpha = 0.4$, $\beta = 0.5$).

## 4. Conclusion

Fractal analysis of topic model behavior under the condition of changing number of topics has shown that self-similar fractal regions exist in the density-of-states function. We have also found out that this function is not a perfect fractal; instead, it always contains at least one transition region that separates self-similar parts of the curve. Next, we have tested the assumption

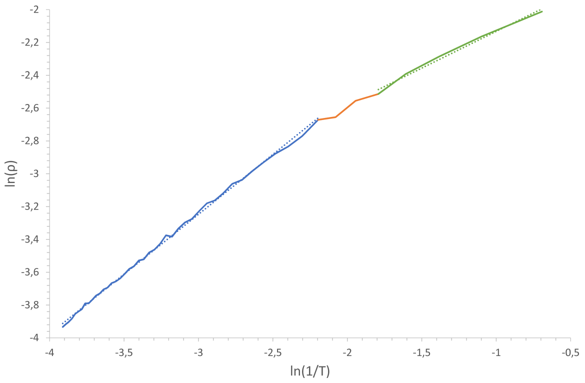

**Figure 7.** Distribution of fractal dimensions for Russian-language dataset (ARTM, sparse_phi=0.01).

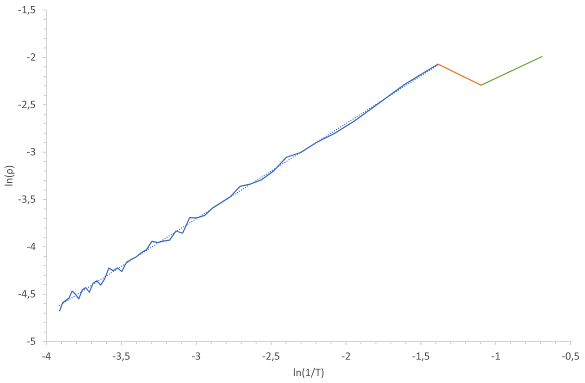

**Figure 8.** Distribution of fractal dimensions for Russian-language dataset (ARTM, sparse_phi=-10).

that this transition region corresponds to the "true" number of topics, as determined by the human mark-up. In our limited number of tests, the best performing algorithm was ARTM with "sparse_phi" = 0.01, although other algorithms have shown promising results, too. More experiments are needed to determine an algorithm (or algorithms) that is best suited for the usage of density-of-states function as an indicator of the optimal number of topics. If such algorithm is found that can perform well on a large number of different datasets, it will give a more solid ground for the determining the optimal number of topics through finding a transition region. This will confirm the applicability of the fractal approach to topic number optimization that has been preliminarily tested in this paper.

### *4.1. Acknowledgments*

The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2018.

## References

[1] Hofmann T 1999 *Proc. of the 22nd Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval (August 15 - 19, 1999, Berkeley)* (New York, NY, USA: ACM) p 50

[2] Koltcov S 2018 *Physica* A: Statistical Mechanics and its Applications **512** 1192

[3] Koltcov S 2017 *Tech. Phys. Letters* **43** 584

[4] Feder J 1988 *Fractals* (Boston: Springer)

[5] Sornette D 2006 *Critical Phenomena in Natural Sciences* (Heidelberg: Springer-Verlag)

[6] Kochedykov D, Apishev M, Golitsyn L and Vorontsov K 2017 *Proc. of the 21st Conf. of Open Innovations Association FRUCT (November 6-10, 2017, Helsinki)* (Helsinki, Finland: FRUCT Oy) pp 182–193

[7] Griffiths T and Steyvers M 2004 *PNAS* **101** 5228

[8] Basu S *et al.* (eds) 2008 *Constrained Clustering: Advances in Algorithms, Theory, and Applications* (Taylor & Francis Group)